Rationalisation de l'Accès aux Produits Naturels Fongiques par une Approche OSMAC in silico : Cas d'étude avec la modélisation du métabolisme de Penicillium rubens

Delphine Negre

COLLEGE VEGETAL
DOCTORAL ANIMAL, ALIMENT
PAYS DE LA LOIRE MER, ENVIRONNEMENT

Nantes Université

THÈSE DE DOCTORAT

NANTES UNIVERSITE

ECOLE DOCTORALE N° 642
Ecole doctorale Végétal, Animal, Aliment, Mer, Environnement
Spécialité : *Génétique, génomique et bio-informatique*

Par
Delphine NÈGRE

Rationalisation de l'Accès aux Produits Naturels Fongiques par une Approche OSMAC *in silico.*

Cas d'étude avec la modélisation du métabolisme de *Penicillium rubens*

Thèse présentée et soutenue à Nantes, le 13 décembre 2024
Unités de recherche : Institut des Substances et Organismes de la Mer, ISOMer, UR-2160, UFR des Sciences Pharmaceutiques et Biologiques, Nantes Université
Laboratoire des Sciences du Numérique de Nantes, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Rapporteurs avant soutenance :

Sabine PERES	Professeure des Universités, Université de Lyon - INRIA
Marco PAGNI	Chercheur, Swiss Institute of Bioinformatics, Lausanne, Suisse

Composition du Jury :

Président :	**Nicolas PAPON**	Professeur des Universités, Université d'Angers
Examinateurs :	**Delphine ROPERS**	Directrice de Recherche, Inria Grenoble
	Jérémie BOURDON	Professeur des Universités, Nantes Université
Dir. de thèse :	**Samuel BERTRAND**	Maître de Conférences – HDR, Nantes Université
Co-encad. de thèse :	**Abdelhalim LARHLIMI**	Maître de Conférences, Nantes Université

REMERCIEMENTS

Je remercie vivement les membres de mon jury de thèse composé de Mme **Sabine Peres** (Professeure des Universités, Université de Lyon – INRIA), M. **Marco Pagni** (Chercheur, Swiss Institute of Bioinformatics, Lausanne, Suisse), Mme **Delphine Ropers** (Directrice de Recherche, Inria Grenoble), M. **Jérémie Bourdon** (Professeur des Universités, Nantes Université) et M. **Nicolas Papon** (Professeur des Universités, Université d'Angers) d'avoir accepté d'évaluer mes travaux de recherche. Merci sincèrement pour le temps consacré à la lecture de ce manuscrit et pour votre présence à ma soutenance.

La rédaction de mon manuscrit et la réalisation de ces travaux n'auraient jamais été réalisables sans le soutien financier de l'Agence Nationale de Recherche (ANR-18-CE43-0013-01 - Induction Rationnelle de Produits Naturels Fongiques – FREE-NPs) et l'accompagnement de plusieurs personnes auxquelles je souhaite exprimer ma reconnaissance.

Je souhaite ainsi remercier les membres de mon comité de suivi individuel, **David Touboul** et **Erwan Corre**, qui, par leur accompagnement bienveillant, leur intérêt à l'égard de mon sujet de thèse ainsi que leurs conseils avisés, ont contribué à orienter et à structurer mes recherches.

J'adresse des remerciements particuliers à **Anne Siegel** et **Jeanne Got** de l'équipe Dylissà l'Irisa de Rennes, à **Laurence Meslet-Cladière** du Laboratoire Universitaire de Biodiversité et d'Ecologie Microbienne à Brest, **Annie Lebreton** et **Erwan Corre** de l'équipe ABiMS de la station Biologique de Roscoff pour leur collaboration et le partage de données essentielles à cette étude.

Je remercie M. **Yves-François Pouchus**, ancien Directeur du laboratoire Mer, Molécules, Santé et directeur de ma thèse lors des premiers mois, de m'avoir accueillie et de m'avoir donné l'occasion de réaliser cette thèse au sein du laboratoire. Je remercie également **les membres des équipes M3** du laboratoire ISOMer et **ceux de l'équipe ComBi** du LS2N.

Je suis profondément reconnaissante à **Samuel Bertrand** et **Abdelhalim Larhlimi**, directeur et encadrant de cette thèse, pour leur accompagnement, leur disponibilité constante et leurs conseils qui ont guidé mes travaux tout au long de ces années. Outre le soutien scientifique que vous m'avez apporté, je vous remercie pour votre patience, votre écoute et votre compréhension.

Mes vifs remerciements à **Aurore**, **Marjorie**, **Simon** et **Toni** pour nos échanges au laboratoire et merci d'avoir égayé, entre autres, mes pauses-café. Merci **Aurore**, Sainte-Patronne des doctorants, pour tes petites et grandes attentions du quotidien ainsi que ton éternel dévouement. Merci **Marjorie** d'avoir été là et de ne pas m'avoir lâchée.

Je tiens à remercier également les trois stagiaires qui ont contribué à ce projet : **Florence Thomas-Giraud**, **Clément Poulain** et **Aïmen El Assimi**.

Les travaux de cette thèse ne se résument pas à ces cinq dernières années, et je tiens à exprimer ma profonde gratitude envers celles et ceux qui ont contribué, bien en amont, à leur réalisation.

Je remercie particulièrement les personnes que j'ai eu l'opportunité de rencontrer durant mes 18 mois à la Station Biologique de Roscoff, et plus spécialement les membres de l'équipe ABiMS. Mes remerciements chaleureux vont essentiellement à **Gabriel Markov** et **Erwan Corre**. Merci à vous de m'avoir fait et donné confiance et Merci pour m'avoir fait découvrir et donner goût au monde de la Recherche.

Et puis, il y a les amis, ceux qui ont compté, ceux qui ont pris des chemins différents, et ceux qui sont toujours là. Merci à toutes celles et ceux qui, à un moment ou à un autre, ont su m'encourager, me soutenir, m'écouter (râler), me secouer, me réconforter, m'héberger, me conseiller, me faire rire, me challenger, me remotiver, m'inspirer, m'accompagner dans les moments de doute et enfin, me comprendre et croire en moi. En résumé tous ceux et celles qui ont su me (sup)porter depuis toutes ces années.

Merci donc à **Mouna** pour avoir su trouvé les mots et me donner le courage de me lancer dans l'aventure. À **Suzanne** pour ta patience et pour m'apprendre à garder les pieds sur terre. À **Marine** pour toutes ces années et moments passés depuis le lycée. À **Guita**, mon petit rayon de soleil sous le ciel gris de Bretagne. À **Émilien** pour nos échanges téléphoniques beaucoup trop rares mais tellement enrichissants. À **Aurélien** pour m'avoir recueillie en galère à mon arrivée sur Nantes. À **Chloé**, pour tous nos moments de partage et de rires. À **Jo**, mon petit hibou, pour nos délires et absurdités qui ne me font rire que nous. À **Sophie**, pour nos aventures hasardeuses et nos souvenirs inoubliables. À **Mathilde**, **Jo**, **Gaëlle** et **à vos « petites » tribus respectives**, pour votre soutien indéfectible. Merci à vous d'être là depuis si longtemps et de faire partie de mes pilliers.

Enfin, mes derniers remerciements, et non des moindres, vont à **mes parents**. Merci **Papa** pour m'avoir enseigné la rigueur et de m'avoir mis un ordinateur dans les mains depuis toute pitchoune. Merci **Maman** pour les heures de relecture, pour ta patience, et ton amour inconditionnel. Merci à vous d'être présents, de m'accompagner, de m'épauler depuis toujours. Merci pour votre soutien sans faille, vos encouragements constants et pour les valeurs que vous m'avez transmises qui m'ont guidée tout au long de ces années. Sans vous, rien de tout cela n'aurait été possible.

“***What I cannot create, I do not understand.***”

Tableau noir de Richard Feynman au moment de sa mort, physicien théoricien américain, 1988

« *[…]*

Hâtez-vous lentement ; et, sans perdre courage,

Vingt fois sur le métier remettez votre ouvrage :

Polissez-le sans cesse et le repolissez ;

Ajoutez quelquefois, et souvent effacez.

[…] »

Nicolas Boileau, poète français, *Art poétique*, 1674, Chant I, v. 147-207

‘*When you have eliminated the impossible, whatever remains, however improbable, must be the truth?*’

Arthur Conan Doyle, écrivain britannique, *A Study in Scarlet*, 1887, Chap. 6, p. 111

‘***All models are wrong, but some are useful.***’

George E. P. Box, statisticien britannique*, Robustness in Statistics*, 1978, pp. 201–236

– Table des Matières –

– Index des Outils –

Cet index répertorie l'ensemble des outils et bases de données utilisés au cours de nos travaux. Ceux inscrits **en gras** font l'objet d'une discussion plus approfondie au sein du Chapitre 2.

– Liste des Figures –

INTRODUCTION GÉNÉRALE

CHAPITRE 1

CHAPITRE 2

CHAPITRE 3

CHAPITRE 4

– Liste des Tableaux –

CHAPITRE 1

CHAPITRE 3

CHAPITRE 4

– Liste des Annexes –

– Liste des Abréviations –

ADN	*Acide Désoxyribonucléique*
ARN	*Acide Ribonucléique*
ATP	*Adénosine Triphosphate*
AuReMe	*Automatic Reconstruction of Metabolic models*
BBH	*Best Blast Hit*
BGC	*Biosynthetic Gene Cluster*
CBM	*Constraint-Based Modelling*
ChEBI	*Chemical Entities of Biological Interest*
CoA	*Coenzyme A*
COBRA	*Constraint-Based Reconstruction and Analysis*
COMBINE	*Computational Modeling in Biology Network*
CYA	*Czapek Yeast Extract Agar*
DMATS	*Dimethylallyl Tryptophan Synthase*
EC	*Enzyme Commission*
FAIR	*Findable, Accessible, Interoperable, Reusable*
FBA	*Flux Balance Analysis*
FBC	*Flux Balance Constraints*
FROG	*Flux variability analysis, Reaction deletion, Objective function values, Gene deletion fluxes*
FVA	*Flux Variance Analysis*
GAM	*Growth Associated Maintenance*
GO	*Gene Ontology*
GPR	*Gene-Protein-Reaction*
GSMN	*Genome-Scale Metabolic Network*
HTML	*Hyper Text Markup Language*
JGI	*Joint Genome Institute*
KAAS	*KEGG Automatic Annotation Server*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
KO	*KEGG Orthology*
MEA	*Malt Extract Agar*
MEMOTE	*Metabolic Model Tests*
Meneco	*Metabolic Network Completion*
MIASE	*Minimum Information About a Simulation Experiment*
MIRIAM	*Minimum Information Requirements in Biochemical Model Annotation*
MMS5	*Marine Mycological Strains*
NAD/H	*Nicotinamide Adenine Dinucleotide / réduit*
NADP/H	*Nicotinamide Adenine Dinucleotide Phosphate / réduit*
NGAM	*Non-Growth Associated Maintenance*
NP	*Natural Product*
NRPS	*Nonribosomal Peptide Synthetases*
ORF	*Open Reading Frame*
OSMAC	*One Strain Many Compounds*
PDA	*Potato Dextrose Agar*
PGDB	*Pathway Genome Database*
PhPP	*Phenotype Phase Plane*
PKS	*Polyketide Synthases*
SBML	*Systems Biology Markup Language*
SBO	*Systems Biology Ontology*
SMILES	*Simplified Molecular Input Line Entry System*
TS	*Terpène Synthase*
UDP	*Uridine Diphosphate*
URI	*Uniform Ressource Identifier*
XML	*Extensible Markup Languag*

– Liste des Communications Scientifiques –

PUBLICATION DANS JOURNAL SCIENTIFIQUE

Nègre D., Larhlimi A., Bertrand S. **Reconciliation and evolution of *Penicillium rubens* Genome-Scale Metabolic Networks–What about specialised metabolism?** *PLOS ONE*. 2023;18(8)
https://doi.org/10.1371/journal.pone.0289757

COMMUNICATION DANS DES CONGRÈS INTERNATIONAUX

Communication orale

Nègre D., Larhlimi A., Bertrand S. **Genome Scale Metabolic Network Reconstruction, a model for specialised metabolites detection? Illustration on the filamentous fungus model *Penicillium chrysogenum***. (Septembre 2022) Analytics 2022, Nantes, France.

Communication affichée

Nègre D., Larhlimi A., Watier E., Lescaut N., Bourdon J., Lortheau E., Siegel A., Nicolas J., Meslet-Cladiere L., Rouillier C., Gentil E., Ruiz N., Grovel O., Pouchus Y-F., Bertrand S. **Genome-Scale Metabolic Network: an interconnected platform for the understanding of basal and specialised metabolism of filamentous fungi *Penicillium chrysogenum*.** (Novembre 2021) Réseau Francophone de Métabolomique et Fluxomique (RFMF), Aussois, France.

COMMUNICATIONS DANS DES CONGRÈS NATIONAUX

Communication orale

Nègre D., Bertrand S., Larhlimi A. **La reconstruction de réseaux métaboliques à l'échelle du génome, une modélisation au service de la détection de métabolites spécialisés. Illustration sur le modèle de champignon filamenteux *Penicillium chrysogenum*.** (Décembre 2021) Journée d'animation scientifique de Biogenouest: Analyse structurale et Métabolomique, visioconférence, France.

Communications affichées

Nègre D., Larhlimi A., Watier E., Bourdon J., Lortheau E., Siegel A., Nicolas J., Meslet-Cladiere L., Rouillier C., Gentil E., Ruiz N., Grovel O., Pouchus Y-F, Bertrand S. **Genome Scale Metabolic Networks of *Penicillium chrysogenum*: evolution, combination and new reconstruction.** (Juin-Juillet 2020) Les Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), visioconférence, France.

Nègre D., Larhlimi A., Siegel A., Bertrand S. **Genome-Scale Metabolic Network: an interconnected platform for the understanding of basal and specialised metabolism of filamentous fungi *Penicillium chrysogenum***. (octobre 2021) GDR Médiation chimique dans l'environnement – Ecologie Chimique (MediaTEC), Toulouse, France.

Nègre D., Bertrand S., Larhlimi A. **A new *Penicillium chrysogenum* Genome Scale-Metabolic Network: reconciliation of previous data and focus on specialised metabolism**. (Juillet 2022) Les Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Rennes, France.

Introduction Générale

La découverte des produits naturels a traditionnellement reposé sur le criblage aléatoire d'extraits d'organismes vivants, générant des molécules essentielles, tant dans le domaine de la santé que dans celui de l'agriculture, puisque de nombreux médicaments et antibiotiques sont constitués de ces molécules naturelles biologiquement actives *(Newman et Cragg 2012; Schueffler et Anke 2014)*. Cependant, au cours des dernières années, la recherche de nouveaux produits naturels tend à décliner, notamment en raison de l'inefficacité des méthodes de bioprospection *(David et al. 2015)*. Face à l'augmentation de la résistance aux antibiotiques, il devient crucial de réorienter ces recherches vers des approches plus rationnelles. L'intégration des approches génomiques et métabolomiques, soutenue par des avancées technologiques, offre un potentiel considérable pour redynamiser la découverte de nouveaux métabolites biologiquement actifs.

Les champignons filamenteux sont largement reconnus pour produire une vaste diversité et une grande quantité de produits naturels *(Wiemann et Keller 2014)*. L'abondance des processus qui régissent leur morphogénèse témoigne directement de la capacité de ces organismes à produire une large diversité de métabolites. Les conditions de culture externes, notamment la composition du milieu de croissance, affectent le développement morphologique des champignons filamenteux *(Grimm et al. 2005)*, qui expriment alors divers phénotypes (**Figure 0-1**). De plus, les modulations qualitatives et quantitatives des milieux nutritifs, même minimes, ont une influence sur la composition et la quantité produite de métabolites par ces organismes *(Camila Dos Santos Dias et al. 2017)*. Des observations également corroborées par les profils chimiques obtenus pour la souche *Penicillium rubens* Wisconsin 54-1255, cultivée sur divers milieux solides au laboratoire ISOMer qui démontrent l'impact du substrat sur leur composition métabolique (**Figure 0-2**). Dès le milieu des années 2000, la problématique liée au besoin de générer des modèles intégratifs qui associeraient le développement morphologique au métabolisme et à la productivité, est apparue comme une voie prometteuse pour une conception rationnelle et ciblée des processus de culture *(Grimm et al. 2005)*. De tels modèles offriraient alors une approche systémique pour rationaliser la bioproduction et exploiter pleinement le potentiel biosynthétique des champignons filamenteux.

L'analyse des génomes des champignons filamenteux a mis en évidence une forte proportion de métabolites spécialisés produits par des enzymes codées par des gènes co-localisés dans des clusters de gènes biosynthétiques (BGC). Ces clusters, comprenant des gènes de résistance, de transport ou de régulation, fournissent des indications sur les activités biologiques et permettent d'orienter les recherches en ingénierie moléculaire *(Eustáquio et Ziemert 2018)*. Toutefois, une grande partie de ces clusters reste non exprimée dans des conditions de culture en laboratoire, représentant ainsi une ressource inexploitée pour la production de produits naturels *(Nielsen et al. 2017)*. Comprendre les conditions d'activation de ces gènes cryptiques est alors essentiel pour débloquer leur potentiel.

La recherche de métabolites spécialisés peut être sensiblement améliorée par l'ajustement des conditions de culture des micro-organismes, notamment à travers des changements dans le régime nutritionnel, les paramètres physiques, la co-culture ou l'ajout d'éliciteurs. L'approche **OSMAC** (« *One Strain-Many Compounds* ») explore la diversité naturelle en fonction de ces variations *(Bode et al. 2002)*. Cependant, bien que prometteuse, cette méthode s'avère chronophage et coûteuse, rendant nécessaire le recours à la modélisation pour optimiser les tests et sélectionner les conditions de culture les plus pertinentes.

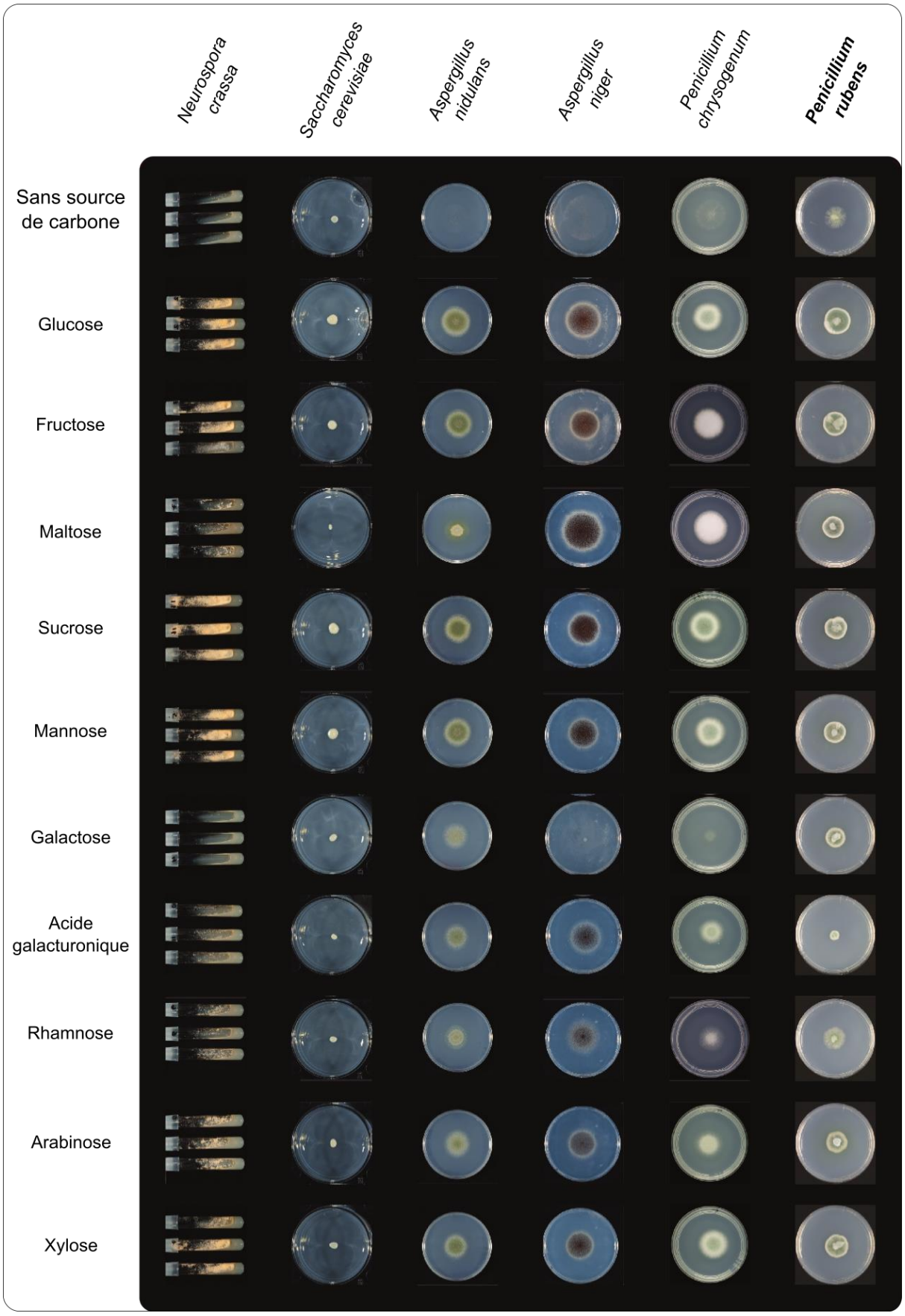


Figure 0-1 : Profils de croissance de diverses espèces de champignons filamenteux sur différentes sources de carbone. *Ces profils permettent d'évaluer l'aptitude de chaque espèce à utiliser des substrats spécifiques, révélant ainsi des variations dans les capacités métaboliques et l'efficacité de l'assimilation des nutriments. Extrait des travaux de (Aguilar-Pontes et al. 2018).*

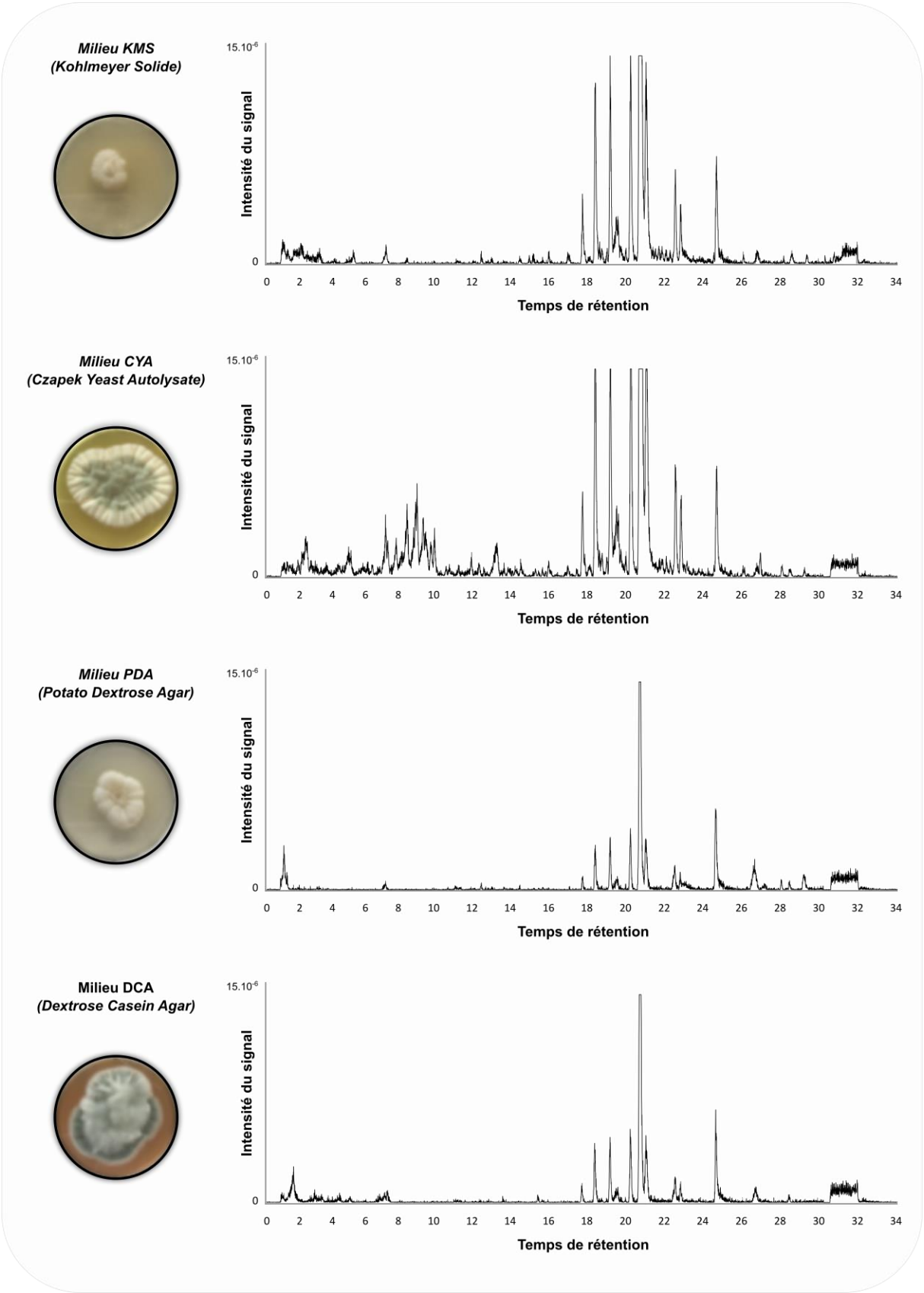


Figure 0-2 : Profils chimiques acquis par chromatographie liquide haute performance couplée à de la spectrométrie de masse haute résolution en mode d'ionisation positif d'extrait de Penicillium rubens Wisconsin 54-1255 obtenus après culture en milieu solide. *Les chromatogrammes sont représentés sous forme de pics de base. Ces profils permettent d'évaluer l'aptitude de Penicillium rubens Wisconsin 54-1255 à produire des métabolites dans des conditions évaluées données. Ici, la condition CYA permet l'obtention de composés spécifiques élués entre 6 et 10 min.*

Dans le cadre de la biologie des systèmes, l'étude du métabolisme d'un organisme peut s'effectuer à travers la modélisation d'un réseau métabolique à l'échelle du génome (**GSMN**). Ce dernier représente un ensemble de réactions soutenues ou non par des données génomiques, reliant divers métabolites et définissant les propriétés biochimiques et physiologiques de l'organisme, en s'appuyant sur les potentialités exprimées par son génome. Ces modèles visent à intégrer et à visualiser les connaissances hétérogènes liées au métabolisme d'un organisme *(Pan et Reed 2018)*. Ainsi, les principes de biologie intégrative, appliquée à la génomique et à la métabolomique, permettent de lier les données de haute qualité obtenues par les technologies analytiques aux modèles métaboliques. La reconstruction de tels modèles repose sur des méthodes mathématiques et sur des données à la fois théoriques, obtenues *in silico*, et expérimentales. Les modèles ainsi proposés fournissent une nouvelle vision des réalités biologiques. À l'échelle d'un organisme, ces modèles offrent une nouvelle perspective sur les systèmes biologiques, permettant de confirmer ou d'infirmer des hypothèses sur les processus biologiques, de prédire des propriétés difficiles à observer expérimentalement, ou encore d'identifier des lacunes dans le réseau métabolique, révélant des limites computationnelles ou des caractéristiques spécifiques à l'espèce étudiée. En résumé, en reliant les observations phénotypiques au génotype, un **GSMN** fournit une plateforme de ressources organisées, reflétant les capacités métaboliques optimales d'un organisme cible au moment de sa reconstruction.

Perturber ces modèles pour accélérer ou freiner la production de composés d'intérêt est une question fondamentale dans l'utilisation des **GSMNs**, souvent motivée par des objectifs industriels dans les domaines pharmaceutiques ou écologiques. Traditionnellement, la plupart des méthodes utilisées à cet effet reposent sur des approches de « *Strain design* », visant à développer de nouvelles souches de micro-organismes aux capacités spécifiques. Ces méthodes, illustrées en **Figure 0-3** sont concentrées essentiellement sur l'analyse des réseaux métaboliques et plus particulièrement sur la modification de réactions internes ; des réactions clés sont ciblées pour être inactivées afin d'éviter les flux indésirables ou activées pour encourager la production d'un métabolite cible.

Ainsi, les stratégies courantes incluent la simulation de modifications génétiques, telles que la suppression de gènes (*e.g. knock-out*) ou l'ajout de gènes introduisant de nouvelles réactions métaboliques (*i.e. knock-in*). En pratique, ces modifications sont souvent testées de manière binaire, et les perturbations internes simulées par ces méthodes s'appuient fréquemment sur des données biologiques comme la transcriptomique qui mesure l'expression des gènes. Ces interventions, qui perturbent le réseau métabolique en modifiant la distribution des flux métaboliques, permettent ainsi de prédire les réponses cellulaires à différentes manipulations génétiques, offrant ainsi une évaluation des conséquences sur l'ensemble du réseau métabolique. L'objectif ultime est d'optimiser ces flux pour rediriger les ressources vers la production de métabolites d'intérêt tout en minimisant la formation de sous-produits indésirables. Ces méthodes se révèlent efficaces pour rationaliser les approches d'ingénierie génétique des micro-organismes.

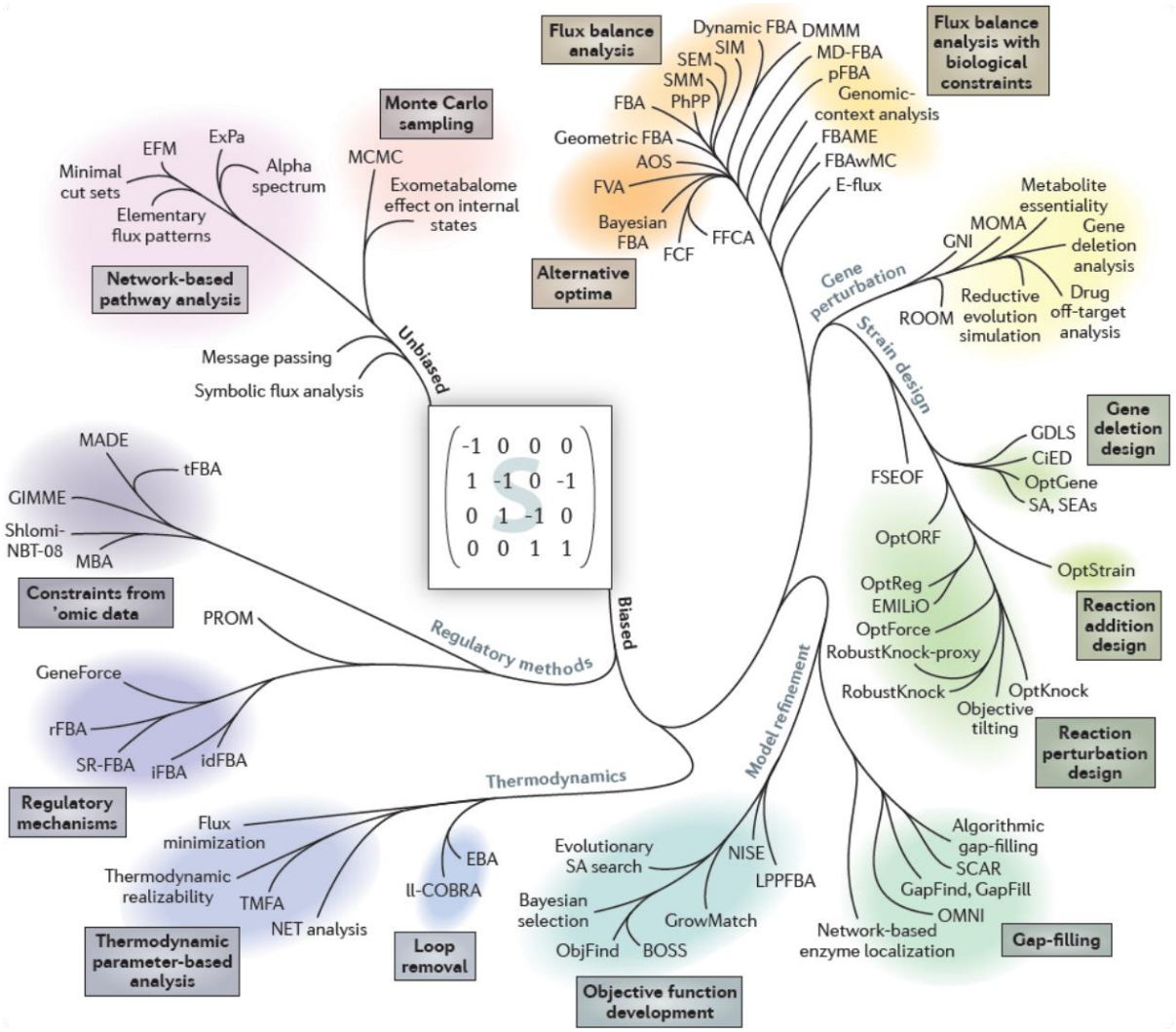


Figure 0-3 : « Phylogénie » des méthodes de modélisation basées sur les contraintes développées au cours de la première décennie des années 2000. *Le répertoire d'outils dédiés à la reconstruction et à l'analyse basées sur les contraintes (COBRA) s'est considérablement élargi, atteignant plus de 100 méthodes en 2012. Ces diverses méthodes exploitent la flexibilité et l'extensibilité des modèles COBRA pour la prédiction et l'analyse, toutes fondées sur l'examen de la structure du réseau métabolique (i.e. matrice stœchiométrique). Les auteurs ont utilisé cet arbre phylogénétique pour illustrer les similitudes entre les applications et les utilisations de ces méthodes, mettant en lumière les algorithmes partagés par plusieurs d'entre elles. Extrait des travaux de (Lewis et al. 2012).*

Les travaux que nous présentons dans ce manuscrit s'intéressent à un aspect différent : **l'impact des variations des conditions environnementales sur les réponses du modèle métabolique**, une approche appelée « *Media design* » qui, à notre connaissance, demeure encore sous-explorée. Notre objectif est donc de perturber le modèle, non pas par des modifications internes, mais en agissant uniquement sur les conditions externes pour simuler différentes situations de culture. Cette approche, qui par définition peut être qualifiée d'**OSMAC *in silico*** (*i.e.* en référence à l'approche « *One Strain Many Compounds* »), se résume par la question suivante : **quelles contraintes ou combinaisons de contraintes sur les imports de nutriments peuvent être appliquées pour activer les voies de biosynthèse d'intérêt chez un organisme modèle de champignons filamenteux : *Penicillium rubens* Wisconsin 54-1255 ?** Nous formulons l'hypothèse que l'utilisation des **GSMNs** constituera une clé essentielle pour de futures stratégies d'optimisation et de rationalisation de la production de produits naturels basées sur les conditions de culture, en reliant les conditions environnementales à l'expression de la chimiodiversité de l'organisme.

Le **Chapitre 1** commence par un examen des produits naturels, de leur diversité et de leurs applications. Il explore ensuite la modélisation du métabolisme à l'échelle du génome, mettant en lumière son rôle fondamental dans l'identification et la production de ces composés. Sous forme d'ébauche de revue de la littérature, ce chapitre détaille les processus de reconstruction des réseaux métaboliques et leur transformation en modèles fonctionnels, avant d'explorer leur application dans les stratégies de *Strain design* et de *Media design* pour maximiser la production de métabolites. Enfin, il introduit notre modèle d'étude *Penicillium rubens* Wisconsin 54-1255.

Le **Chapitre 2**, structuré comme un appui méthodologique pour les principales ressources utilisées pendant nos travaux, fournit un glossaire détaillé des outils et des bases de données utilisés, facilitant ainsi la compréhension des concepts abordés dans le manuscrit. Cette section aide à contextualiser les méthodologies employées, en expliquant les raisons de leur sélection, leurs avantages ainsi que leurs limites respectives.

Le **Chapitre 3** est concentré sur le cœur de nos travaux, portant sur la reconstruction d'un réseau métabolique à l'échelle du génome de haute qualité pour *Penicillium rubens* Wisconsin 54-1255, désigné sous le nom d'***i*Prub22**. Ce chapitre décrit en détails les différentes étapes nécessaires à la reconstruction, y compris les méthodes utilisées pour élaborer un modèle fonctionnel. Nous discuterons les améliorations apportées à la reconstruction, les défis rencontrés, notamment ceux liés au *gap-filling*, à la modélisation des échanges de l"organisme avec l'environnement, ainsi que les éléments permettant de qualifier ***i*Prub22** de modèle de haute qualité. En outre, une attention particulière sera portée sur le métabolisme spécialisé, avec une exploration des voies de biosynthèse et des facteurs influençant la production de ces molécules. À travers ce chapitre, notre objectif est de fournir une compréhension approfondie des choix méthodologiques et des considérations sous-jacentes qui ont conduit à cette reconstruction, tout en soulignant l'importance d'une ressource accessible et interopérable pour les futures études.

Enfin, le **Chapitre 4** explore les perspectives futures d'évolution et d'utilisation d'***i*Prub22**. Nous examinerons diverses pistes d'amélioration pour des reconstructions à venir, notamment l'intégration automatique de la modélisation intracellulaire. Le chapitre se conclura par une réflexion sur l'application du modèle à une souche marine de *Penicillium rubens* et sur la pérennité des approches méthodologiques adoptées, mettant en évidence les opportunités offertes par ***i*Prub22** pour progresser dans la recherche sur la production des métabolites spécialisés.

CHAPITRE 1 :

Étude Bibliographique

1 - Les produits naturels : origines, diversité et applications

Les produits naturels sont des molécules organiques produites, comme leur nom l'indique par des organismes vivants. Ces éléments sont une source majeure de molécules biologiquement actives *(Newman et Cragg 2020)* et pour cette raison, ils ont été exploités de tout temps, de l'antiquité à l'époque moderne, pour leurs diverses propriétés *(Demain et Fang 2000)*. À ce titre, ils ont joué un rôle central dans le développement des sociétés humaines, tant sur le plan médical qu'industriel et culturel *(Dias et al. 2012)*.

Définies historiquement et subjectivement par leur petite taille (*i.e.* masse moléculaire inférieure à 1500 daltons), ces molécules se subdivisent en deux classes selon le rôle qu'elles occupent au sein d'un organisme *(Walsh et Tang 2017)*. Nous distinguons ainsi les métabolites constitutifs qui assurent la croissance et le développement de l'organisme, des métabolites spécialisés qui confèrent aux organismes un « succès reproductif » ou un « taux de survie » accrus en leur octroyant, par exemple, une meilleure résistance aux stress biotiques ou abiotiques à travers le développement de mécanismes de défense. Généralement, en raison de leur spécificité – métabolites non ubiquitaires exprimés sous des conditions spécifiques avec une gamme d'organismes producteurs qui englobe les trois règnes du vivant – ce sont ces derniers qui sont considérés dans le domaine de la chimie des produits naturels.

Depuis la première décennie du XIX^{eme} siècle l'étude des produits naturels a évolué avec l'émergence de la chimie moderne. Le premier isolement d'un produit naturel est attribué au pharmacien allemand Friedrich Wilhelm Adam Sertürner vers les années 1805 avec la purification de la morphine à partir de l'opium *(Huxtable et Schwarz 2001)*. Dès lors, les principes actifs ont pu être isolés et caractérisés de manière systématique et ces découvertes ont jeté les bases de la pharmacognosie moderne.

À l'image de l'identification du premier antibiotique en 1929 par Alexander Fleming, la découverte et la caractérisation des produits naturels ont souvent été issues de criblages aléatoires d'extraits végétaux, fongiques, bactériens, *etc.* *(Li et al. 2009)*. Cependant, depuis le début des années 2000, la découverte de nouveaux produits naturels se raréfie et les méthodes de bioprospection, longues, inefficaces et coûteuses ont alors été délaissées *(Scheffler et al. 2013; David et al. 2015)*. Corrélé à la résistance de plus en plus accrue aux antibiotiques *(Gould 2009)*, il est essentiel de relancer ces recherches par des approches plus rationnelles et prédictives. Ce besoin est d'autant plus urgent que le rejet de déchets pharmaceutiques dans l'environnement *(Patel et al. 2019)* contribue à l'essor de la résistance aux antimicrobiens, un des principaux enjeux de santé publique actuels *(Prestinaci et al. 2015; World Health Organization 2023)*. L'émergence de la génomique couplée à la métabolomique permet désormais d'associer les métabolites avec leurs données structurales et leurs gènes biosynthétiques. Associées à l'automatisation et à l'évolution qualitative des outils observées ces dernières années *(Wolfender et al. 2019)*, ces approches devraient jouer un rôle de plus en plus central dans la découverte de nouveaux produits naturels.

À ce jour, la base de données *Dictionary of Natural Products*® est l'un des inventaires les plus exhaustifs qui retrace et recense les propriétés ainsi que l'historique complet de la littérature concernant ces composés. Cette ressource aide les chimistes à explorer et à développer des solutions basées sur des produits naturels pour diverses industries, notamment pharmaceutiques *(Rodrigues et al. 2016)*. À la fin des années 2010, plus de 300 000 composés y sont recensés, accompagnés de données descriptives et numériques sur leurs propriétés chimiques, physiques et également biologiques. Principalement utilisée dans les domaines de la chimie, de l'agro-alimentaire, de la cosmétique, de la pharmacologie et de la biotechnologie, cette base de données de référence, continuellement mise à jour, constitue ainsi l'une des sources d'information les plus complètes sur les produits naturels *(Chassagne et al. 2019)*, témoignant de la vaste diversité chimique connue. Néanmoins, la chimiodiversité encore à explorer est immense et la richesse des organismes dans la nature offre de nombreuses opportunités pour la bioprospection et la découverte de nouveaux composés thérapeutiques.

En effet, les produits naturels suscitent un grand intérêt en pharmacologie en raison de leur capacité à fournir des molécules biologiquement actives essentielles à la formulation de nombreux médicaments *(Schueffler et Anke 2014; Newman et Cragg 2020)*. Ces composés naturels présentent un vaste éventail d'activités pharmacologiques, incluant des propriétés antibiotiques, antitumorales, toxiques, analgésiques ou immunosuppressives *(Walsh et Tang 2017)*. Il est donc peu surprenant de constater que l'espace chimique occupé par les médicaments est largement similaire à celui des produits naturels *(Feher et Schmidt 2003)* et que ces derniers représentent près de la moitié des nouveaux médicaments approuvés chaque année (**Figure 1-1**). En outre, il est estimé qu'environ 50 % des produits naturels caractérisés ne possèdent pas d'équivalents synthétiques *(Rodrigues et al. 2016)*, renforçant ainsi l'idée selon laquelle les produits naturels constituent une source irremplaçable et essentielle pour la découverte de nouvelles molécules bioactives et médicales.

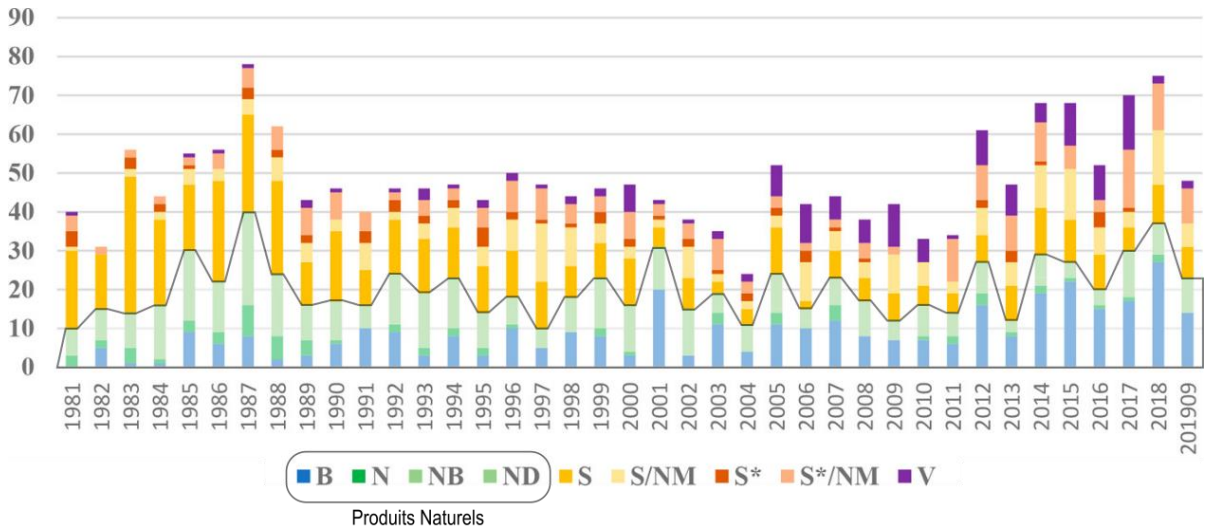


Figure 1-1 : Nouveaux médicaments annuels approuvés selon leur source depuis le début des années 1980. *Les travaux de (Newman et Cragg 2020), dont est extraite cette figure, couvrent la période de 1981 à septembre 2019 et représentent 1 881 nouveaux médicaments. Les produits naturels au sens strict correspondent aux quatre premières catégories suivantes : B : macromolécule biologique ; N : produit naturel inaltéré ; NB : médicament à base de plantes (mélange défini) ; ND : dérivé de produit naturel. Les autres catégories sont : S : médicament de synthèse ; S/NM : médicament de synthèse imitant un produit naturel ; S* : médicament synthétique (pharmacophore des produits naturels) ; S*/NM : médicament synthétique (pharmacophore des produits naturels imitant un produit naturel) ; V : vaccin.*

En outre, bien que les produits naturels se distinguent par une grande diversité structurelle, souvent liée à des activités biologiques spécifiques, les étapes menant à leur biosynthèse sont relativement simples et peu nombreuses *(Demain et Fang 2000)*. Notons également que les voies de biosynthèse des métabolites spécialisés produisent généralement plusieurs produits et non un seul, suggérant alors que ces voies ont évolué en favorisant la diversité moléculaire *(Fischbach et Clardy 2007)*, sans compter que dans certains cas, une certaine souplesse à recruter des précurseurs variés est observée *(D'Ari et Casadesús 1998)*. Les voies biosynthétiques des produits naturels comprennent un nombre restreint de réactions qui se détachent du métabolisme primaire en un nombre limité de points, à partir d'une dizaine de précurseurs communs (*e.g.* composés non azotés tels que l'acétyl-CoA, l'acide shikimique, l'acide mévalonique, le méthylérythritol phosphate, ou des acides aminés tels que la phénylalanine, la tyrosine, le tryptophane, l'ornithine ou la lysine). La structure chimique des produits naturels résulte ainsi de voies biosynthétiques spécifiques à un organisme source, et ces structures ont souvent servi de modèles pour la synthèse de nouvelles molécules en chimie médicinale et en pharmacologie *(Rodrigues et al. 2016)*. En conséquence, l'étude des produits naturels et de leurs structures joue un rôle crucial dans le processus de découverte et de développement de médicaments.

La comparaison des mécanismes génétiques et évolutifs (*e.g.* spéciation, duplication, transfert de gènes horizontaux) qui sous-tendent l'évolution de la chimiodiversité eucaryote et procaryote démontre l'existence de similarités structurales et organisationnelles entre ces deux domaines *(Chevrette et al. 2020; Rokas et al. 2020)*. Plus précisément, une forte proportion de métabolites spécialisés sont produits par des enzymes dont les gènes sont co-localisés et co-régulés *(Romano et al. 2018)*. Ces regroupements de gènes peuvent représenter plus de la moitié du génome des bactéries et fonctionne sous forme d'opérons où ils sont traduits simultanément. Des regroupements similaires, dans une proportion moindre, sont également observés chez les levures, les champignons filamenteux et les plantes, mais les gènes y sont transcrits indépendamment *(Nützmann et al. 2018)*. L'histoire évolutive de ces ensembles, dénommés clusters de gènes biosynthétiques (*i.e. biosyntetic gene cluster*, BGC), associés à la nature des gènes qui les composent, gènes de résistance, de transport ou de régulation par exemple *(Keller 2019)*, sont des indices potentiels de leurs activités biologiques qui permettraient de guider les recherches d'ingénieries moléculaires *(Eustáquio et Ziemert 2018)*.

Cependant, les différences observées entre les modèles théoriques *in silico* et les cultures *in vitro* indiquent que les clusters de gènes non exprimés en laboratoire sont nombreux (**Figure 1-2**) et forment une ressource inexploitée de production de produits naturels, notamment chez les champignons filamenteux *(Nielsen et al. 2017; Keller 2019)*. Deux hypothèses sont alors généralement avancées pour expliquer ce phénomène : soit la biosynthèse des produits naturels est finement régulée et activée uniquement en réponse à des conditions de cultures spécifiques (*i.e.* la production de métabolites spécialisés est un processus coûteux pour un organisme), soit il existe effectivement une production, mais sous les seuils de détection analytiques *(Romano et al. 2018)*. En outre, l'un des problèmes majeurs rencontrés par le chimiste des produits naturels est la forte tendance à la mise en évidence systématique de molécules déjà connues *(Eustáquio et Ziemert 2018)*. L'intérêt est alors de comprendre quelles sont les conditions sous lesquelles les clusters de gènes, dits silencieux, orphelins ou cryptiques, pourraient être activés.

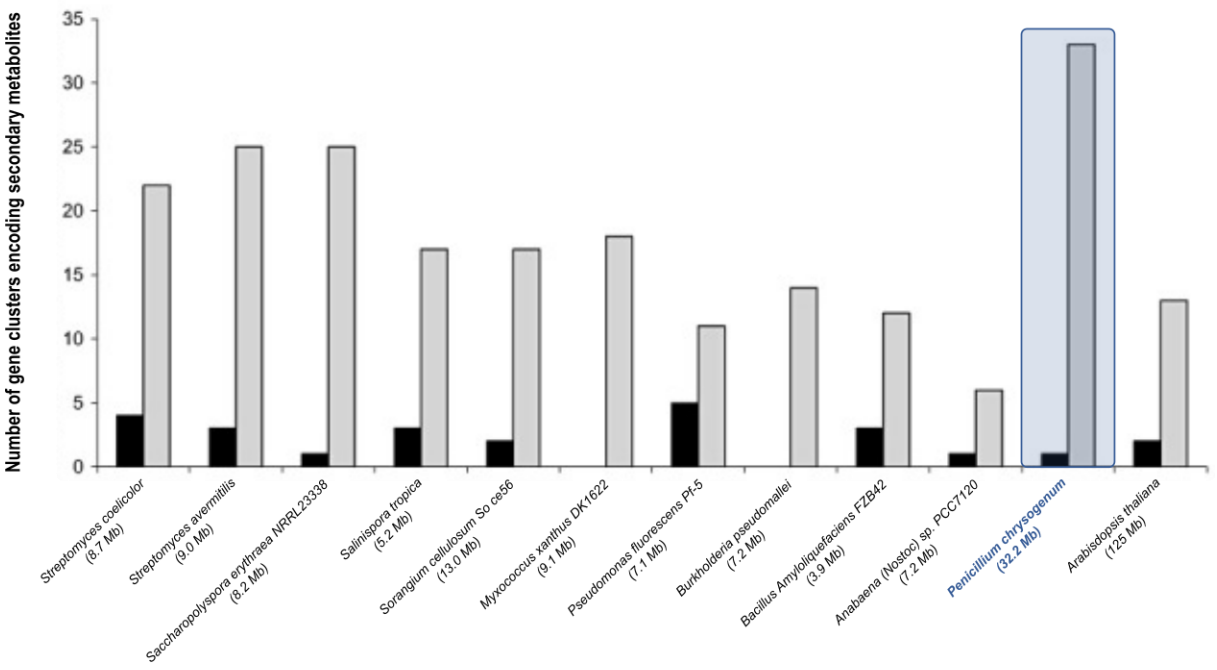


Figure 1-2 : Nombre de clusters de gènes reliés à la production de métabolites spécialisés de douze organismes dont les génomes ont été entièrement séquencés. *Les barres noires représentent les clusters de gènes connus pour être associés à des métabolites isolés, tandis que les barres grises indiquent le nombre de clusters de gènes prédits sur la base des données de séquençage du génome. Au début des années 2010, il était estimé qu'en moyenne, seulement 4 % des clusters de gènes d'un organisme étaient liés à un ou des produits isolés en laboratoire. Extrait des travaux de (Gross 2009; Pettit 2011).*

À l'instar de ce qui a été effectué pour des produits naturels bactériens *(Zarins-Tutt et al. 2015)*, le décodage du génome est une étape primordiale dans les processus de compréhension de déblocage de ces gènes cryptiques pour relancer la découverte de nouveaux antibiotiques *(Scheffler et al. 2013)*. Les avancées continuelles en ressources informatiques, de plus en plus performantes, permettent de mieux exploiter les données génomiques et chimiques, ouvrant la voie à l'exploration des génomes (*i.e. genome mining*) pour activer et caractériser les clusters de gènes inexplorés, offrant désormais des outils de plus en plus puissants pour traiter les volumes croissants de données génomiques et chimiques *(Challis 2008; van der Lee et Medema 2016; Kim et al. 2017)*.

Indépendamment des connaissances liées à la structure du génome, la recherche de métabolites spécialisés peut également s'envisager en explorant empiriquement les différences d'expression des composés qui résultent de la variabilité des conditions de culture des micro-organismes *(Pan et al. 2019)*. Cette approche, connue sous l'acronyme **OSMAC** pour « *One Strain-Many Compounds* » *(Bode et al. 2002)*, propose ainsi d'explorer la diversité naturelle d'un organisme en agissant sur quatre leviers principaux (**Figure 1-3**). Ces derniers incluent **(1)** la modification du régime nutritionnel, en ajustant, par exemple, les sources et les apports de carbone, d'azote, de phosphore, de soufre ou d'éléments traces, **(2)** la variation des paramètres physiques de culture, tels que la température, la salinité, le pH ou les supports de culture, **(3)** la co-culture pour favoriser les interactions entre différentes espèces (*i.e.* mise en évidence des phénomènes de compétition ou de symbiose) ainsi que **(4)** l'ajout d'éliciteurs aux cultures (*i.e.* molécules extrinsèques induisant un stress à l'origine de réponses défensives). Ces stratégies permettraient ainsi de révéler la production de composés encore non observés *(Romano et al. 2018; Arora et al. 2020)*.

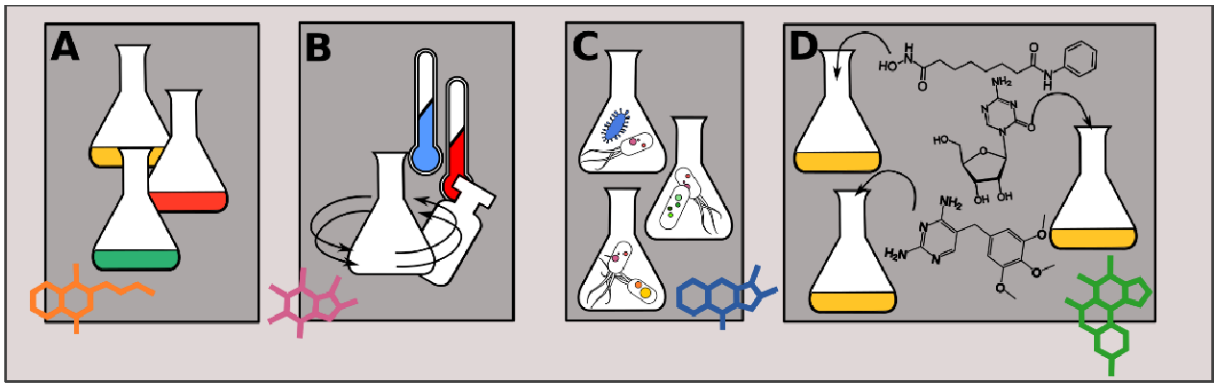


Figure 1-3 : Susciter la production de métabolites spécialisés cryptiques via une approche « One Strain Many Compounds » (OSMAC). *Un processus empirique effectué généralement en variant : les régimes nutritionnels* ***(A)****, en modifiant les paramètres physiques* ***(B)****, en réalisant des co-cultures* ***(C)*** *ou en ajoutant des éliciteurs chimiques* ***(D)****. Figure et légende issues des travaux de (Romano et al. 2018).*

Cependant, aussi séduisante que puisse être l'idée de maximisation de l'expression du potentiel de chimiodiversité à travers les approches **OSMAC**, il n'en demeure pas moins qu'elles s'inscrivent dans des processus chronophages et coûteux. En effet, tester *in vitro* l'ensemble des possibilités et modifications offertes par cette approche démultiplie les expériences, et même si les applications effectuées essentiellement sur des bactéries ou des champignons ont été concluantes *(Pan et al. 2019)*, l'aspect aléatoire des résultats limite de surcroît le développement de processus commun pour l'ensemble des micro-organismes *(Bode et al. 2002)*. Le recours préalable aux outils de la biologie des systèmes et plus précisément à la modélisation du métabolisme serait donc une solution pour restreindre le nombre de combinaisons à tester et sélectionner les conditions de culture jugées les plus pertinentes.

2 - La modélisation du métabolisme au service de la détection, de la caractérisation et de la production de produits naturels

Pour aborder cette thématique, cette section, rédigée en anglais, présente une ébauche de review portant sur l'utilisation des **GSMNs** pour l'accès rationalisé aux produits naturels. L'objectif est de montrer comment les **GSMNs** sont actuellement employés dans la mise en évidence de ces produits. Cette review est destinée à un journal dont le public cible est peu familier avec ces réseaux.

Notre réflexion s'articule autour de quatre axes principaux: **(1)** une introduction aux **GSMNs**, **(2)** leur utilisation dans le *Strain design*, **(3)** leur application au *Media design*, et **(4)** une discussion sur les limitations actuelles dans le domaine de la chimie des produits naturels. Cette dernière section, centrale aux travaux du manuscrit, ne sera cependant pas présentée dans ce Chapitre.

Genome Scale Metabolic Network, A Model Of Metabolism Guiding The Discovery Of Natural Products?

2.1 INTRODUCTION

Natural products (NPs) constitute a significant reservoir of novel bioactive compounds, offering substantial potential for drug discovery and development *(Newman et Cragg 2020)*. The sheer diversity of organisms in nature provides ample opportunities for bioprospecting *(Zhu et al. 2011)* and then identifying previously uncharacterised compounds with therapeutic potential *(Geers et al. 2022)*. The inherent ability of nature to generate unique and structurally complex molecular frameworks continues to serve as an unmatched source of inspiration for new drug design and synthesis *(Stratton et al. 2015; Stone et al. 2022)*. Consequently, the NPs exploration remains a pivotal strategy in advancing contemporary drug discovery efforts.

Recent advances in OMIC technologies have provided new insights into the vast chemical diversity of NPs, underscoring the significant gap in our understanding of natural bioactive metabolite richness. The rapid expansion and widespread application of genome sequencing have uncovered numerous biosynthetic gene clusters, many of which remain unexplored *(Gavriilidou et al. 2022; Caesar et al. 2023; Yee et al. 2023)* despite extensive curation and annotation efforts *(Terlouw et al. 2023)*. This emphasis that many NP gene clusters have yet to be linked to any produced compounds, underlining the need for a continuous effort toward this direction *(Guzmán-Chávez et al. 2018)*.

Similarly, metabolomics has revealed a substantial lack of structural annotation for detected chemical signals *(Wolfender et al. 2019; Allard et al. 2023)* despite significant efforts to gather and organise this data *(Rutz et al. 2022; van Santen et al. 2022)*. In particular, the time-consuming steps required for the complete chemical characterisation of NPs *(Berlinck et al. 2019)* have hindered the growth of NP databases, limiting our ability to exploit their full potential. Nonetheless, both genomics and metabolomics approaches have demonstrated the incredible and expected chemical diversity from natural sources, emphasising the need for continued investment and innovation in this field.

Focusing on the microbial NP discovery process, one of the reasons for such observation is the presence of cryptic biosynthetic pathways *(El-Hawary et al. 2023)*, explained by two factors: **(1)** under traditional laboratory conditions, either their genes are not expressed, or **(2)** their products fall below analytical detection thresholds. Thus, various approaches have been developed to address these limitations, broadly considered as culture-based and non-culture-based strategies. Non-culture-based approaches, often rooted in molecular biology, involve engineering organisms to produce target molecules *(Volk et al. 2023)*, with techniques such as heterologous expression being widely explored *(Chiang et al. 2023; Woodcraft et al. 2023)*. However, despite numerous examples in the literature, these methods have typically yielded only a few novel compounds at once after substantial and valuable effort. On the other hand, several culture-based strategies have also been exploited to enhance chemical diversity through methods like the **OSMAC** (One Strain Many Compounds) approach *(Bode et al. 2002; Romano et al. 2018)*, epigenetic modifications *(Bind et al. 2022)* or co-culture *(Arora et al. 2020)*. Over the past decade, these approaches have increasingly been coupled with chemical profiling to identify newly formed compounds and facilitate their annotation *(Wolfender et al. 2019)*. While effective in isolating previously unreported molecules *(Bertrand et al. 2013; Le et al. 2021)* and despite the significant effort invested in screening for chemical novelty before compound isolation *(Pham et al. 2021; Quiros-Guerrero et al. 2022)*, the results are often viewed as insufficient. Furthermore, culture-based approaches tend to be considered random and heavily reliant on exhaustive chemical profiling *(Romano et al. 2018; Allard et al. 2023)*.

For ages, natural product scientists have been engaged in discovering new molecules *(Dias et al. 2012)*. However, more recently, the focus has shifted to understanding their biosynthesis and discovering their associated gene clusters. Significant efforts are currently directed towards deciphering these gene clusters and annotating the function of each enzyme *(Terlouw et al. 2023)* to unravel their chemical logic. Despite advancements, a substantial gap remains between the number of reported compounds and the understanding of their specific biosynthesis. In contrast, core metabolic pathways are nowadays well-documented in

databases such as ModelSEED *(Seaver et al. 2021)*, KEGG *(Kanehisa et al. 2022)*, and BioCyc *(Karp et al. 2019)*. Recently, these extensive metabolic maps have been employed to interpret metabolomics data *(Amara et al. 2022)*. However, such systems biology approaches are still rarely applied in the field of natural products research.

In Systems Biology, Genome-Scale Metabolic Networks (**GSMNs**) provide a detailed representation of an organism's biochemical and physiological properties. *(Klipp et al. 2016)*.These models are constructed through computational approaches that integrate both *in silico* and experimental data. With increasing model complexity, the heterogeneous knowledge of metabolism can be efficiently integrated and visualised *(Pan et Reed 2018)*, thereby deepening our understanding of metabolism at the system level *(Hattwell et al. 2020)*. This rise of integrative biology, thus, has facilitated the linkage of high-quality analytical data (metabolomics, transcriptomics, proteomics) to these models. **GSMNs** serve as resource platforms that summarise the metabolic capabilities of the target organism at the time of reconstruction, essentially offering a “knowledge snapshot”. However, widespread adoption by the scientific community is crucial for their sustainability, evolution, and application, as exemplified by the challenges in reconstructing the *Caenorhabditis elegans* metabolic network *(Witting et al. 2018)*. To date, hundreds of manual and automatic reconstructions are available for model and non-model organisms, spanning all three domains of life: prokaryotes, archaea, and eukaryotes *(Gu et al. 2019)*.

Considering the systems biology approach, once an accurate **GSMN** reconstruction is achieved *(Thiele et Palsson 2010)*, appropriate data analysis strategies should be able to explain the observed production of specific NPs. It creates a unique opportunity to deepen our understanding of the mechanisms governing NP biosynthetic pathways. Consequently, the potential for optimising NP production is evident. However, the applications of this approach extend beyond this aspect, and systems biology could be a powerful and reasonable tool to anticipate synthetic biology strategies *(Chen et al. 2020; Volk et al. 2023)*. Unfortunately, using systems biology to access chemical novelty remains widely unexplored and poorly documented.

After detailing the reconstruction and analysis of **GSMNs**, this review explores their possible applications in the microbial NP discovery process, from the design of optimised strain to the selection of appropriate culture conditions. Finally, the review addresses current challenges and bottlenecks in applying systems biology to the large-scale discovery of specialised metabolites, highlighting future perspectives in this emerging field of research.

2.2 WHAT IS A GENOME-SCALE METABOLIC NETWORK?

2.2.1 Evolution and Applications of GSMNs

In systems biology, metabolic network modelling involves, from a holistic point of view, considering all the knowledge about the biosynthetic pathways of an organism *(Klipp et al. 2016)*. Metabolic pathways are described as a series of enzyme-catalysed reactions that result in the synthesis of various compounds. Therefore, **GSMN** could be considered as a knowledge platform that defines biochemical and physiological properties within a cell. Moreover, it represents a biological system in a format that displays usable information to guide experimentation. The construction of such models is based on a set of mathematical methods and both theoretical, obtained *in silico*, and experimental data. Thus, these models provide a new vision of biological realities that helps interpret chemical and/or biological observations. For instance, at the level of an organism, the use of **GSMN** allows, among other things: to investigate relationships between phenotypic observations and genotype *(McCloskey et al. 2013)*; to predict properties of the system that would be difficult to observe experimentally; to validate or refute hypotheses regarding the dynamics of biological processes, for instance, in response to environmental perturbations or modifications; to guide the identification of potential antimicrobial drug targets by pinpointing vulnerabilities in metabolic pathways or highlighting essential genes and proteins involved in pathogen growth *(Chung et al. 2020)*; to provide a robust approach for predicting gene-to-phenotype linkages, thereby enhancing our understanding of human health by identifying relationships such as drug side effect associations *(Shaked et al. 2016)*; to raise new questions by applying the concept of “dead ends” (*i.e.,* metabolites that are neither produced nor consumed based on network topology)

(Mackie et al. 2013; Pan et Reed 2018); to study complex microbe-microbe and host-microbe interactions *(Heinken et Thiele 2015)*, revealing symbiosis, competition, and their effects on host health and disease; to give insights into evolutionary concepts by tracing genetic adaptations, functional divergence, and conserved pathways across species.

As a preamble, we would like to draw attention to the terminology of reconstruction and model. Although the terms are often interchanged, it is accepted in the community that a reconstruction expresses the potentialities expressed in the genome, and a model is a parameterisation (*i.e.,* a specification) of this reconstruction. Historically, the GEM abbreviation was used to refer to a genome-scale metabolic network model. Here, we will specifically prefer to use the acronym **GSMN** reconstruction or **GSMN** model only when the distinction is necessary; otherwise, the simple abbreviation **GSMN** will be employed.

Historically, the first publication of a model was assigned to the bacterium *Haemophilus influenzae* in 1999 *(Edwards et Palsson 1999)*. Thus, over the last 25 years, numerous reconstructions have been produced, starting essentially with reconstructions related to prokaryotic, then eukaryotic organisms and even archaea *(Gu et al. 2019)*. At present, and to our knowledge, there is no exhaustive reference to reconstructions and models generated by the scientific community. For example, the Palsson lab website (https://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms) lists 113, 57 and eight reconstructions for prokaryotic, eukaryotic and archaeal organisms, respectively (last update in February 2018). As of November 2022, BiGG *(Norsigian et al. 2020)*, a historical reconstruction database, offers 108 models, while BioCyc *(Karp et al. 2019)* reports 19,441, 470, and 39 databases for prokaryotes, eukaryotes, and archaea, respectively. BioModels *(Malik-Sheriff et al. 2020)*, which is a repository of mathematical models of biological and biomedical systems, offers 2,830 models distributed mainly in **SBML** format and 37% of which are manually curated. Finally, it is also possible to find reconstructions on the specific pages of tools that have generated these models or of teams working on these themes. To illustrate this point, we can cite the GitHub of Systems and Synthetic Biology at Chalmers University of

Technology, where **GSMNs** of eukaryotic models are available, such as *Homo sapiens* (Human-GEM) *(Robinson et al. 2020)*, *Danio rerio* (Zebrafish-GEM) *(Wang et al. 2021)*, *Caenorhabditis elegans* (Worm-GEM) *(Wang et al. 2021)*, *Drosophila melanogaster* (Fruitfly-GEM) *(Wang et al. 2021)*, *Rattus norvegicus* (Rat-GEM) *(Wang et al. 2021)*, *Mus musculus* (Mouse-GEM) *(Wang et al. 2021)*, *Saccharomyces cerevisiae* (yeast-GEM) *(Zhang et al. 2024)* or the **AuReMe** *(Aite et al. 2018)* website which hosts 11 reconstructions for single organisms and two multispecies networks.

The **GSMN** reconstruction relies on different concepts and, nowadays, consists of well-established protocols which are regularly improved *(Thiele et Palsson 2010)*. Still, all reconstructions are based on genomic information, as further discussed. Detailed procedures and tools for **GSMN** reconstruction are far beyond the scope of the review, and in-depth information is available in Thiele's methodology.

Since this review focuses on the potential and future applications of **GSMNs** in the natural product discovery field, a description of the reconstruction process and available analytical methods is necessary and provided below. First, a **GSMN** reconstruction is divided into two major parts: **(1)** the establishment of a first draft **GSMN** reconstruction obtained from the genomic sequences of the organism under study and **(2)** the refinement of this draft by an extensive manual curation to obtain a high-quality **GSMN** model. Subsequently, in addition to reconstruction, different bioinformatics network analyses can be employed based on the specific scientific question, which will be addressed at the end of this section. Further details on these elements can be found in **Figure 1-4**.

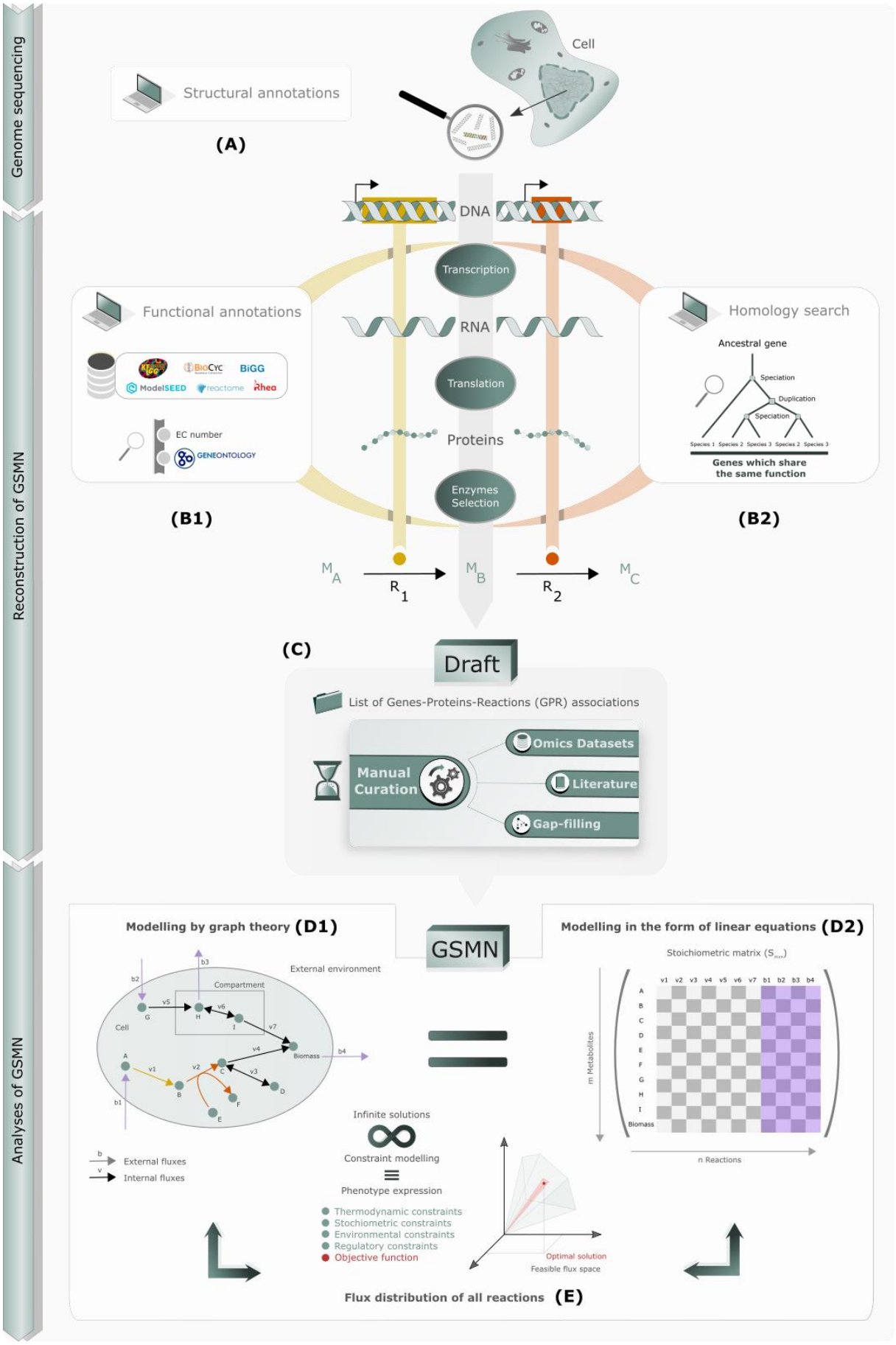

Genome sequencing
Reconstruction of GSMN
Analyses of GSMN
Structural annotations
(A)
Cell
DNA
Transcription
RNA
Translation
Proteins
Enzymes
Selection
Functional annotations
BiGG
ModelSEED
reactome
Rhea
EC number
GENEONTOLOGY
(B1)
Homology search
Ancestral gene
Speciation
Duplication
Speciation
Species 1 Species 2 Species 3 Species 2 Species 3
Genes which share the same function
(B2)
M A
R 1
M B
R 2
M C
(C)
Draft
List of Genes-Proteins-Reactions (GPR) associations
Manual Curation
Omics Datasets
Literature
Gap-filling
GSMN
Modelling by graph theory (D1)
External environment
Compartment
Cell
Biomass
External fluxes
Internal fluxes
Modelling in the form of linear equations (D2)
Stoichiometric matrix (S mxn)
m Metabolites
n Reactions
Infinite solutions
Constraint modelling
Phenotype expression
Thermodynamic constraints
Stochiometric constraints
Environmental constraints
Regulatory constraints
Objective function
Optimal solution
Feasible flux space
Flux distribution of all reactions (E)

Figure 1-4: Modelling an organism's metabolism *via* reconstruction and analysis of its Genome Scale Metabolic Network (GSMN). A metabolic network represents a collection of biochemical reactions that capture the physiological and biochemical functions within a cell. Specifically, a **GSMN** is a reconstructed model derived through a bottom-up approach, using genomic information to map these reactions. **(A)** Genome sequencing of the studied species provides a list of genes following structural annotation. Biochemically, these genes express proteins after transcription and translation, including enzymes that catalyse metabolic reactions. On the bioinformatics level, the reconstruction of a **GSMN** first requires establishing simplifying assumptions, as explained in **(B1)** and **(B2)**. **(B1)** Genome functional annotations are cross-referenced with various existing databases to search for reported enzymes and their related reactions. For instance, the metabolism-related genes identification is achieved using Gene Ontology annotations and EC numbers (Enzyme Classification System) that connect reactions within a draft network. **(B2)** Another simplifying assumption is based on evolutionary concepts, which propose that homologous genes (*i.e.* those derived from a common ancestor) perform similar functions. By conducting homology searches in existing **GSMN** models, it is possible to infer and integrate these reactions into the draft network. These two independent, complementary, and often automated approaches (**(B1)** and **(B2)**) yield a list of Gene-Protein-Reactions associations that define the initial GMSN draft. **(C)** At this stage, this list is neither entirely accurate nor complete and requires extensive curation, often carried out manually. For example, corrections are frequently needed to account of reaction stoichiometry, charge or mass balance, as well as redundancy in reactions. Additionally, omics data (*e.g.* transcriptomics, proteomics) and insights from the scientific literature can be integrated to refine the model by adding or removing reactions, thereby enhancing its comprehensive. Most of the time, the **GSMN** topology also requires gap-filling strategies to ensure the network's functionality and its completeness. These iterative and time-consuming processes are essential for establishing a consistent and reusable **GSMN**. **(D1 & D2)** One **GSMN**, two analytical frameworks. The final **GSMN** can be visualised as a hypergraph **(D1)** where metabolites are nodes and reactions are edges, or equivalently, as a stoichiometric matrix **(D2)**, with rows representing metabolites and columns corresponding to reactions. The stoichiometric matrix allows for the mathematical modelling of metabolic fluxes. Both representations support complementary network analysis approaches: graph traversal algorithms for the hypergraph and mathematical modelling using first-order linear equations for the stoichiometric matrix. **(E)** To accurately replicate observed phenotypic behaviours and associated reaction fluxes, it is essential to incorporate constraints, such as stoichiometric, thermodynamic, or environmental, into the **GSMN**. This constraint modelling not only enhances the precision of the simulation but also simplifies the mathematical problem by narrowing the flux distribution space. Thus, the ability of the **GSMN** to accurately reproduce phenotypic traits, such as biomass production or the synthesis of specific compounds, reflects the overall quality and effectiveness of the active network.

2.2.2 Reconstruction Process: From Genome to Reconstruction

The reference protocol for obtaining a high-quality **GSMN** was established in 2010 *(Thiele et Palsson 2010)*. Although there are areas for improvement and completion, particularly related to the proliferation of data and advancements in technical capabilities, it is organised into five key areas as follows:

- Creating a draft reconstruction
- Manual reconstruction refinement
- Conversion from reconstruction to mathematical model
- Network evaluation = “Debugging mode”.
- Prospective use

The proliferation of data, associated with the advancement of “omics” technologies *(Pinu et al. 2019)*, has led to numerous **GSMN** generations. The stakes represented by the automation of these approaches are currently no longer preponderant *(Terzer et al. 2009)*, since, on the one hand, it is now possible to establish network models for non-model organisms. On the other hand, these approaches have become increasingly integral to large-scale reconstruction processes. For instance, in recent years, simultaneous reconstructions have been carried out for about twenty-four species of *Penicillium* *(Prigent et al. 2018)* and hundreds of *Salmonella* strains *(Seif et al. 2018)*. Among the best-known and most widely used general reconstruction databases are BiGG *(Norsigian et al. 2020)*, KEGG *(Kanehisa et al. 2022)* and BioCyc *(Karp et al. 2019a)*. Resources such as MetaNetX *(Moretti et al. 2021)* facilitate the compatibility of their identifiers through mapping steps. Additionally, the tools available to carry out all or part of the reconstructions include the RAVEN Toolbox *(Wang et al. 2018)*, the Model SEED pipeline *(Seaver et al. 2021)*, Pathway Tools *(Karp et al. 2019b)*, CarveMe *(Machado et al. 2018)* and AuReMe *(Aite et al. 2018)*. For further details, the performance of some of these tools was benchmarked by Mendoza *et al.* *(2019)*. It is important to note that recent years have seen a growing emphasis on the reproducibility and traceability of data *(Carey et al. 2020)*.

2.2.2.1 Integrating genomic data into comprehensive metabolic models

A **GSMN** is a platform of organised and summarised resources (*i.e.,* a set of reactions supported by genomic data that link various metabolites) aimed at reflecting the metabolic capabilities of an organism at the time of reconstruction. The aggregated genomic information proposed with the following protocol is integrated and encoded within **GSMN** as Gene-Protein-Reaction (GPR) associations, meaning that the reactions are entirely catalysed by enzymes derived from proteins. It is important to note that the GPR associations aim to represent as accurately as possible the biological reality by modelling individual enzymes, isoenzymes or enzyme complexes.

As its name suggests, the **GSMN** reconstruction lies on the organism's genomic information, elucidated through genome decoding. The latter is usually based on structural annotations (**Figure 1-4.A**), which lead to functional annotations (**Figure 1-4.B1**) that help predict candidate metabolic functions. Among the available functional annotations, two are classically fundamental to determining the GPR associations: the EC numbers (Enzyme Commission numbers) and terms from the Gene Ontology (GOT) *(The Gene Ontology Consortium 2021)*. The former provides information on the nature of enzymes and the type of reactions they catalyse, while the latter organises gene information of biological systems into a directed graph format. In parallel, this annotation process can be enriched by including concepts related to evolution (**Figure 1-4.B2**). Indeed, if we consider that two homologous genes, especially orthologs, share the same function *(Fang et al. 2010)*, then by conducting a homology search between the genes or proteins of the species of interest and those of a model organism with an established **GSMN**, it becomes possible to infer the reactions present within these network models to the data of the species under study. Moreover, the use of already well-established models facilitates the propagation of annotations with greater confidence, thereby aiding in the consolidation and acceleration of the reconstruction process *(Notebaart et al. 2006)*.

2.2.2.2 The essential and unavoidable curation

The fusion of all collected data is referred to as a “draft” (**Figure 1-4.C**), but this preliminary version faces significant challenges in accurately representing biological complexity. As a result, various curation steps are essential to refine and evolve this draft into a high-quality **GSMN**. Curating a **GSMN** is a critical process aimed at enhancing the accuracy and functionality of the metabolic network. This step involves thorough examination, correction, and improvement by integrating heterogeneous data from various scientific domains and sources, including biochemistry, genomics, and systems biology, to ensure an accurate representation of metabolic reactions, enzyme activity, and metabolite concentrations. Core steps in this process include validating reactions, correcting errors, and addressing network gaps. Curation also requires an in-depth literature review, experimental validation of metabolic functions, and evaluation of the overall coherence between metabolites and reactions.

Curating a **GSMN** is, therefore, an essential step. However, the term itself is overly generic. This “suitcase word” is a guarantor of the quality of both reconstructions and models, but it encompasses a wide range of tasks, each varying depending on the scope and the scale considered. Is the curation focusing on the reconstruction or the model itself? Is it carried out automatically, semi-automatically, or manually? At which stage of the **GSMN** life cycle does it occur? Are the improvements aimed at biological aspects or technical refinements? These considerations highlight the critical nature of curation. Unfortunately, many of these elements have often been insufficiently or inadequately documented, even fully undocumented, in publications concerning **GSMN** generation. What emerges from the consensus is that curation is a complex, time-consuming and iterative process, yet this process is essential for producing a high-quality model that provides detailed and accurate metabolic insights. As a comprehensive examination of these elements is beyond the scope of this manuscript, we will limit ourselves to sum up the following aspects.

Curation aims to refine the initial draft of the network, tackling the challenges of accurately reflecting biological reality. This refinement typically involves verifying metabolites and reactions for consistency in mass, charge, stoichiometry, and directionality, elements now systematically integrated into most biochemical databases *(Caspi et al. 2020; Norsigian et al. 2020; Kanehisa et al. 2021)*. Next, the curation stages focus specifically on the study of network structure, with the most well-known and documented step being gap-filling through dead-end identification by connectivity or functionality-based approaches *(Orth et Palsson 2010)*.

In addition, it is necessary to also consider the existence of spontaneous reactions and exchange reactions, which are used to model secretion and absorption phenomena and to identify substrates and cofactors *(Thiele et Palsson 2010; Xu et al. 2017; Jeffryes et al. 2021; Sun et al. 2023)*. Moreover, in eukaryotes, it might be relevant to provide compounds and reaction compartmentation to reflect the organisation of the cell. Since some reactions may occur specifically or not in different organelles, such as nuclei, mitochondria, or peroxisomes, this implies adding information on the location of gene products and ensuring the presence of adequate intracellular transport reactions.

Another critical aspect of curation involves the incorporation of metadata to trace the origin of the various components in the reconstruction *(Pham et al. 2019)*. Ideally, the addition of elements is accompanied by confidence scores to ensure traceability and enhance the reusability of the **GSMN**. Finally, defining a robust biomass reaction for growth simulations *(Feist et Palsson 2010)*, identifying blocked reactions or infeasible energy cycles *(Schellenberger et al. 2011)*, and applying appropriate constraints to assess the predictive capacity of models are also essential tasks in the curation process *(Heirendt et al. 2019)*.

However, while curation grants researchers some flexibility in verifying reactions, metabolites, and GPR associations, this process is not immune to biases and errors introduced by those responsible for generating **GSMNs**. The potential for human error highlights the importance of carefully managing and documenting each step of the curation process to ensure network reliability *(Carey et al. 2020)*.

2.2.3 Reconstruction Process: From Reconstruction to Model

2.2.3.1 Models functionality and format compatibility

Converting the reconstruction into a model is not without obstacles. This includes searching for blocked reactions and gap metabolites *(Ponce-de-León et al. 2013)*, addressing redundant reactions despite this being a natural phenomenon *(Sambamoorthy et Raman 2018)* and identifying infeasible cycles *(Schellenberger et al. 2011; De Martino et al. 2013)*. As highlighted in the review by Lewis *et al.* *(2012)*, numerous algorithms and methods have been implemented over the years. These tools for constraint-based reconstruction and analyses (COBRA) are integrated into the COBRA Toolbox software suite *(Heirendt et al. 2019)*, available under MATLAB or as the Python version COBRApy *(Ebrahim et al. 2013)*.

At last, one of the criteria for validating the functionality of a **GSMN** is its ability to produce biomass. The quantitative modelling of this behaviour is usually conducted using flux-based approaches such as flux balance analysis (FBA) *(Orth et al. 2010)* or flux variability analysis (FVA) *(Gudmundsson et Thiele 2010)*, under given conditions while considering the system in a steady state. Mathematically, this analysis leads to a linear optimisation problem where the function to be optimised generally models the synthesis of essential biomolecules such as amino acids, membrane lipids or cofactors by the organism under study, collectively referred to as the biomass function *(Feist et Palsson 2010)*. Besides this crucial point, model validity should and can be improved through comparisons between the predicted and known physiological properties. Ideally, the **GSMNs** would be validated with experimental data, such as growth rate or compound production rate. All these aspects are essential before considering using the **GSMNs** for its intended purpose and applications.

Finally, a model must be standardised to facilitate collaborative development, evaluation and sharing. Frequently, data related to biochemical reactions are stored in the well-established XML-based format known as Systems Biology Markup Language (**SBML**). This format, associated with software libraries and tools, has become established over the past decade and continues to evolve *(Keating et al. 2020)*.

2.2.3.2 Analysis Strategies in GSMN Exploration

Depending on the problem under consideration, there are various approaches to conducting **GSMN** prospecting. Similarly, the methods and tools available for its analysis, much like those used in reconstruction processes, are diverse *(Lacroix et al. 2008; Wang et al. 2017; Vijayakumar et al. 2018)*. Formerly, two historical and complementary approaches can be distinguished *(Pitkänen et al. 2009)*, with the differences primarily arising from how the concept of metabolic pathways is translated into a mathematical model.

In the first case (**Figure 1-4.D1**), the analyses are carried out using graph theory, whereas in the second case (**Figure 1-4.D2**), they rely on the constraint-based modelling of the metabolic network *(Rezola et al. 2015)*. The choice between these two approaches depends on which perspective metabolism is being considered: from a structural or dynamic viewpoint, yielding qualitative or quantitative approaches, respectively.

The graph-based analysis is qualitative, focusing on the network structure, commonly referred to as topology, whereas the constraint-based modelling is quantitative, primarily based on flux analyses. From a quantitative perspective, several layers of precision and complexity can be added and superimposed onto this **GSMN** graph, whether oriented or not. For instance, incorporating stoichiometric data related to biochemical reactions enables a more detailed characterisation of metabolic states through metabolic flux quantification. These “Constraint-Based Modelling” (CBM) approaches *(Bordbar et al. 2014)* applied to **GSMNs** (**Figure 1-4.E**) allow the description of a functional metabolic network. The mathematical modelling of such a network is expressed as a set of ordinary differential equations that describe the temporal derivatives of chemical concentrations in a given state, explicitly encoding cause-and-effect relationships. Thus, this information helps identify the active reaction network under specific conditions or for particular cell types. Moreover, to focus on specific pathways, this model can be further refined by integrating kinetic data that considers enzymatic regulation mechanisms and metabolite concentrations, which are two dynamic factors.

Both modelling approaches are described thereafter. Wang *et al.* *(2017)* provide an overview of some of the existing computational tools behind these analyses, offering practical insights into the software landscape. In recent years, graph theory concepts and flux analyses have increasingly been used together to push the intrinsic limits of each of these approaches.

2.2.3.2.1 Qualitative Analyses Focused on Network Topology

In its simplest form, a graph is defined as a set of nodes connected by edges. In the context of the representation of metabolism, these elements represent biological entities such as metabolites, reactions, enzymes or genes *(Lacroix et al. 2008)*. From the selected GPR associations, the overall or partial metabolism of an organism can be modelled by three families of graphs, which will influence subsequent analysis choices *(Lacroix et al. 2008; Cottret et Jourdan 2010)*. The first intuitive representation is a graph centred on a biological entity. The nodes represent either metabolites, reactions or genes, while the edges indicate the existence of links between these objects. Bipartite graphs, on the other hand, depict two distinct node types which coexist, typically metabolites and reactions, with edges representing the interactions between these entities. This representation is more biologically grounded, as it mirrors the natural relationships between enzymatic reactions and their associated substrates or products. Finally, hypergraphs provide a more comprehensive depiction, connecting reactant metabolite nodes to product metabolite nodes *via* hyperedges. This approach closely aligns with the stoichiometric modelling, capturing the necessary input-output relationships of biochemical transformations. Once a network representation is selected, additional considerations must be addressed to provide a more detailed and dynamic view of the metabolic system. For instance, edges can be oriented to indicate reaction directionality or weighted to capture the magnitude of their influence. Comprehensive and detailed descriptions of the architecture and characteristics of biological network models were reviewed by Barabási *et al.* *(2004)*.

Graph theory serves as the cornerstone of network science, enabling the analysis of large networks using structural characteristics *(Majeed et Rauf 2020; Dusad et al. 2021; Wang et al. 2022)*. Graph-based models encompass more than just simple representations of reactions and metabolites. Concerning overall network connectivity, the degree distribution and the presence of hubs (*i.e.,* highly connected nodes) suggest a scale-free topology commonly observed in biological networks *(Erciyes 2023)*. As the removal of low-degree nodes has minimal impact and the loss of hubs can lead to significant network disruption, understanding the structural organisation of a metabolic network is fundamental to grasping the robustness of metabolic systems. Broadly, the analysis of centrality, including degree, betweenness, and closeness centrality, further highlights the relative importance of nodes within the network *(Wang et al. 2022)*. Nodes with high centrality frequently act as control points, playing crucial roles in network stability and functionality, underscoring their significance in maintaining coordination across the system.

In graph-based modelling of metabolic networks, one approach to capture metabolite producibility from nutrient resources is through the use of network expansion algorithms, which connect structural features to functional properties *(Kruse et Ebenhöh 2008)*. The minimal set of compounds that enables the synthesis of all other metabolites in the network is termed “seed compounds”. Defining these seeds is crucial for examining network growth and expansion, as it enables researchers to explore how an organism’s metabolism adapts or evolves in response to different environmental conditions. Predicting metabolic capacity involves studying the network’s scope, defined by the metabolite set that can be produced from a specific set of initial seeds. Besides examining organism-environment interactions, network expansion methods could be employed in metabolic network reconstruction, particularly during gap-filling phases *(Prigent et al. 2017)*. Thus, **GSMN** topology may provide insights into compound sets exogenously acquired *(Borenstein et al. 2008)* (*i.e.,* seeds), offering valuable guidance for designing culture media.

Clustering within the network offers additional insights into hidden subsystems or modules (*i.e.,* breakdown into sub-networks), representing functional groupings of metabolites or reactions These clusters often reveal pathways and subsystems that function semi-independently within the broader metabolic network. Additionally, chokepoints are reactions or metabolites that serve as critical passageways in a network, creating bottlenecks in flux flow *(Yeh et al. 2004; Oarga et al. 2020)*. These points are essential for maintaining metabolic balance, and targeting them can disrupt metabolic function. Such chokepoints are particularly relevant for drug targeting or metabolic engineering strategies *(Rahman et Schomburg 2006; Singh et al. 2007, 2009; Perumal et al. 2009; Kim et al. 2010; Taylor et al. 2013)*. Another crucial concept is small-world properties, which describe networks where most nodes can be reached from any other by a few steps, reflecting a highly efficient and modular structure. This small-world organisation enhances the system's resilience and adaptability *(Fell et Wagner 2000; Jeong et al. 2000; Wagner et Fell 2001)*.

In addition to mathematical graph traversal algorithms commonly used for network analysis, visual exploration offers an interactive and flexible approach for investigating **GSMNs**, tailored to specific research objectives. For instance, tools developed for automatic visualisation and interactive exploration facilitate the integration of omics data, especially fluxomics, and streamline network gap identification, significantly reducing curation time and supporting comparative analyses. While some tools are not exclusively designed for visualisation and also address aspects of network reconstruction, such as gap-filling guidance or flux optimisation, numerous platforms have been developed for these purposes. Notable examples include Cytoscape *(Su et al. 2014)*, CellNetAnalyser *(Klamt et al. 2007)*, Omix *(Droste et al. 2011)*, ModelExplorer *(Martyushenko et Almaas 2019)*, MetExplore *(Cottret et al. 2018)*, LMME *(Aichem et al. 2021)* and Fluxer *(Hari et Lobo 2020)*. Recent advancements have enabled the shift for transitioning from traditional 2D visualisation to virtual reality environments, offering enhanced capabilities for immersive and comprehensive exploration of global metabolic models *(Aichem et al. 2022)*.

2.2.3.2.2 *Quantitative Analyses Focused on the Flux Distribution*

Quantitative analyses use the stoichiometric matrix which represents the relationships between the reactions and metabolites within a metabolic network. In steady-state conditions, metabolites do not accumulate or deplete, and this equilibrium is mathematically expressed through a linear system of equations, where the rates of reactions consuming or producing each metabolite are balanced. This steady-state assumption simplifies the analysis of metabolic fluxes and allows us to focus on the relationships between reactions without the complications of dynamic changes in metabolite concentrations. However, since **GSMNs** typically contain more metabolites than reactions, the resulting system of equations is underdetermined, leading to an infinite set of possible flux distributions. Therefore, the solution space becomes prohibitively broad for practical use without the imposition of additional constraints which narrows the solution space *(Llaneras et Picó 2008)*.

To further refine the model, an objective function is often incorporated, guiding the system towards physiologically or bioengineering-relevant goals. The choice of the objective function in **GSMN** modelling can be summarised in three key points: **(1)** it must allow a simplified exploration of the solution space, **(2)** it should represent a bioengineering objective, and **(3)** it must have a meaningful physiological aim *(Kauffman et al. 2003)*. The most common physiological target is organismal growth, which is commonly monitored through a reaction known as biomass production.

Constraints define the space of all the feasible flux distributions of a system *(Orth et al. 2010)* and aim to represent biological reality as accurately as possible while allowing the resolution of the underlying mathematical problems. Linear programming algorithms are typically employed to solve these problems efficiently, identifying optimal flux distributions within the constrained solution space. Historically, linear programming in metabolic modelling began in the early 1980s *(Papoutsakis 1984; Watson 1986)* and was formalised in 1992 by Savinell and Palsson's work *(1992a, 1992b)*.Traditionally, solution space exploration is carried out by flux analyses, with FBA being the most widely used method. FBA allows for predictions of metabolic phenotypes under various conditions *(Anand et al. 2020)*.

Constraints in **GSMNs** can be broadly classified into four main categories: physico-chemical, topological, thermodynamic, and environmental *(Kauffman et al. 2003; Lee et al. 2006)*. The primary physicochemical constraint is mass conservation. According to the steady-state hypothesis, this key stoichiometric constraint assumes that the intracellular concentration of metabolites remains constant over time, ensuring mass balance within the model. Topological constraints pertain to restrictions on metabolites due to subcellular compartmentalisation, while thermodynamic constraints define the irreversibility of specific reactions. Environmental constraints are determined by the conditions of the culture media in which the cell is grown *(Lee et al. 2006)*.

Additionally, based on known biological data, maximum and minimum flux rates can be applied to specific reactions. A particular case involves limiting the directionality and reversibility of fluxes. In this context, it's noteworthy that flux rates can be directly measured in organisms using fluxomics strategies. These constraints ensure that the metabolic model closely aligns with biological reality while also allowing the model to be computationally tractable for analysis.

Constraint-based approaches can be categorised into two families: biased and unbiased approaches *(Lewis et al. 2012; Vijayakumar et al. 2018)*. Each of these families encompasses distinct strategies for interpreting the data and drawing conclusions about the underlying biological or chemical systems. Biased approaches typically incorporate prior knowledge or assumptions about the system under study. They can include predefined constraints or specific hypotheses that guide the analysis. The strength of biased approaches lies in their ability to focus on particular pathways or processes that are hypothesised to be significant. For example, in metabolic engineering, a researcher may prioritise pathways known to influence product yields. In contrast, unbiased approaches such as sampling fluxes aim to minimise preconceived notions about the system, allowing the data to dictate the conclusions drawn. This method fosters a more comprehensive exploration of the stoichiometric model, potentially revealing interactions and relationships that would otherwise remain hidden.

In terms of formalism, flux sampling methods can be applied to deterministic and stochastic models, whereas constraint-based models are initially formulated as deterministic equations. However, it is possible to introduce stochasticity into these models, which allows for a more accurate representation of the “living” nature of a cell by incorporating experimental noise or relaxing the steady-state assumption. Thus, flux sampling under steady-state conditions is one unbiased method for exploring the capabilities of a metabolic network by generating probability distributions rather than relying on fixed values. Unlike traditional flux analysis techniques (*e.g.,* FBA or FVA), sampling methods do not require the maximisation or the minimisation of an objective function *(Herrmann et al. 2019)*.

On the other hand, FBA is used to generate a set of steady-state fluxes for all reactions within a biochemical network by optimising an objective function under specific constraints. This approach allows the exploration of the **GSMN**'s capacity to produce biomass or specific compounds under varying environmental scenarios. FBA serves as a powerful tool for answering fundamental metabolic questions shaped by the choice of the objective function *(Anand et al. 2020)*. For instance, maximising metabolite production provides insights into biochemical capabilities, while minimising ATP generation or nutrient uptake helps assess metabolic efficiency. Moreover, maximising biomass production can highlight trade-offs between growth and metabolite overproduction. Consequently, flux analysis approaches offer numerous benefits, including characterising system behaviour, predicting drug targets, identifying biomarkers, optimising functionalities, designing metabolic phenotypes, and rationalising knock-out experiments or, more broadly, genome editing procedures. Over the years, FBA has become a widely approach for metabolic network analysis and has been extensively used to interrogate metabolic networks. Recent reviews, such as Anand *et al. (2020)*, delve into conventional FBA and its variants. Since then, the algorithms proposed have achieved significant milestones in the 2000s by integrating various biological concepts, including thermodynamics, regulatory mechanisms, and omics constraints. For a comprehensive overview of these methodologies and their connections, readers can consult reviews by Lewis *et al. (2012)* and Vijayakumar *et al. (2018)*.

2.2.3.2.3 Bridging Graph Theory and Flux Analysis for Enhanced Metabolic Modelling

Finally, while graph-based approaches offer valuable structural insights for modelling metabolism, it is critical to recognise their limitations. One major drawback is that graph-based approaches often ignore stoichiometric coefficients—required elements that define the quantities of metabolites involved in reactions—and do not account for reaction fluxes, which quantify the rates of these transformations. As a result, these approaches may offer less specificity compared to constraint-based approaches *(Pitkänen et al. 2009)*.

Moreover, graph traversal algorithms, such as depth-first search, k-shortest past, or breadth-first search *(van Helden et al. 2002)*, tend to generate a broader number of alternative pathways, many of which may lack biological relevance. This requires careful filtration steps to refine the outputs and draw meaningful conclusions about the system's behaviour.

While graph theory provides a valuable toolkit for visualising and exploring metabolic networks, it is often used alongside other modelling methods that incorporate stoichiometric and flux data, resulting in a more complete understanding of metabolic behaviours. By combining structural concepts with quantitative data, such as reaction fluxes and metabolite concentrations, graph theory offers a compelling framework for analysing metabolic networks. When paired with CBM, these approaches are used to rationalise experimental designs, optimise metabolic pathways, and predict the effects of gene knock-outs or environmental perturbations on overall system performance. For example, fragility coefficients of reactions derived from Minimal Cut Sets (MCSs) serve as a graph-based characteristic that quantifies vulnerability within metabolic networks *(Klamt 2006; Hädicke et Klamt 2011).* Coefficients close to 1 indicate high sensitivity and highlight crucial roles in maintaining network functioning. This integration aids in identifying essential reactions vital for metabolic performance and designs metabolic engineering strategies, enabling targeted modifications that can significantly alter metabolic output.

Furthermore, advancements in machine learning present exciting prospects for achieving a deeper and more comprehensive understanding of metabolic processes. These approaches will be only slightly discussed in this review, for more details, we invite the reader to refer to these two complete reviews: Antonakoudis *et al. (2020)* to understand how unsupervised and supervised machine learning algorithms are used together with FBA and that of Jin *et al. (2020)* for an overview of the possibilities of these methods applied to biological systems.

2.2.4 Towards a Better Overview Organism Functioning Understanding with Multicellular and Multiscale-Modelling

As powerful as it is, **GSMN** is only one layer among others that grasp organism complexity. The eukaryotic organisms modelling is complex since, in addition to the difficulty – and not trivial task – of assigning a gene to its reaction in the appropriate intracellular compartment (*i.e.,* nucleus, peroxisome, chloroplast, mitochondrion, *etc.*), the cells are grouped into tissues which together form organs with various functions. These spatial and functional differences lead to the emergence of multicellular metabolic models, which require, among other things, omics data integration for a better understanding of their interactions. For example, Lewis *et al. (2010)* simulate the energy metabolism of the brain by integrating gene expression data, proteomics data and manual curation to mimic the interaction between astrocytes and various neuron types relevant to Alzheimer's disease; Bogart *et al. (2016)* propose to model CO_2 assimilation within the leaves of C4 plants by correlating fluxes and expression data of RNA-seq. As proposed by Li *et al*. *(2022)*, the algorithms associated with the study of these multicellular **GSMNs** can be classified according to their objective, classical "phenotype predictors" and "network builders" that aim to extract the subsets necessary for the modelling of a particular tissue in a specific context (*i.e.,* enzyme expression levels are different from one tissue to another).

Mimicking the phenotypic behaviour of an organism, tissue or cell through a metabolic model is done through the integration and reconciliation of omics data multiple layers *(Ramon et al. 2018)*. In this sense, the reconstruction of **GSMN**, which expresses the metabolic potential

encoded in the genome, is, therefore, the first cornerstone necessary for global understanding and abstraction of cell metabolism. Data obtained with transcriptomic, proteomic, metabolomics or fluxomics allow a more accurate simulation and understanding of the biological process. One way to predict dynamic behaviour is mediated by adding descriptive parameters (*e.g.,* reaction rate, metabolites and enzyme concentrations) to reaction fluxes characterisation through kinetic models. The temporal dynamics of metabolites are described by studying their kinetics. However, due to the large number of parameters required for this type of approach, this is reserved for the study of specific pathways and cannot be generalised to the scale of a **GSMN**. In this regard, Saa and coworkers *(2017)* propose a review of modelling kinetic frameworks which require the structure of a network associated with thermodynamic constraints. However, this idea of mixing statistical and dynamic knowledge offered by a continuous or discrete model is not new *(Watson 1986)*, but it is now no longer limited by performance constraints.

Multi-formalism simulation is a real alternative as it involves the integration of **GSMN** data with, among others, the analysis of large clinical transcriptome and dynamic simulation studies. Recently, these challenges of hybrid modelling of biological systems have been addressed by the implementation of the MUFINS *(Wu et al. 2016)* software that reconciles multiple kinetic models with signalling and regulatory networks, omics data and FBA. Pienaar and co-workers *(2016)* present a multiscale model *of Mycobacterium tuberculosis*, which bridges metabolic scale to tissue scale by combining various data and models obtained both *in vivo*, *in vitro* and *in silico*. These layers' combination allows a better understanding of the interactions between the bacteria, its host and the environment in which it evolves. Among the results presented, let us mention, for example: **(1)** that the outcome of inflammation is more strongly correlated to bacterial properties than to those of the environment, **(2)** that the attenuation phenomenon is influenced temporally by the moment of inhibition of the enzymes (following the realisation of virtual knock-outs, the authors propose a quarantine of enzymes as antibiotic targets) and **(3)** that the identification of metabolic bypass pathways mitigate growth defects and may represent auxiliary drug targets. The integration of **GSMN** and Quasi-Steady State Petri Net

has improved our understanding of molecular interactions (*e.g.,* gene regulation and signalling pathways) in recent years *(Fisher et al. 2013)*. Examples include the work of Kędzia *et al. (2018)*, who modelled the communication system of liver cells around cytokines to study tumour growth as a function of the spatial location of a cell population and Maldonado *et al. (2018)*, who modelled adaptation to fats and sugars in non-alcoholic fatty liver disease. In public health and nanotoxicology context, Cordes *et al. (2018)* are developing a model that predicts the cellular response to drugs by proposing a workflow that reconciles a comprehensive drug-specific whole-body Physiologically Based Pharmacokinetic (PBPK) model and a cellular-level organ-specific **GSMN**. By validating their simulation by predicting phenotypic responses to a first-line antibacterial agent against *M. tuberculosis*, known to cause drug-induced liver damage, they are paving the way for an *in silico* simulation of drug side effects. With this in mind, Maldonado *et al. (2017)* present a tutorial to mix **GSMNs**, dynamic models of Gene Regulatory Network and PBPK. Finally, integrating machine learning with multiscale modelling opens up numerous possibilities but also challenges in biological and biomedical research *(Alber et al. 2019)*. These combined approaches enable more precise predictions of complex biological systems, ranging from molecular interactions to whole-organism behaviour, by leveraging data-driven techniques alongside mechanistic models. This synergy could revolutionise areas like drug discovery, allowing for more efficient identification of therapeutic targets and potentially improving the precision of medical treatments.

Ultimately, **GSMNs** serve as *in silico* platforms designed to simulate the phenotypic behaviour of cells or organisms. They integrate, systematise, structure and consolidate heterogeneous knowledge about an organism's metabolism, offering a comprehensive, organism-specific overview of metabolic pathways and functional capacities. By unifying this complex information into a single framework, **GSMNs** empower researchers to systematically explore metabolic potential, predict cellular responses to various environmental or genetic perturbations, and optimise experimental designs. These capabilities contribute significantly to advancements in both fundamental biological research and applied biotechnological innovation.

2.3 LEVERAGING GENOME SCALE METABOLIC NETWORK FOR OPTIMISED AND RATIONAL EXPERIMENTAL DESIGN

2.3.1 Unlocking Natural Product Biosynthesis

GSMNs' range of applications extends from enhancing our fundamental understanding of biological systems to fostering innovations in biotechnology and pharmaceutical development. *(O'Brien et al. 2015; Kim et al. 2017; Gu et al. 2019)*. With the advent of genome mining tools, the search for new natural products is shifting away from traditional screening methods and opening new research avenues *(Weber 2014; Medema et Fischbach 2015)*. These developments enable complete genome analysis to identify patterns markers for specific BGCs within genomic sequences *(Kim et al. 2017; Yee et al. 2023)*. Some of the most commonly used tools include the antiSMASH *(Blin et al. 2023)*, fungiSMASH *(Blin et al. 2023)*, and plantiSMASH *(Kautsar et al. 2017)* tool suites, designed to detect BGCs in bacteria, fungi, and plants, respectively *(Blin et al. 2021)*. Another notable tool is PRISM, which focuses specifically on detecting microbial BGCs *(Skinnider et al. 2020)*. The practical goal now is to understand the underlying mechanisms of these BGCs to induce and optimise their activation, which is responsible for NP production.

Typically, compound producibility in a **GSMN** is a response that can be assessed by topological analysis and flux-based studies. Given the diversity of tools and algorithms available for **GSMN** analysis, essentially developed based on the formalisms presented in section 2.2.3.2 *Analysis Strategies in GSMN Exploration,* we have chosen to categorise NPs research approaches as follows: those based on internal perturbations (*e.g.,* gene inactivation or addition, which affect reactions) and those focused on external conditions (*e.g.,* simulating various nutrient media, adding cofactors or elicitors). The first approach aligns with strain design, while the second pertains to media design (**Figure 1-5**).

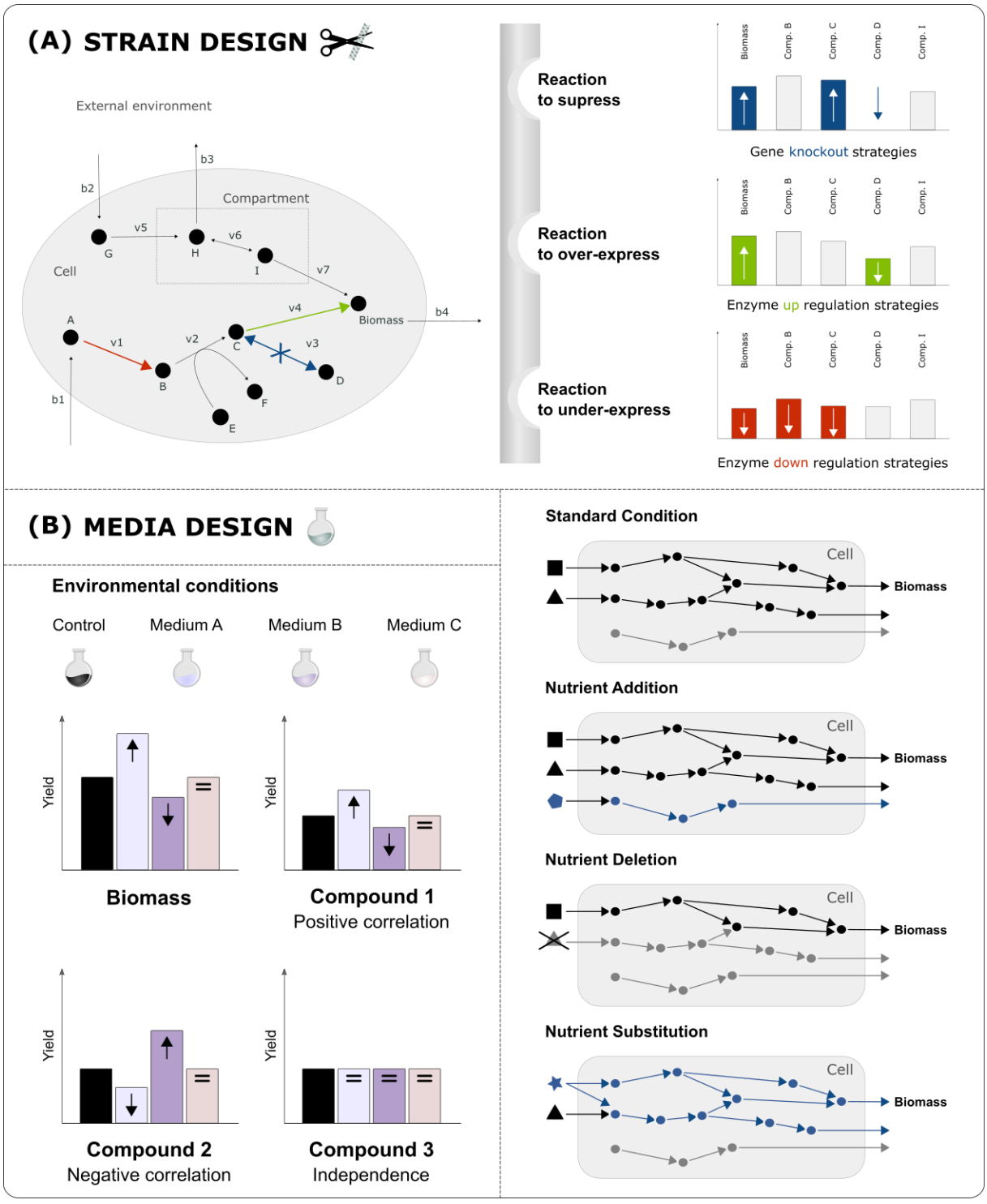


Figure 1-5: Overview of strategies for modulating an organism's metabolic capacity through GSMN analysis. Strain design **(A)** and Media design **(B)** are driven by internal and external parameters, respectively. Strain design involves internal perturbations, such as gene knock-outs or additions which directly modify the organism's metabolic pathways. In contrast, Media design focuses on external modifications, like altering nutrient availability or environmental conditions, to affect metabolic activity. By leveraging the predictive power of **GSMNs**, these strategies enable the rational design of experiments to improve growth conditions, enhance product yields, and discover new metabolic potentials.

The drug development pipeline comprises three key stages: discovery, development, and clinical trials *(Dougherty et al. 2017)*. **GSMNs** provide a structured systems biology framework that can potentially aid in the discovery phase by streamlining the process of identifying optimal drug targets. However, how might these approaches enable the access, production and optimisation of interest NPs? Conceptually, specific or general strategies can be envisaged to answer these questions. Using prior genomic knowledge to focus on a group of specifically chosen genes, such as targeted BGCs, to understand which parameters influence the production of their gene products and how to modulate them is an approach that can be likened to a "bottom-up" approach. This method allows precise tuning of genetic and/or environmental elements to enhance NP yields. Conversely, more general "top-down" approaches aim to explore the broader effects of network perturbations, such as recreating experimental conditions *via* methods like **OSMAC**, which seeks to induce diverse metabolic responses across varying environmental conditions and, thus, assess the perturbation consequences.

2.3.2 Metabolic Network for Strain Design

2.3.2.1 Metabolic Engineering: From Evolutionary to Computational Strategies

Formally introduced in the early 1990s *(Bailey 1991)*, “metabolic engineering” expression describes the manipulation of cellular enzymatic, regulatory, and transport processes through recombinant DNA technology aimed at increasing the yield of targeted products *(Stephanopoulos et Sinskey 1993)*. Over the past thirty years, various terms (*e.g.,* evolutionary, systems, computational, synthetic, environmental) have been associated with metabolic engineering, reflecting the evolving concepts and techniques in the field *(Jang et al. 2019; Sohn et al. 2020)*. From early methods relying on natural selection and random mutagenesis, the techniques have advanced significantly. Today, *in silico* modelling allows for the rational identification and optimisation of genetic targets, streamlining the search for improved metabolic pathways. For a thorough historical overview of the development of these techniques and the key milestones achieved over the past three decades, we strongly recommend the review by Kim *et al. (2023)*.

Metabolic engineering entails the precise manipulation of cellular processes to transform micro-organisms into efficient cell factories for producing valuable compounds. By rewiring their metabolic pathways, engineers can redirect metabolic flux towards chemical production and enhance the bioconversion of industrial by-products into high-value products *(Rangel et al. 2020)*. Rational and intuitive approaches in metabolic engineering are often combined to achieve multiple goals, such as enhancing carbon source utilisation, modifying transporters for efficient product export, reducing by-product synthesis, and rerouting metabolic pathways for better precursor conversion *(Lee et al. 2012)*. However, due to the inherent complexity of cellular networks, these methods alone are insufficient to optimise cellular performance. Consequently, system-wide approaches, including *in silico* and omics-based analyses, have been developed to enable more precise target gene selection and fine-tuning of metabolic pathways *(Lee et al. 2012)*.

Metabolic engineering harnesses a variety of biosynthetic pathways to enhance the production of valuable compounds. These biosynthetic pathways can be classified into three categories: **(1)** native pathways, which rely on the natural metabolic mechanisms of an organism, allowing for targeted improvements in productivity; **(2)** existing non-native pathways, which involve the introduction of metabolic pathways from other organisms adapted to function in the target host; and **(3)** created non-native pathways, which are artificially designed through retro-biosynthesis by combining various enzymes to establish a new metabolic flux *(Cho et al. 2022)*. In strain design, the concepts of metabolic engineering focus on identifying genes to delete, add, or modulate (*i.e.,* overexpression or downregulation) to optimise the production of a target compound while supporting the organism's growth (**Figure 1-5.A**). These actions, typically carried out through well-established genome editing methods *(Oost et Patinios 2023)*, affect the organism's functioning and allow for targeting different metabolic components: reactions, enzymes, and metabolites. Within the **GSMNs**, these various layers of understanding are ensured by the GPR associations. Over the years, the concepts of strain design have been bolstered by a plethora of methodologies and algorithms, with CBM methods being among the most prevalent *(Machado et Herrgård*

2015; Maia et al. 2016). However, a comprehensive description of all the methods developed is beyond the scope of this review, as these topics have been extensively covered in numerous publications. For a quick overview, we recommend consulting the works of Machado *et al. (2015)* and Long *et al. (2015)*. For a more in-depth understanding of the primary *in silico* constraint-based strain design strategies and algorithms, we advise referring to the review by Maia *et al. (2016)* and the references cited therein.

Finally, modelling **GSMNs** is crucial for advancing our understanding of biological systems. However, it remains essential to complement *in silico* findings with experimental validation. The insights gained from these analyses can generate hypotheses that guide experimental strategies, such as evolutionary engineering techniques (*e.g.,* chemical and physical mutagenesis, adaptive laboratory evolution, directed evolution, assisted genome evolution, and high-throughput screening) *(Shepelin et al. 2018; Zhu et al. 2018; Sandberg et al. 2019)*. These methods, which target either specific genes or the entire genome, enable the characterisation of phenotypes of interest while testing the robustness and viability of the resulting mutants. Ultimately, these mutation effects can be elucidated through multi-omics data studies (*e.g.,* transcriptomics, proteomics, metabolomics, fluxomics) *(Yizhak et al. 2010; Kim et al. 2018; Ramon et al. 2018)*. Together, these approaches provide a systematic workflow for the rational development of cell factories.

2.3.2.2 Targeting Genes for Deletion: The Principle of Essentiality

In a biological system, the essentiality of an element refers to its critical role in the survival or growth of an organism. A gene, enzyme, reaction, or metabolite is deemed essential if its deletion or absence significantly compromises the organism's viability or hinders its development under specific culture conditions. This state can be effectively simulated simply by removing a reaction from the **GSMN** and assessing the changes in predicted growth rates before and after this modification *(Joyce et Palsson 2008)*. Research on essentiality is valuable during the early phases of drug development, as it aids in novel therapeutic target identification *(Dougherty et al. 2017)*.

The most straightforward conceptual approach for identifying drug targets using **GSMNs** is the gene/reaction-centric approach, also known as knock-out strategies. For instance, Kumelj *et al.* *(2019)* explore robust gene-knock-out strategies to increase the production of acetyl-CoA, a precursor of polyketides, in *S. coelicolor*. Reactions essential for maintaining a non-zero flux through a specific objective function may serve as a potential goal to be achieved; thus, considering compound production. Once identified, the GPR rules can be employed to pinpoint the appropriate enzyme to target for inhibiting the corresponding reaction *(Rawls et al. 2020)*. Numerous methods (*i.e.,* CBM-based, Elementary-Modes Analysis -based, graph-based) and algorithms have been developed to enhance the efficiency of gene knock-out research. For the purposes of this review, we will only mention OptKnock, which identifies growth-coupled, non-intuitive gene deletion strategies *(Burgard et al. 2003)*, and OptGene, an evolutionary programming-based algorithm that relies on heuristic searches *(Patil et al. 2005)*. OptKnock has been successfully applied to predict gene deletions in *Escherichia coli* for enhanced succinate and lactate production *(Burgard et al. 2003)*. Meanwhile, OptGene demonstrated its utility by identifying gene deletion strategies in a recombinant *S. cerevisiae* strain capable of producing vanillin, a natural flavour compound *(Patil et al. 2005)*. On the other hand, we recommend the review by Valderrama-Gomez *et al.* *(2017)*, which provides an overview of methods for increasing succinate production (*i.e.,* spanning 26 theoretical studies from 2003 to 2016) and presents the algorithms available for strain optimisation.

In those cases, optimisation relies on identifying pathways linked which, when deleted, do not affect growth, or minimally, while increasing the targeted compound production. Two cases may thus be related to the overproduction of targeted compounds: **(1)** highlighting reaction(s) responsible for targeted compound consumption and **(2)** highlighting reaction(s) responsible for the consumption of precursor, thus diverting the metabolic flux away from targeted compound production. Conceptually, removing both types of reactions yields the increased production of the compound of interest.

Elementary-Modes Analysis (EMA) is one of the foundational methods in the constraint-based modelling and analysis of metabolic reaction networks *(Schuster et Hilgetag 1994)*. It identifies minimal sets of enzyme-catalysed reactions (*i.e.,* elementary modes) that operate in steady-state and link substrates to biomass and products in a metabolic network. EMA facilitates the enumeration and analysis of all possible metabolic pathways, allowing for comparing these routes based on various criteria such as efficiency. Unlike FBA, which focuses on optimising specific objectives, EMA provides a comprehensive map of all feasible pathways in the network. Additionally, it helps identify key reactions and potential gene knock-out targets to block competing pathways and redirect metabolic flux toward the production of desired compounds *(Trinh et al. 2009)*. However, as the complexity of the network increases, the number of possible elementary modes grows exponentially, making the analysis computationally intensive and more difficult to interpret. To address this challenge, Machado *et al. (2012)* developed a sampling approach that computes elementary modes in large-scale networks without generating the entire set. Their method, applied to *E. coli* to enhance succinate production as a proof of concept, successfully predicted the most optimal knock-outs and achieved near-optimal solutions. Furthermore, Hädicke *et al. (2011)* advanced these techniques with their generalised Minimal Cut Sets (MCSs) approach, called “Constrained MCSs”, which determine robust knock-out strategies for simultaneously enhancing biomass and product synthesis. They demonstrated this by improving ethanol production in *E. coli*. Another significant contribution came from Toya *et al. (2015)* with the development of SSDesign, a tool specifically for predicting gene knock-out combinations that enforce target compound production under non-growing conditions, once again tested on succinate production in *E. coli*.

2.3.2.3 Targeting genes to modulate, the principle of metabolic regulation

However, gene deletion is only one of many strategies available in genetic manipulation. Within a biological system, the gene function is not only a state of being switched off or on. *In vivo*, metabolic flux can be redirected toward a target metabolite by modulating gene

expression through either overexpression or underexpression. Unlike simple gene knock-outs, these approaches allow for a more nuanced control, either concentrating metabolic fluxes on a specific production pathway or attenuating an essential function. This flexibility enables fine-tuning of metabolic pathways to enhance production or disrupt critical cellular processes.

Studying and detecting the impact of flux variation resulting from gene overexpression is, however, more challenging than the simple gene deletion knock-out described so far. In addition to the fact that sets of genes and reactions are neither bijective sets nor one-to-one relationships, the presence of multifunctional enzymes, enzyme complexes and isoenzymes can complicate the strain design strategies resulting from *in silico* knock-out approaches.

The overproduction of ethanol and succinate in *E. coli* was used as a case study to validate the feasibility of metabolic engineering algorithms such as OptReg *(Pharkya et Maranas 2006)*, OptForce *(Ranganathan et al. 2010)*, and OptORF *(Kim et Reed 2010)*. OptReg aims to identify the optimal combination of regulatory interventions, OptForce provides targeted strategies for adjusting metabolic fluxes, and OptORF simultaneously optimises knock-outs and overexpressions by incorporating regulatory and metabolic data. These algorithms were evaluated for their effectiveness in optimising metabolic pathways by identifying gene knock-outs, overexpression, and regulatory modifications in common compound production. Together, these methods have expanded the toolkit for strain improvement, enabling more efficient production of biofuels and other valuable chemicals. However, the limited availability of regulatory data has severely constrained the dissemination and application of the OptOrf approach. In contrast, OptForce has enabled the comprehensive identification of minimal genetic intervention strategies, leading to various intriguing characterizations. For instance, Xu *et al. (2011)* identified interventions in *E. coli* that cooperatively direct carbon flux toward malonyl-CoA while simultaneously preventing its diversion to by-products. The predicted genetic interventions result in a significant enhancement of yields for this fundamental building block in the biosynthesis of numerous natural products, exemplified by the production of naringenin.

In 2010, Choi *et al.* (2010) developed a strategy for *in silico* selection of genome-wide gene amplification targets called “flux scanning based on enforced objective flux” (FSEOF). The FSEOF algorithm imposes additional constraints during flux analysis to optimise cell growth and production of the desired product, thus providing a more reliable alternative. These objectives, which are sometimes mutually exclusive, especially in the search for specialised metabolites, are then compatible since the newly set constraints are based on a gradual increase from the initial flux value to a value adjacent to the theoretical maximum value for product formation. They demonstrated which genes should be overexpressed in *E. coli* to induce improved lycopene production. In 2012, this algorithm was extended by Park *et al.* (2012) to include an iterative notion of flux variability (*i.e.,* “flux variability scanning based on enforced objective flux” – FVSEOF). Using physiological omics data with grouped reactions, they were able to demonstrate the impact of the amplification of some genes inducing increased production of shikimate acid (*i.e.,* natural products precursor) and putrescine in *E. coli*. In 2021, Raajaraam *et al.* (2021) proposed an improvement of the original algorithm by focusing on the cooperative production of metabolite pairs in *E. coli* and *S. cerevisiae*. Their study, named co-FESOF (*i.e.,* “co-production using Flux Scanning based on Enforced Objective Flux”), showed an improvement in the production of multiple metabolite pairs, including essential precursors of specialised metabolites, by overexpressing a few genes. They also noted a better production of compounds under anaerobic conditions compared to aerobic conditions for both species, with a particular attraction for *S. cerevisiae* (*i.e.,* metabolites significantly more interesting for the industry).

Thus, coupled, for example, with knock-out analyses or essential gene searches, the FSEOF algorithm, which is mainly used on micro-organisms, can either focus more specifically on the general precursors of the metabolites of interest or directly on the latter. The following work based on the target gene expression amplification has led to improved metabolite production, such as the carotenoid astaxanthin in *E. coli* (Park et al. 2018), the antibiotic actinorhodin in *Streptomyces coelicolor* (Kim et al. 2014), chorismite derivatives in *Streptomyces albus*

(Kittikunapong et al. 2021), cyanide in *Bacillus megaterium* *(Aminian-Dehkordi et al. 2019)*, chaxamycin A and chaxalactin A (*i.e.,* molecules with antibiotic and anticarcinogenic activity) in *Streptomyces leeuwenhoekii* *(Razmilic et al. 2018)*, *α*-tocopherol in *Helianthus annuus* *(Srinivasan et al. 2019)*, Gibberellin GA_3 in *Fusarium fujikuroi* *(Li et al. 2023)* and hyaluronan in *Lactococcus lactis* *(Badri et al. 2019)*.

2.3.2.4 Targeting "synthetic" reactions to be added, the non-native biosynthetic pathways

Non-native biosynthetic pathways involve the integration of enzymes or metabolic intermediates from various organisms into microbial production hosts such as *E. coli*, *Pseudomonas* strains, *Bacillus* strains, or diverse yeast species *(Wendisch 2016)*. *In vitro*, the introduction of these pathways generally takes two forms: either by expressing a gene cluster heterologously *(Meng et al. 2022; Chiang et al. 2023; Cox 2024)* or by adding a new pathway to generate specific precursors not originally synthesised by the host *(Choumpi et al. 2020; van der Hoek et al. 2022)*. Incorporating such pathways in a microbial host allows researchers to bypass the limitations of the host's native metabolism, creating opportunities to synthesise complex molecules more efficiently by novel biosynthetic routes creation. This method is particularly valuable when the target compound requires unique intermediates that the host cannot readily produce.

To implement non-native biosynthetic pathways, researchers first identify the desired reactions for integration within the metabolic network. The process starts with pathway identification, wherein researchers outline the biosynthetic pathway necessary for the target compound. This often involves utilising enzymes from diverse organisms and conducting literature reviews, accessing pathway databases, and employing computational modelling. Afterwards, specific non-native enzymes are selected for critical biosynthetic steps, sometimes introducing new metabolic branches to support pathway integration.

Then, computational tools, such as OptStrain *(Pharkya et al. 2004)*, could be employed to simulate the integration, predicting how genetic modifications may impact metabolic fluxes and identifying gene knock-outs or overexpressions that would optimise production. For example, OptStrain has enabled *E. coli* to produce vanillin by integrating a non-native pathway *(Pharkya et al. 2004)*.

However, despite these advances and to the best of our knowledge, prior *in silico* modelling use for enhancing natural product yields remains limited, pointing to an area with substantial potential for further research.

2.3.3 Designing Media Culture Through Genome Scale Metabolic Network Analyses

2.3.3.1 General Principles and Guidelines

The chemical diversity exhibited by a micro-organism is meaningful only if the experimental conditions leading to compound production are clearly defined *(Herbert 1961; Egli 2015)*. Indeed, a biological organism's ability to adapt and express various phenotypes in response to environmental changes is a rudimentary characteristic. As genetic factors and environmental conditions shape the expressed phenotype, exposing the organism to diverse culture environments is crucial for unlocking its metabolic capabilities (*i.e.,* modelling distinct phenotypes). Therefore, the starting point for metabolic engineering lies in the tailored culture media development for the organisms under study *(Egli 2015)*.

The design and optimisation of culture media have long been critical factors in enhancing cell culture performance *(Jerums et Yang 2005)*. The composition of culture media directly influences several aspects, including cell growth rate, production of target compounds, by-product formation, fermentation equipment, and the size and number of downstream processing units *(Bode et al. 2002; Rangel et al. 2020)*. A cell culture media is a complex blend of essential nutrients, including amino acids, minerals, salts, inhibitors, vitamins, yeast extracts, meat extracts, and growth factors, which are vital for supporting the growth and productivity of the organism *(Jerums et Yang 2005)*. Effective media formulations are crucial

for achieving optimal cell density, viability, yield, and product quality, regardless of the chosen culture method - batch, fed-batch, or continuous culture. However, developing a culture media that promotes the growth of micro-organisms or cells presents significant challenges, particularly in finding approaches that balance efficiency, cost-effectiveness, and performance outcomes. One promising solution involves *in silico* design of media based on prior metabolic knowledge of the targeted cell line, reducing the experimental burden in the laboratory. Studying the media composition can assist the rational design of experiments, helping to optimise costs and simulate various environmental conditions, such as anaerobic versus aerobic states or nutrient-rich versus nutrient-limited settings. However, designing an ideal culture media is a combinatorial problem unsuited to an exhaustive experimental exploration. In this context, using **GSMNs** as prospective and predictive tools is a powerful approach to simulating an organism's metabolic potential. For instance, FBA and its variations can model the organism's growth under different perturbations, ensuring that its metabolic demands are satisfied despite changes in environmental conditions. Thus, limiting conditions, such as auxotrophic, can be pinpointed.

In addition to nutrient enrichment or depletion, another fundamental parameter for **GSMN** modelling is the management of energy cofactors. These chemical substances maintain the cellular redox balance, operating as substrates or catalysts for biochemical reactions. The most common are acetyl coenzyme A, NAD(P)H/NAD(P)+ and ATP/ADP *(Sun et al. 2023)*. They serve as redox carriers, electron carriers, energy carriers or functional groups in anabolic and catabolic reactions. Cofactor engineering, a significant aspect of metabolic engineering, involves modifying the form or intracellular concentration cofactors to manipulate metabolic fluxes. In media design, such modulations are achieved by adding substrates with different oxidation potentials, activators, inhibitors, competitors, synthetic analogues or precursors of their biosynthesis, factors that can significantly influence cell growth and metabolic performance *(Xu et al. 2017)*.

Fundamentally, in the context of **GSMN** modelling, two principal factors govern the simulation of environmental media: the nature and quantity of elements that constitute the medium. These external parameters can be modulated through specific interventions, summarised by adding, removing or substituting elements (**Figure 1-5.B**) or by altering the import rates of one or more components. Such disturbances provide a framework for systematically studying the organism's metabolic responses to its environment changes, enabling a targeted approach to media design and metabolic optimisation.

Within a **GSMN** model, the organism's response to environmental changes is modelled *via* exchange reactions, enabling the uptake of nutrients and secretion of metabolic by-products. However, accurately defining the nutrient environment in which the organism evolves is a complex process. Organisms' natural environments are often dynamic and variable, the compositions of those used in laboratory conditions are not always accurately known, and nutrient inputs, both in terms of nature and concentration, are highly specific to the organism studied. Since **GSMN** predictions of metabolic capabilities are sensitive to the precise definition of the nutritional context, it is crucial to develop clear, standardised guidelines for nutrient modelling to ensure reliable and reproducible outcomes *(Marinos et al. 2020; Bernstein et al. 2021)*. Nevertheless, once this framework is well-established, numerous modelling scenarios can be envisioned. It is well-established that exploring different environments can uncover a wide range of behaviours in organisms. Consequently, a thorough understanding of minimal nutrient requirements can help in identifying precursors that may enhance growth or, conversely, impede development. Finally, comparing flux profiles from nutrient-rich and nutrient-poor simulations can provide valuable insights into which elements influence the activity of specific regions within **GSMN** models.

To our knowledge, most studies assessing the effects of the outlined processes focus primarily on cellular growth responses to environmental disturbances or on the by-products production not typically classified as high-value NPs. Thus, as the variations are studied regarding the biomass reaction flux evolution, three correlation scenarios — positive

correlation, negative correlation, or independence (**Figure 1-5.B**) — must be considered for extending these methods to NPs. However, many studies assume a linear relationship, overlooking the potential combinatorial effects of nutrients, feedback loops *(Chung et al. 2021)* and inherent trade-offs *(Hashemi et al. 2021; Ekkers et al. 2022)*. Consequently, this section aims to provide essential insights as proof of concept, highlighting necessary adjustments to enhance the discovery and optimisation of valuable NPs in this field of application.

2.3.3.2 Enhancing Culture Media

An organism's inability to grow in a simulation model suggests that at least one component necessary for biomass production cannot be synthesised (*i.e.,* the absence of metabolic flux in the relevant reactions). Thus, when a nutrient is removed from the uptake list, the persistence of biomass flux indicates that the nutrient is non-essential or that an alternative compensatory pathway exists. Conversely, if the biomass flux is absent, this may suggest potential auxotrophy of the organism. Such auxotrophy can be confirmed or refuted through topological analyses of compound productivity (*e.g.,* by verifying the absence of specific transmembrane proteins associated with a substrate, evaluating the seed list in a network, and examining the connectivity of metabolites). Nonetheless, it is essential to acknowledge that this approach constitutes only a preliminary and overly simplistic step toward identifying this much more complex phenomenon *(Seif et al. 2020)*. Moreover, it is essential to ensure that these auxotrophies are not the result of biases introduced during the gap-filling steps in the curation process *(Borer et Magnúsdóttir 2023)*. Identifying the auxotrophies of a cell or organism is crucial for determining the components necessary for a culture medium, thus informing essential metabolic needs, optimising culture conditions, designing enhanced strains, and assessing their viability.

Pathogen culture requires controlled media to ensure assay reproducibility and to provide deeper insights into their mechanisms of action, such as adaptation, stress resistance, and virulence. Seeking favourable conditions for their growth is then an objective to prioritise, especially as fastidious growth is a paradoxical property of some prokaryotic pathogens *(Gerlin*

et al. 2020). In this regard, the work of Tejera *et al.* *(2020)* defines a simple media for the growth of *Campylobacter jejuni*, which is a cause of foodborne bacterial gastroenteritis. By blocking media transporters one at a time and using constraint modelling the authors identified methionine, niacinamide, and pantothenate as auxotrophic substrates. Supported by experimental validation they finally provide a suitable environment generalised to eight other genus strains by supplementing the culture media with pyruvate, cysteine, serine, and glutamine. However, as evidenced by the **GSMN** analyses of the plant pathogen *Xylella fastidiosa*, which possesses a minimalist network with limited robustness and flexibility, slow growth cannot always be attributed solely to the composition of the culture media *(Gerlin et al. 2020)*.

Another way to refine the selection of compounds for exploration is based on their productibility as determined by network topology and the seed concept. Different from essential compounds (*i.e.,* compounds required for the survival of the organism in a given environment), the seed set of an organism corresponds to a minimal compounds collection in the network that allows the synthesis of all other compounds in the network (*i.e.,* union of essential sets required in all environments) *(Kruse et Ebenhöh 2008)*. As we know, the metabolic activity of micro-organisms is deeply dependent on the biochemical environment in which they evolve *(Monk et al. 2013)*, and results provided by topological analyses are the witness of the organism's interaction with its environment. **GSMN** topology may insight compound set exogenously acquired and thus help in culture media design. Moreover, one of the advantages of using algorithms derived exclusively from graph theory is that they are free of model stochiometry and can therefore be used on incomplete networks, or organisms, with partial information. For instance, Borenstein *et al.* *(2008)* propose a framework based on the decomposition of Strongly Connected Compounds and the Kosaraju algorithm to determine the set of compounds absorbed by an organism in various environments and thus take up the seed set concept. Ignoring the stoichiometry, dynamics and concentration of reactants and products, their study, conducted on a cross-species analysis of 478 networks, reveals seed sets of different sizes and nature depending on the environment of the organism studied

(*e.g.,* obligate parasites present a seed set of smaller size than species with larger adaptive capacities). Using the seed set framework on uncultivated micro-organisms from the incomplete genome (*i.e.,* collection of metagenome-assembled and single-cell genomes) from freshwater actinobacterial lineage, Hamilton *et al.* *(2017)* retained 31 metabolites in the final set of proposed auxotrophy and nutrients *in silico* and then refined them by experimental measurement. Concepts of "reverse ecology" *(Levy et Borenstein 2012)* (*i.e.,* inferring environmental characteristics from information in biological systems) have paved the way for exogenously acquired compounds prediction and organism metabolic capacity forecasting. Thus, these approaches contribute to an understanding of the mechanisms deployed by an organism in its natural environment but also highlight its synthetic capacities in other habitats.

2.3.3.3 Depleting Culture Media

Questions regarding the removal, substitution, and concentration reduction of nutriment are crucial for optimising culture media design (*i.e.,* minimising costs while maximising the yield of desired products). *In silico* media design aims to lower costs while maintaining optimal biomass yields. Primarily motivated by industrial and economic factors, various simulations have been conducted to reduce production costs while ensuring optimal biomass yield. For instance, as demonstrated in their work, using the genome-scale model of the CHO-K1 mammalian cell line and analyses with an extension of FBA that considers molecular crowding and assumes that enzymes have a maximum turnover rate, Pérez-Fernández *et al.* *(2021)* have identified an optimal culture media compared to standard one. This optimised media reduces overall production costs at high cell densities by pinpointing non-essential amino acids, proposing more cost-effective alternatives, such as glycine, or diminishing some amino acid concentrations, while maintaining a growth rate equivalent to that achieved with the reference medium. Since the cost of each component in the media varies based on factors such as their availability or accessibility, finding substitutable compounds highlights the potential to reduce production costs.

In our quest to pinpoint targets of interest, these strategies are meaningful only if a positive correlation is presumed between biomass production and the production of a relevant by-product. In their validation of the model for *Saccharopolyspora erythraea* producing the antibiotic erythromycin, Licona-Cassani *et al.* *(2012)* investigated the impact of amino acid supplementation on both growth and antibiotic yield. Their *in silico* analyses validated experimentally, demonstrated that high erythromycin yields can be achieved without compromising biomass production. This finding underscores the potential of targeted nutrient supplementation in optimising antibiotic production while maintaining cellular growth, contributing to more efficient bioprocessing in industrial applications. In their work, Deantas-Jahn *et al.* *(2024)* proposing different nitrogen source mixtures for optimal growth, analyses of the **GSMN** of the haloalkaliphilic bacterium *Halomonas campaniensis* improve by a factor of 1.5 and 2.5, respectively, the biomass yield along with polyhydroxybutyrate titer, an alternative for conventional fossil fuel-based plastics. This dual benefit underscores the potential for economically efficient solutions that boost productivity in biotechnological applications.

However, as specialised metabolites are by definition non-essential for the organism's growth and are used, for instance, to improve survival rate in a competitive environment, an inherent trade-off between basal and specialised metabolism occurs. As a result, optimising for growth and maximising NP production can also present conflicting objectives and thus being antagonistic. Furthermore, an increase in nutrient availability is not a guaranteed means of enhancing biomass. Isidro *et al.* *(2016)* propose a method for extracting meaningful metabolic knowledge from culture media screening data adapted from their previous development on Projection to Latent Pathways (PLP), a constraint version of Partial Least Square (PLS) regression at the level of metabolic pathways regulation. PLP analysis minimises the number of active elementary flux modes while maximising correlations with fluxome and envirome measurements. By extending this concept to the analysis of media screening data (*i.e.,* understanding how media components up- or down-regulate key metabolic pathways) the

authors highlight conditions that have a significant impact on the growth, maintenance and secretion of relatively common by-products (*e.g.,* ethanol, citrate, pyruvate and acetate) in *Pichia pastoris*. Thus, due to an inhibition of the overall metabolic activity of this yeast in the presence of defined concentration ranges of iron and manganese, this work suggests, among other things, decreasing their concentration to favour its development.

2.3.3.4 Study the Correlation Between Nutrient, Biomass and Bioproduct Production

To establish an optimal growth medium, when evaluating biomass production as a function of nutrient availability, a proportional relationship should be observed between nutrient concentration and growth *(Egli 2015)*. Additionally, optimal nutrient levels can be identified by observing a distinct transition between nutrient-limited (*i.e.,* growth phase) and non-limited (*i.e.,* saturation phase) conditions. Any deviation from these principles may suggest the influence of additional factors *(Egli 2015)*. A straightforward yet effective way to evaluate the effects of environmental variables is through Phenotype Phase Plane (PhPP) analyses *(Edwards et al. 2002)*.

The PhPP serves as a robust computational tool in metabolic network studies, allowing researchers to visualise and analyse the dynamic behaviours of metabolic pathways under various conditions. By mapping the interactions between different metabolic fluxes and cellular states, PhPP provides valuable insights into how changes in substrate availability, enzyme activity, and other factors impact metabolic outcomes. Initially, PhPP analyses were studied under oxygen and glucose-limiting conditions in *E. coli* *(Edwards et al. 2002)* and *S. cerevisiae* *(Duarte et al. 2004)* to understand the function and capacity of the organisms' metabolic machinery. Ethanol obtained from fermentation processes initiated under anaerobic conditions is a fundamental component in biofuel production

By utilising PhPP, scientists can assess the robustness and adaptability of metabolic systems, thereby gaining a better understanding of how organisms react to environmental changes. For example, Acevedo *et al.* *(2017)* performed a PhPP analysis on the yeast *Scheffersomyces stipitis* and confirmed the essential role of cofactors in enhancing

fermentation processes. Another instance is the work of Nanda *et al*. *(2020)* on the yeast *Lachancea kluyveri,* demonstrating how ethyl acetate production can be influenced by varying the carbon source under controlled oxygen uptake rates. This study highlights the role of substrate and oxygen availability in modulating product yield, providing insights into optimising metabolic outputs through environmental adjustments.

PhPP has also proven valuable in identifying limiting metabolites in specific pathways. For example, in *E. coli*, sensitivity and PhPP analyses identified NADPH and tryptophan as bottlenecks in violacein biosynthesis, which has therapeutic potential *(Immanuel et al* 2018).

In summary, the PhPP is a crucial framework for understanding metabolic dynamics and supporting decision-making in metabolic engineering, ultimately facilitating the optimisation of valuable compound production in biotechnology and biomanufacturing.

2.4 CONCLUSION ET PERSPECTIVES

En résumé, la modélisation du métabolisme par les **GSMNs** a trouvé, au fil du temps, diverses applications biotechnologiques d'intérêt avec notamment les succès rencontrés pour l'optimisation de produits à hautes valeurs ajoutées. Le potentiel prédictif des **GSMNs** a été largement exploré dans la littérature, avec de nombreuses publications présentant des méthodes qui démontrent l'applicabilité et la faisabilité de ces analyses. Cependant, il est important de noter que ces éléments restent, d'une part, trop souvent limités à des micro-organismes largement étudiés tels que *E. coli*, *S. cerevisiae* ou *M. tuberculosis*, ainsi qu'à la production de composés de base et, d'autre part, lorsque nous nous concentrons exclusivement sur la définition stricte des produits naturels au sens biologique, les éléments disponibles se font de plus en plus rares.

Néanmoins, la modélisation du métabolisme par les **GSMNs** demeure une option de choix pour rationaliser les approches typiques de strain design. En revanche, outre diverses études axées essentiellement sur la croissance des micro-organismes et de la production de bioproduits simples associés, à notre connaissance, peu de travaux ont clairement exploré à grande échelle l'obtention de produits naturels parallèlement à la biomasse en fonction de la disponibilité des nutriments par un organisme en simulant l'environnement de culture dans lequel il évolue.

En conséquence, à travers l'étude du métabolisme du champignon filamenteux *Penicillium rubens* et la reconstruction de son réseau métabolique à l'échelle du génome, nous proposons au cours des travaux présentés dans ce manuscrit de nous concentrer sur la recherche de contraintes environnementales qui seraient susceptibles de contraindre le modèle vers la production de métabolites spécialisés.

3 - *Penicillium rubens* Wisconsin 54-1255, un organisme modèle ?

Cette section vise à introduire l'organisme d'étude en abordant dans un premier temps la notion d'organismes modèles, leur histoire et leurs rôles dans la compréhension des processus biologiques fondamentaux. Nous expliquerons pourquoi les champignons filamenteux sont reconnus aujourd'hui comme des organismes modèles. Nous poursuivrons en donnant une définition générique des champignons avant d'examiner leurs enjeux économiques et leur place dans la recherche scientifique. Nous poursuivrons par une description succincte des caractères morphologiques qui ont guidé leur classification taxonomique. Ce point nous mènera alors directement à la présentation de notre organisme de recherche, la souche *Penicillium rubens* Wisconsin 54-1255. À travers les informations relatives à son histoire, à son statut d'usine de production de produits naturels, et à l'existence de nombreuses données acquises et accumulées après plusieurs décennies d'exploitation, nous justifierons ainsi le choix de cet organisme comme sujet d'étude.

3.1. Les organismes modèles : piliers de la compréhension biologique

La majorité de nos connaissances sur l'hérédité, le développement, la physiologie, le métabolisme, ainsi que sur les processus cellulaires et moléculaires, provient de l'étude d'organismes dits modèles ou de référence *(Müller et Grossniklaus 2010)*. Ces espèces, non humaines, sont étudiées en profondeur pour comprendre des phénomènes biologiques spécifiques, avec l'idée que les résultats obtenus puissent être extrapolés à d'autres organismes. Initialement, le concept d'organisme modèle est donc basé sur le principe de conservation de l'évolution et se différencie, à ce titre, des organismes expérimentaux utilisés en recherche qui incluent des centaines d'espèces qui sont étudiées soit pour explorer un phénomène biologique spécifique, soit pour leur intérêt scientifique propre *(Ankeny et Leonelli 2011)*.

La sélection des organismes modèles a été guidée par leur position phylogénétique et leur adaptabilité expérimentale à des intérêts de recherche spécifiques *(Müller et Grossniklaus 2010)*. Ces organismes ont été choisis pour étudier des processus biologiques complexes, tels que la génétique, le développement, la physiologie, l'évolution et l'écologie. En servant de référence pour des comparaisons inter-espèces et en développant des ressources interdisciplinaires autour d'un organisme spécifique, cette approche favorise les études comparatives à grande échelle *(Leonelli et Ankeny 2013)*.

Assurer à long terme le succès d'un organisme modèle repose sur quelques principes essentiels et incluent la sélection et la préparation de l'organisme (*i.e.* choix de l'organisme, assemblage de son génome et profilage de son expression génique), la simplicité et l'universalité des protocoles expérimentaux (*i.e.* techniques de culture et d'expérimentation), la stabilité génétique assurant une constance des résultats, l'accessibilité aux ressources génomiques et transcriptomiques, ainsi que la disponibilité d'outils d'édition du génome pour réaliser des expériences de perte ou de gain de fonction (*i.e.* manipulation génétique, mutagénèse et ingénierie génétique) *(Matthews et Vosshall 2020)*. La sélection d'organismes appropriés, où l'extrapolation des

connaissances acquises permet d'étendre la compréhension de phénomènes biologiques, se base également sur un socle de critères communs tels que leur petite taille, leur cycle de vie court, leur faible coût d'élevage, leur haute fertilité et leur mutabilité élevée *(Leonelli et Ankeny 2013)*. Enfin, l'un des critères majeurs qui amène à l'attribution du terme modèle réside dans l'appropriation de ces espèces par les communautés de recherche. Inciter les chercheurs à adopter des normes collectives de collaboration et d'investissement dans l'infrastructure sont autant d'aspects qui garantissent une politique coordonnée favorisant le partage immédiat et le maintien des ressources, techniques et données. Cette dynamique sociale de coopération et d'investissement est l'essence même de ce qu'est, et doit être, un organisme modèle *(Leonelli et Ankeny 2013)*.

En dépit de la diversité de la vie, peu d'espèces répondent à l'ensemble des exigences susmentionnées et la plupart des grandes découvertes biologiques du XX^{e} siècle ont été réalisées à l'aide d'un socle commun restreint d'espèces *(Davis 2004)*. En 2010, seulement 13 espèces étaient officiellement reconnues comme « organismes modèles » (cf. encart : *Liste canonique des principaux organismes modèles historiques*, page 63) par le National Institutes of Health *(Leonelli et Ankeny 2013)*. Une décennie plus tard, avec l'essor des technologies de séquençage et l'exploration à grande échelle en transcriptomique et protéomique, le concept même de « modèle de référence » a évolué. En effet, bien que ces organismes modèles hautement standardisés aient grandement contribué à notre compréhension des principes biologiques, ils ne peuvent à eux seuls refléter toute la complexité de la biodiversité *(Goldstein et King 2016)*.

Dès lors, avec l'émergence et l'avènement des outils permettant la génération et l'analyse de données omiques à grande échelle *(Pinu et al. 2019)* les différences de concepts entre espèces expérimentales et modèles s'atténuent *(Leonelli et Ankeny 2013)*. Désormais, l'expression galvaudée « d'organismes modèles » s'élargit pour englober, entre autres, toute espèce dont l'étude approfondie permet d'améliorer la compréhension des mécanismes biologiques essentiels à la santé et aux maladies humaines, incluant l'analyse de la pathogénicité et des processus liés à l'émergence de la résistance *(Bertile et al. 2023)*.

Les espèces appartenant au groupe des champignons filamenteux s'inscrivent pleinement dans cette évolution du concept d'organisme modèle, désormais centré sur la génétique. Ces organismes sont particulièrement adaptés à des expérimentations à grande échelle en raison de l'expression de caractéristiques communes *(Naranjo-Ortiz et Gabaldón 2020)* telles que leur diversité morphogénétique *(Lin et al. 2015)*, leur mode de reproduction à la fois sexuée et asexuée *(Powers-Fletcher et al. 2016)*, leur croissance multicellulaire *(Nagy et al. 2020)*, leur facilité de manipulation et d'édition génomique *(Meyer 2008)*, ainsi que leur capacité à se développer sur tout environnement écologique *(Chambergo et Valencia 2016; Chaudhary et al. 2022)*. Leur utilisation a ainsi permis de surmonter les limites des modèles traditionnels en représentant des aspects spécifiques de la biologie eucaryote tout en offrant une grande flexibilité expérimentale. Par exemple, les découvertes clés réalisées avec des espèces telles que *Neurospora crassa* *(Roche et al. 2014)*, *Aspergillus nidulans* *(Brandl et Andersen 2017)* ou *Penicillium chrysogenum* *(Fierro et al. 2022)* ont non seulement éclairé des mécanismes fondamentaux de la génétique, du cycle cellulaire et de la résistance aux antibiotiques, mais ont également ouvert de nouvelles perspectives en biotechnologie, écologie et pharmacologie. En outre, avec leur diversité métabolique et leur capacité à s'adapter à une variété de conditions environnementales *(Klein*

et Paschke 2004), ces organismes continuent d'élargir notre compréhension des processus biologiques complexes. En effet, ces espèces sont au cœur, entre autres, des études sur la compréhension des interactions hôte-pathogène, la régulation épigénétique ou la production de métabolites spécialisés *(Keller 2019)*. Enfin, leur capacité à produire des composés bioactifs, tels que des antibiotiques, des immunosuppresseurs et des toxines, en font également des modèles d'intérêt cruciaux en pharmacologie et écologie *(Keller et al. 2005)*. De ce fait, les champignons filamenteux, en tant que modèles, participent à repousser les frontières de la recherche biomédicale et environnementale *(Meyer et al. 2020)*, tout en illustrant la valeur de la diversification des organismes étudiés pour une compréhension plus approfondie de la vie.

Liste canonique des principaux organismes modèles historiques (d'après les travaux de *Müller et Grossniklaus 2010)* **:**

Escherichia coli : sa facilité de culture et la simplicité de son génome ont permis de jouer un rôle crucial dans l'élucidation des concepts fondamentaux de la régulation transcriptionnelle et dans la compréhension des processus cellulaires basiques. Les recherches en biologie moléculaire et en biotechnologie ont également largement bénéficié des techniques de transformation et d'expression génétique élaborées et optimisées à partir de cette bactérie.

Schizosaccharomyces pombe et ***Saccharomyces cerevisiae*** : les études portées sur ces deux levures, des organismes unicellulaires, ont permis d'identifier les composantes du cycle de division cellulaire et notamment ses points de contrôle. À noter également que *S. cerevisiae* a été l'un des premiers organismes eucaryote dont le génome a été intégralement séquencé dans le courant des années 1990.

***Neurospora crassa* :** de par son cycle de vie haploïde et sa culture aisée, ce champignon filamenteux a longuement été populaire pour étudier, entre autres, le métabolisme, les mécanismes de régulation et d'intégrité génomique ainsi que les processus de développement biologique. En outre, les expérimentations de mutations génétiques par Beadle et Tatum dans les années 1940 ont mené à l'hypothèse « one gene-one enzyme » selon laquelle chaque gène code pour une enzyme spécifique.

Arabidopsis thaliana : la simplicité de son génome associée à son cycle de développement rapide et sa forte proportion à la production de graines ont établi cet embryophyte au rang de modèle principal du règne végétal. Ce modèle a ainsi permis des avancées significatives en biologie moléculaire, éclairant des processus comme le développement des organes, la biologie des cellules souches, le rythme circadien et les réponses des plantes à des stimuli environnementaux tels que la production de phytohormones.

Zea mays et ***Oryza sativa*** : ces plantes à haut intérêt dans le secteur de l'agriculture ont été historiquement des modèles de choix pour les études génétiques et moléculaires. Les échanges physiques de chromosomes, l'existence des transposons ainsi que l'impact des phénomènes épigénétiques résultent de l'utilisation de ces modèles.

Caenorhabditis elegans : les phénomènes biologiques impliqués dans la mort cellulaire programmée ou les mécanismes de l'interférence ARN (*i.e.* un outil puissant qui permet de générer des mutations ciblées afin de réduire l'expression d'un gène - knock-down) ont été élucidés à partir de ce nématode, l'un des plus petits eucaryotes multicellulaires.

Drosophila melanogaster : au début du XX^{e} siècle, à l'aide de l'élevage facilité de ces mouches, les travaux de Morgan regroupant des concepts de cytologie et de génétique établissent que les gènes sont situés sur les chromosomes, les unités fondamentales de transmission des traits héréditaires.

Danio rerio : de par sa transparence et son développement embryonnaire externe, ce petit poisson consistue un modèle vertébré idéal pour l'étude de la génétique du développement.

> ***Xenopus laevis*** : les manipulations expérimentales des embryons de cet amphibien, facilitées par leur taille et leur accessibilité, ont permis d'établir des principes fondamentaux de la biologie du développement, tels que l'induction neuronale et la formation de motifs lors du développement des appendices.
>
> ***Mus musculus*** : modèle mammifère de premier ordre en raison de ses caractéristiques physiologiques étroitement liées à celles de l'Homme. Outre sa capacité à exprimer des pathologies à haut intérêt en santé humaine (*e.g.* cancer, hypertension, diabète, ostéoporose, *etc.*), la souris constitue un organisme de choix pour l'étude des cellules souches et donc pour les expériences de génétique ciblée.

3.2. Qu'est-ce qu'un champignon ?

3.2.1. Définition populaire et générique

Dans l'imaginaire collectif, les champignons occupent une place significative dans diverses sphères culturelles en raison notamment de leur croissance rapide et fugace et de la variété de formes et de couleurs qu'ils expriment. Outre leur intérêt culinaire, les champignons sont populairement perçus comme des êtres fantastiques, symbolisant mystère, transformation et magie. Présents dans toute forme de divertissements, illustrations, dessins animés, jeux vidéo, peintures, littérature, ils sont également fortement liés aux légendes et aux pratiques folkloriques en évoquant la métamorphose ou les mystères de la nature. En tant que sujets artistiques, leur présence dans l'art peut être à la fois esthétique, symbolique ou conceptuelle en explorant des idées telles que la nature, la croissance, la décomposition ou l'interconnectivité écologique.

Dans le langage courant, outre l'existence d'expressions telles que « croître comme un champignon » ou « champignon atomique » qui soulignent leur mode de croissance, le terme champignon est défini comme étant une « plante charnue, comestible ou non, dont la forme évoque généralement un chapeau muni d'un pied, qui sont les développements extérieurs de l'appareil végétatif* ». Littéralement, le terme champignon est issu du bas latin *campania, -orum, n* ou *campaniia, -ae, f* qui signifie les champs, la plaine† faisant ainsi référence uniquement et une fois encore aux formes macroscopiques observées dans des espaces ouverts. Cependant, ces définitions, d'une part ne couvrent qu'une fraction restreinte de la diversité du règne fongique, et d'autre part, ont englobé des organismes appartenant à des taxons distincts tels que les les oomycètes, les actinomycètes ou les myxomycètes *(Cavalier-Smith 2001)*. Ainsi, le terme même de champignon est alors d'un point de vue scientifique ambigu, voire obsolète.

D'un point de vue historique, les mycètes, ou champignons, sont des organismes eucaryotes reconnus depuis la fin des années 1960 comme un règne distinct des plantes et des animaux *(Whittaker 1969)*. Longtemps assimilé aux végétaux en raison de leur mode de vie sédentaire, ce clade d'organismes multi-variés s'en distingue par l'absence de chlorophylle, principal pigment photosynthétique, et par leur capacité à se nourrir de matière organique. La diversité des espèces de ce règne se manifeste également par la présence d'une paroi cellulaire chitineuse, la perte des capacités phagotrophiques, ainsi que par une variété de structures allant d'organismes unicellulaires à pluricellulaires *(Naranjo-Ortiz et Gabaldón 2019; Nagy et al. 2020)*. En tant qu'hétérotrophes, ces organismes absorbent des molécules organiques synthétisées par d'autres organismes vivants pour puiser leur énergie.

* CNRTL. (s.d.). Champignon. Dictionnaire de l'Académie française (8e édition)

† Gaffiot, F. (1967). *Campania*. Dictionnaire illustré latin-français. Paris : Hachette.

3.2.2. Applications bio-industrielles et exploitations commerciales inhérentes

Les champignons filamenteux présentent des aspects contrastés, tantôt bénéfiques, comme en témoigne leur utilisation dans la production de métabolites utiles ou d'antibiotiques *(Demain et Sanchez 2009)*, tantôt nuisibles, en provoquant des infections ou en se développant de manière invasive *(Egbuta et al. 2017; Hoenigl et al. 2024)*. En raison de leur capacité à décomposer la matière organique et de leur impact dans la chaîne alimentaire *(Grimm et al. 2005)*, ces organismes, utilisés seuls ou en synergie avec d'autres agents biologiques tels que les bactéries *(Peleg et al. 2010; Deveau et al. 2018)*, les algues *(Chu et al. 2021)* ou d'autres espèces fongiques *(Sperandio et Ferreira Filho 2019)* sont largement étudiés et exploités, dans de nombreux secteurs industriels (**Figure 1-6**).

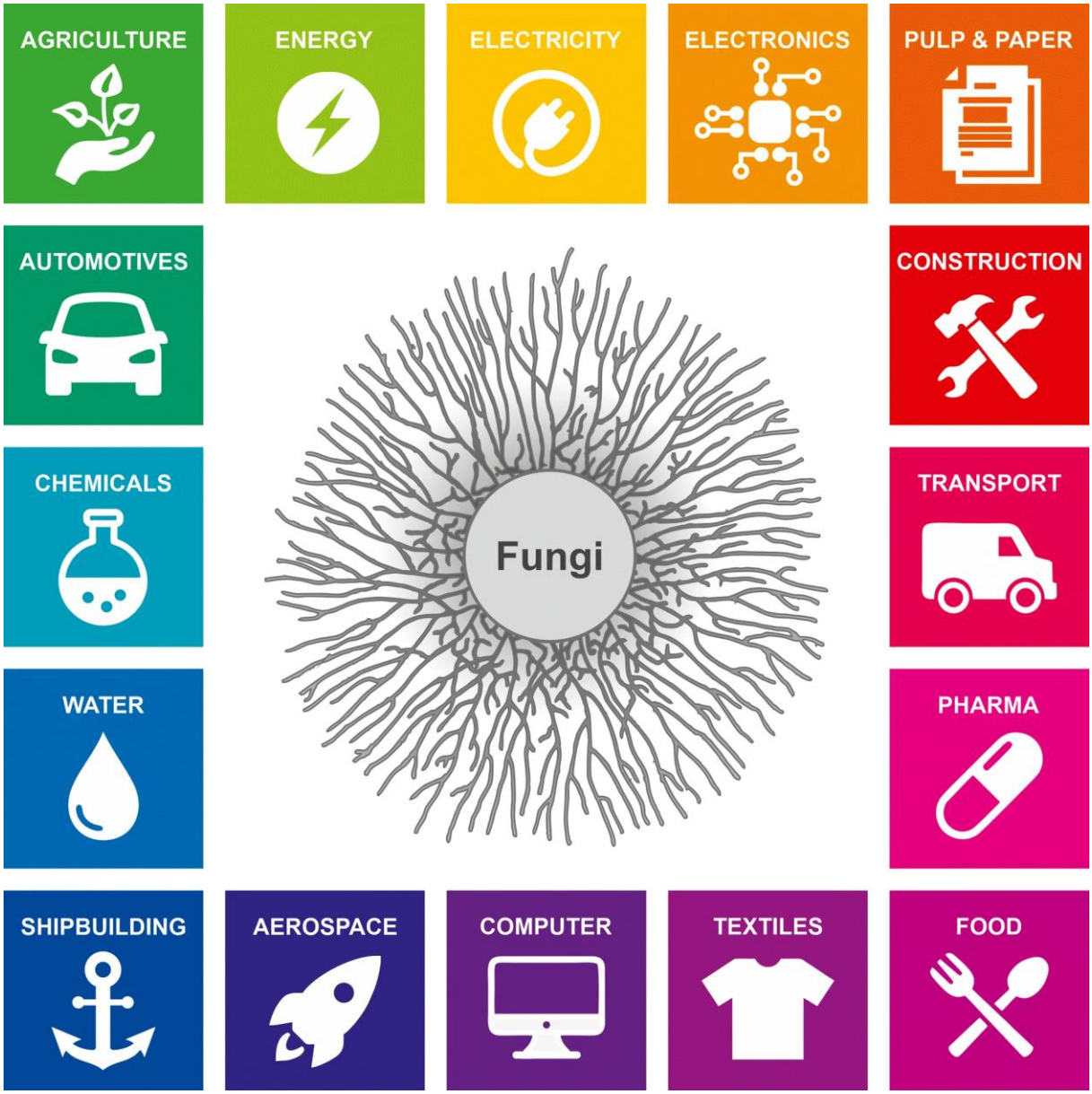


Figure 1-6 : Applications industrielles des capacités métaboliques des champignons filamenteux. *Seules quelques douzaines d'espèces de champignons filamenteux sont exploitées en biotechnologie comme usines cellulaires et leurs produits enzymatiques sont utilisés dans divers secteurs industriels avec des applications variées dans l'agroalimentaire, les biens de consommation, l'énergie et les carburants, la santé et les produits pharmaceutiques, l'industrie chimique, la technologie et l'électronique, ainsi que la construction et le transport. La diversité de ces secteurs souligne la polyvalence et l'importance économique de ces organismes qui contribuent à une économie circulaire. Extrait du white paper : Growing a circular economy with fungal biotechnology (Meyer et al. 2020).*

Au cours des dix dernières années, le nombre de publications scientifiques relative à la biologie synthétique *(Ausländer et al. 2017)* traitant de la biotechnologie des champignons, que ce soit sous les termes « *fungal biotechnology* », « *fungal genomics* », « *fungal gene editing* » ou « *mycoremediation* », n'a cessé d'augmenter, illustrant l'intérêt croissant pour ces organismes *(Roth et al. 2023)*. Leur facilité de manipulation et leur accessibilité en font également des modèles privilégiés dans divers marchés industriels mondiaux, qu'ils soient pharmaceutiques ou alimentaires (cf. encart : *Influence économique et marchés mondiaux : le poids croissant des champignons filamenteux*, page 67).

En raison des nombreux mécanismes exprimés par les champignons filamenteux, qu'ils soient enzymatiques ou non (*e.g.* absorption des composés toxiques, production de biosurfactants, biominéralisation, ou bioprécipitation), ces organismes sont des candidats de premier intérêt pour traiter les questions relatives à une dépollution respectueuse de l'environnement et permettent de restaurer les écosystèmes à moindre coût *(Abatenh et al. 2017; Omokhagbor Adams et al. 2020)*. Ce procédé, connu sous le terme de bioremédiation, utilise des organismes vivants présents naturellement dans l'environnement ou ajoutés intentionnellement pour dégrader, transformer ou éliminer des contaminants émergents, des substances toxiques ou polluantes (*e.g.* hydrocarbures, métaux lourds, solvants organiques, pesticides, produits chimiques industriels, déchets agricoles, déchets pharmaceutiques). Par exemple, de nombreuses espèces de *Penicillium* sont connues pour être des bio-absorbants naturels, efficaces dans la réduction des métaux lourds et des hydrocarbures pétroliers dans l'environnement. Ces organismes ont la capacité de remédier à une large gamme d'hydrocarbures aromatiques polycycliques (*e.g.* pyrène, fluorène) présents naturellement dans le charbon, le pétrole brut et l'essence, contribuant ainsi à l'assainissement de l'environnement *(Ghosh et al. 2023)*.

Répartis en 6 axes - stratégies contre les maladies humaines, stratégies contre les maladies des plantes, amélioration des cultures et de la sylviculture, aliments et boissons à base de champignons, protection de l'environnement et production de biens - la review *The amazing potential of fungi: 50 ways we can exploit fungi industrially (Hyde et al. 2019)* expose les attributs uniques des champignons filamenteux qui contribuent à leur succès dans les domaines de la biotechnologie et de l'industrie. Outre les intérêts pharmaceutiques, largement documentés et évoqués précédemment, nous souhaitons ici souligner la diversité de ces applications avec des exemples sur la production d'enzymes cellulolytiques pour les industries du papier, du textile et du bio-éthanol *(Singh et al. 2021)*, la conception de carburants et de produits chimiques à partir de biomasse lignocellulosique *(Madhavan et al. 2022)*, la production de biopigments et de colorants naturels pour l'alimentation, le textile et les cosmétiques *(Dufossé et al. 2014; Kalra et al. 2020)*, l'utilisation de mycoprotéines comme produits alimentaires pour humains et animaux *(Barzee et al. 2021; Strong et al. 2022)*, la production de biopesticides *(De la Cruz Quiroz et al. 2015)* et la lutte biologique contre les nuisibles *(Baron et al. 2019)*, ainsi que la fabrication de matériaux bio-composites à base de mycélium pour des matériaux de construction écologiques et biodégradables *(Mandal et Krishnan 2021; Sierra et al. 2023)*. Notons à cet égard que les espèces du genre *Penicillium* sont largement représentées dans ces divers secteurs *(Ashtekar et al. 2021)*.

L'attrait et le plébiscite des champignons filamenteux s'expliquent notamment par leur aptitude intrinsèque à sécréter une large gamme de molécules bioactives telles que des peptides antimicrobiens, des hydrophobines, des enzymes hydrolytiques (*e.g.* protéases, amylases, xylanases, cellulases) ainsi que des protéines recombinantes qui en ont fait des organismes de choix pour le développement de « *cell factories* » *(Lübeck et Lübeck 2022)*. Ce concept de biotechnologie, consistant à modifier ou à utiliser des cellules vivantes afin de produire en grandes quantités des substances spécifiques, est appliqué dans divers processus industriels, tels que la production d'enzymes, la bioconversion et la synthèse de produits recombinants. Depuis les années 2000, de nombreuses usines cellulaires fongiques puissantes ont ainsi été développées, où les principaux hôtes appartiennent aux genres *Aspergillus*, *Trichoderma* et *Penicillium,* et dans une moindre mesure aux genres *Myceliophthora*, *Fusarium*, *Rhizopus* ou *Mucor* *(Liu et al. 2023)*. Bien que la production de protéines recombinantes par les champignons filamenteux pour la synthèse de métabolites basaux et spécialisés ait rencontré un succès relatif *(Grimm et al. 2005; Ward 2012)*, l'ère post-génomique a permis de surmonter plusieurs obstacles en offrant une meilleure compréhension des mécanismes cellulaires et génétiques (*e.g.* optimisation des systèmes d'expression, amélioration de la stabilité des protéines produites, réduction de leur dégradation enzymatique). Ces avancées, bien qu'encore perfectibles *(Liu et al. 2023)*, ont renforcé l'efficacité et la viabilité industrielle des champignons filamenteux.

Influence économique et marchés mondiaux : le poids croissant des champignons filamenteux

Pour illustrer l'importance et le poids économique des études sur les champignons filamenteux, cet encart présente cinq grands marchés mondiaux dans lesquels ces organismes interviennent. ***Grand View Research*** est une entreprise américaine qui fournit des rapports détaillés sur les tendances du marché, les analyses de croissance et les prévisions dans divers secteurs industriels, tels que la santé, la technologie, les produits chimiques, et l'énergie. Leurs rapports intègrent des analyses qualitatives et quantitatives, couvrant des aspects tels que la taille du marché, les parts de marché, les tendances actuelles et futures, ainsi que les dynamiques concurrentielles. Les informations chiffrées présentées ci-après proviennent de leurs rapports.

En 2021, la taille du **marché mondial de la biorémédiation** était estimée à 12,38 milliards de dollars américains (USD) avec une croissance projetée de 9,93 % par an jusqu'en 2030. Ce marché se subdivise en six segments technologiques clés : la biostimulation qui vise à améliorer l'activité microbienne naturelle pour dégrader les contaminants, les bioréacteurs qui contrôlent les conditions environnementales pour optimiser ce processus, la bio-augmentation qui implique l'ajout de micro-organismes spécifiques pour accélérer la dégradation des polluants, les traitements basés sur le sol qui regroupent diverses techniques pour la décontamination des sols ainsi que la phytoremédiation et la mycoremédiation qui utilisent les caractéristiques de ces organismes spécifiques pour la décontamination des sols *(Grand View Research 2023a)*.

En 2023, la taille du **marché mondial de la microbiologie et de la culture bactérienne pour les essais industriels** était estimée à 6,74 milliards de dollars américains (USD) avec une croissance projetée de 7,09 % par an jusqu'en 2030. La croissance du marché de la microbiologie est stimulée par les avancées en biotechnologie, qui favorisent l'identification des microbes et l'étude de leur rôle dans la pathogenèse des maladies. La résistance aux antimicrobiens, exacerbée par leur usage excessif dans divers secteurs, pose un problème de santé mondial, menaçant la sécurité alimentaire et hydrique. Cette résistance conduit à l'apparition de souches pathogènes résistantes, compliquant la gestion des maladies d'origine alimentaire. La nécessité de détecter ces agents pathogènes résistants dans les aliments et l'eau pour prévenir les infections alimente également la demande sur ce marché, dont les études sont essentiellement centrées sur les bactéries, les algues et les champignons *(Grand View Research 2023b)*.

En 2022, la taille du **marché mondial des micro-organismes agricoles** était estimée à 6,63 milliards de dollars américains (USD) avec une croissance projetée de 14,2 % par an jusqu'en 2030. Les micro-organismes agricoles, comprenant des bactéries, des champignons et divers microbes, jouent un rôle crucial dans l'agriculture durable et respectueuse de l'environnement en améliorant la croissance des plantes, la fertilité des sols et la protection des cultures contre les ravageurs et les maladies (*e.g.* lutte contre les maladies fongiques des racines ou les invasions parasitaires, *etc.*). Même si en 2022 le segment des bactéries, utilisées notamment pour la formulation de biopesticides, de biofertilisants et de biostimulants a représenté la plus grande part de revenus (52,8 %) le marché devrait connaître une utilisation accrue des champignons en raison de leurs bénéfices pour la germination, le rendement et la floraison. En effet, diverses espèces sont utilisées en raison de leur capacité à effectuer des associations symbiotiques avec les racines des plantes, améliorant ainsi l'absorption des nutriments, d'autres agissent comme des agents de biocontrôle, suppriment les agents pathogènes des plantes et améliorent la santé du sol et de manière générale, les champignons assurent le cycle des nutriments dans l'écosystème en contribuant à la décomposition de la matière organique. *(Grand View Research 2023c)*

En 2023, la taille du **marché mondial des enzymes industrielles** était estimée à 7,42 milliards de dollars américains (USD) avec une croissance projetée de 6,3 % par an jusqu'en 2030. Cette même année, le segment des micro-organismes a dominé ce marché avec la part de revenu la plus élevée, soit 85,49 %. Ces enzymes sont classées selon leur provenance, enzymes bactériennes ou fongiques. Ces organismes sont particulièrement exploités et favorisés en raison de leur faible coût de production et de leur accessibilité pour les fabricants d'enzymes. La majorité des produits résultants sont utilisés dans les détergents, les aliments et les applications médicales. Les enzymes fongiques qui comprennent principalement les phénols oxydases, les estérases et les hydrolases, connaissent une demande croissante, en particulier dans les industries alimentaires pour la préparation de produits tels que la sauce soja, la bière, les produits de boulangerie, les fruits transformés et les produits laitiers *(Grand View Research 2023d)*.

En 2023, la taille du **marché mondial des antibiotiques** était estimée à 50,91 milliards de dollars américains (USD) avec une croissance projetée de 4,2 % jusqu'en 2030. L'augmentation de la prévalence des maladies infectieuses est un facteur majeur contribuant à la croissance de ce marché. Le segment de la pénicilline représentait une fois encore en 2023 la plus grande part de revenus avec 23,8 %. La pénicilline, première classe d'antibiotiques découverte, reste largement utilisée pour traiter diverses infections, notamment celles causées par les staphylocoques, les streptocoques, le Clostridium, et la Listeria. Ces antibiotiques agissent en inhibant la synthèse de la paroi cellulaire ou en empêchant la formation de la couche de peptidoglycane. Représentant la première ligne de traitement, ces médicaments sont essentiels pour le traitement d'infections telles que la pharyngite, les infections cutanées ou la toux bronchique *(Grand View Research 2023e)*.

3.2.3. Taxonomie fongique et caractéristiques morphologiques des *Penicillium*

Depuis la classification Linnéenne du XVIII[e] et son attribution au « *Regnum Vegetable* », la taxonomie fongique n'a eu de cesse d'évoluer *(Hibbett et al. 2007)* au gré des connaissances acquises sur les caractéristiques morphologiques, reproductives, moléculaires et génomiques de ces organismes *(Naranjo-Ortiz et Gabaldón 2019)*. À ce jour, la classification la plus récente se divise en neuf grandes lignées qui ensemble forment le clade monophylétique des « champignons vrais » : *Opisthosporidia*, *Chytridiomycota*, *Neocallimastigomycota*, *Blastocladiomycota*, *Zoopagomycota*, *Mucoromycota*, *Glomeromycota*, *Ascomycota* et *Basidiomycota* *(Naranjo-Ortiz et Gabaldón 2019)*. Les ascomycètes, en représentant plus de 74 % des plus de 8 600 genres du règne fongique, constituent la divison qui exprime le plus de diversité d'espèces et de caractéristiques *(Adl et al. 2019)*. Notre attention se porte ici exclusivement sur les champignons filamenteux *P. rubens* et *P. chrysogenum*, dont la taxonomie est présentée dans le **Tableau 1-1**. Il convient de noter à cet égard que la souche Wisconsin 54-1255, longtemps considérée comme appartenant à l'espèce *P. chrysogenum*, a été reclassée sous *P. rubens* (cf. section *3.3.1* *Historique de la production de pénicilline : de la découverte à l'industrialisation*, page 72).

Tableau 1-1 : Classification taxonomique des espèces *Penicillum chrysogenum* et *rubens*.

Rang Taxonomique	Taxons	Étymologie et description
Règne	*Fungi*	Du latin ***fungus, -i, m*** signifiant « champignon ». Ce terme désigne un vaste groupe d'organismes eucaryotes hétérotrophes, comprenant les levures, les moisissures et les champignons filamenteux qui se nourrissent par absorption.
Division	*Asco**mycota****	Du grec **askós - ἀσκός, οῦ (ὁ)** signifiant « outre ou sac », en référence à la structure en forme de sac, les asques où sont produites les spores caractéristiques de ce groupe.
Sous-division	*Pezizo**mycotina****	Du grec **pezis - πέζις, ιος (ἡ)** désignant un type de champignon à pied réduit c'est à dire sans stipe, et de l'adjectif **pezos - πεζός, ή, όν** signifiant « qui ne s'élève pas de terre », en référence à la morphologie de ces champignons.
Classe	*Eurotio**mycete****	Du grec **eurys - εὐρύς, εὐρεῖα, εὐρύ** signifiant « large, qui s'étend en largeur », en référence au mode de croissance de ces organismes. Les taxons de cette classe et de cet ordre dérivent du nom de genre *Eurotium*, néologisme provenant probablement de ce mot grec.
Ordre	*Eurotiale**	
Famille	*Trichocomaceae**	Du grec **trikhós - τριχός, θριξί (ἡ)** signifiant « cheveu », et du verbe latin ***como, -are (coma)***, signifiant « être chevelu », en référence à l'apparence filamenteuse des hyphes de ces champignons.
Genre	*Penicillium*	Du latin ***penicillum, -i, n*** qui signifie « petit pinceau », en référence à la forme des structures reproductives.
Espèce	*Chrysogenum*	Du grec **khrusós - χρυσός, οῦ (ὁ)** qui signifie « or » et du suffixe latin ***genus, -eris, n*** signifiant « origine, extraction, naissance », en référence à la production d'un pigment jaune doré.
	Rubens	Du latin ***rubeo, -bui, -ere (ruber)*** signifiant « être rouge », en référence à la couleur des colonies de cette espèce. *Rubens* et le participe présent de ce verbe

***** En taxonomie fongique, la classification des champignons est basée sur l'utilisation de suffixes spécifiques qui sont associés à chaque rang taxonomique**.** Les suffixes **-mycota**, **-mycotina** et **-mycetes** issus du grec ancien **múkês - μύκης, ητος (ὁ)** signifiant « champignon », sont utilisés respectivement pour désigner les divisions, les sous-divisions et les classes. Les suffixes latin *-ales* (« qui ressemble à ») et *-aceae* (« semblable à ») sont utilisés pour former respectivement les noms des ordres et des familles fongiques.

Sources - les étymologies présentées dans ce tableau sont extraites des dictionnaires suivants :

Gaffiot, F. (1967). Dictionnaire illustré latin-français. Paris : Hachette.

Le Bailly, A. (2020). Dictionnaire grec-français Le Bailly. Paris : Éditions Les Belles Lettres.

Les champignons filamenteux qui se distinguent par leur mode de croissance basé sur l'extension apicale de filaments appelés hyphes *(Nagy et al. 2020)*, optimisent l'exploitation du substrat environnant en maximisant l'absorption des nutriments. Ce processus est facilité, d'une part par la multiplication des hyphes, structures multicellulaires formant un réseau complexe et filiforme appelé mycélium *(Powers-Fletcher et al. 2016)*, et d'autre part, par l'excrétion d'exoenzymes qui catalysent la dégradation de molécules complexes telles que la lignine et la cellulose en molécules plus simples comme le glucose, ré-absorbées ensuite par les hyphes.

Traditionnellement, la classification des espèces de *Penicillium,* et plus largement des organismes de l'ordre des Eurotiales, reposait sur l'analyse des caractères phénotypiques *(Tsang et al. 2018; Houbraken et al. 2020)*, conduisant parfois à une multiplicité de noms pour une même souche notamment en raison de la variation des formes observées. En tant qu'ascomycètes, ces champignons présentent une reproduction à la fois asexuée, *via* la production de conidies, et sexuée, par la formation de spores méiotiques dans les asques *(Powers-Fletcher et al. 2016)*. Le recours à des noms différents pour désigner les formes anamorphe (*i.e.* phase asexuée), télomorphe (*i.e.* phase sexuée) ou holomorphe (*i.e.* ensemble des phases du cycle de vie) des champignons a alors été source de confusion dans l'établissement des nomenclatures *(Frisvad 2015)*. Les avancées en séquençage moléculaire, en concepts phylogénétiques et en profilage des métabolites ont depuis remis en question les classifications exclusivement phénotypiques. Ainsi, toute espèce présentant des conidophores en forme de pinceaux n'est plus systématiquement assimilée à des espèces de *Penicillium (Visagie et al. 2014; Houbraken et al. 2020)*.

Ces nouvelles techniques ont non seulement renforcé l'identification des espèces de *Penicillium*, mais ont également permis de reclasser et de délimiter les espèces connues, les espèces cryptiques et les complexes d'espèces en sections et clades bien définis *(Tsang et al. 2018)*. Abandonnant la multiplicité des noms pour un même organisme *(Hawksworth et al. 2011)* conformément aux recommandations du Code international de nomenclature pour les algues, les champignons et les plantes *(Turland et al. 2018)*, une réforme de la classification de *Penicillium* a été initiée au début des années 2010 *(Houbraken et Samson 2011)*. Tout comme pour les *Aspergillus*, les classifications relatives aux *Penicillium* évoluent rapidement et sont donc fréquemment révisées *(Visagie et al. 2014; Tsang et al. 2018)*. En 2020, les derniers travaux de *Houbraken et al. (2020)* ont établi que le genre *Penicillium* est désormais subdivisé en deux sous-genres, 32 sections et 89 séries, recensant ainsi 483 espèces.

Pour conclure, et afin d'illustrer les caractéristiques morphologiques macroscopiques et microscopiques ainsi que les différences phénotypiques résultant des conditions de croissance, la **Figure 1-7** expose diverses photographies des colonies et des structures reproductives de *Penicillium rubens*.

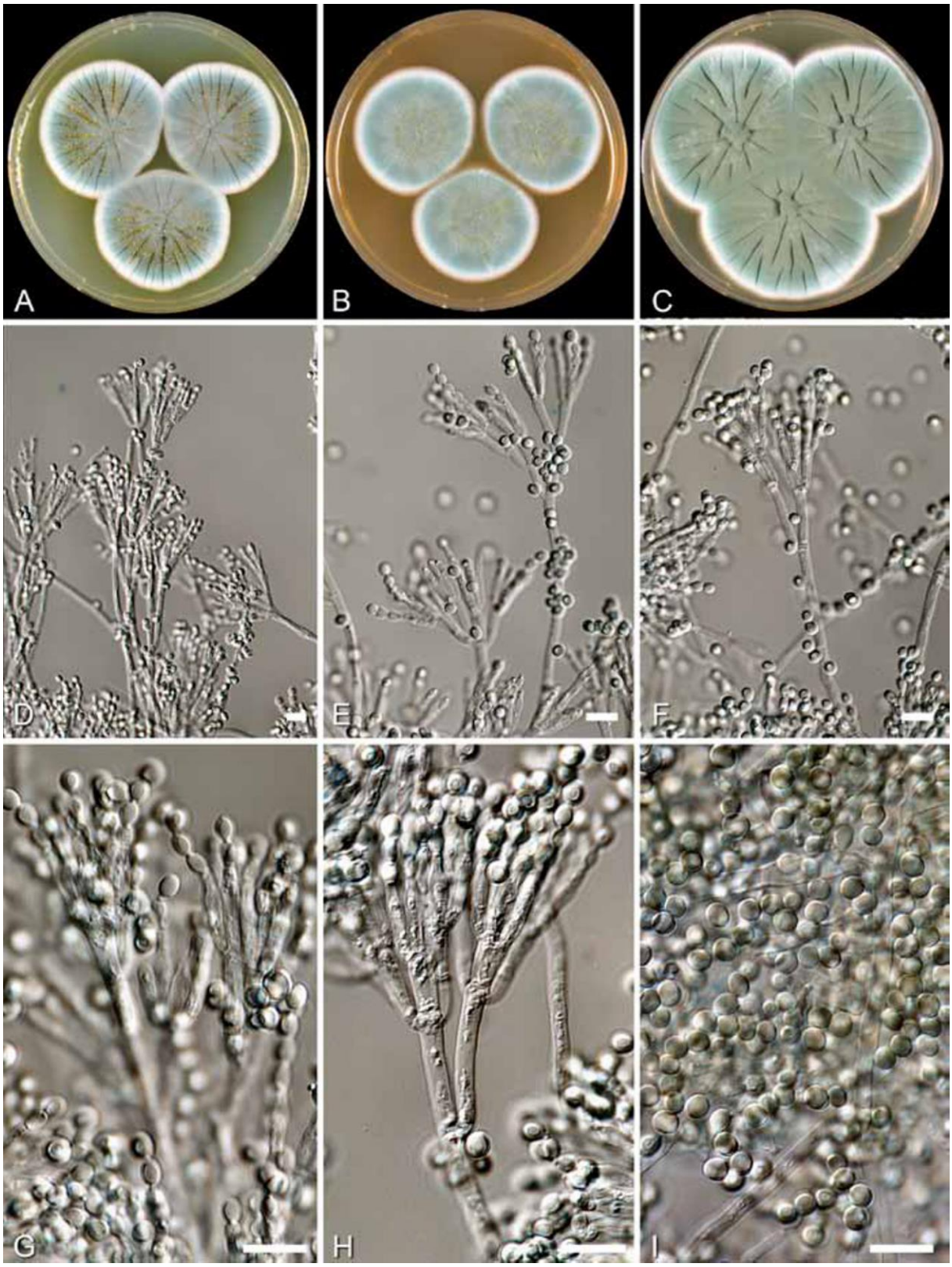


Figure 1-7 : Penicillium rubens : colonies en croissance sur divers milieux de culture et observation de la morphologie microscopique. *Photographies de la souche CBS 205.57, premier producteur de pénicilline identifié par Fleming et proche parente phylogénétique de la souche Wisconsin 54-1255.* ***(A-C)*** *Colonies à 7 jours cultivées à 25 °C sur différents milieux : Czapek Yeast Autolysate Agar* ***(A)****, Malt Extract Agar* ***(B)*** *et Yeast Extract Sucrose Agar* ***(C)****.* ***(D-H)*** *Conidiophores - structures fongiques spécialisées, généralement érigées, sur lesquelles se forment les conidies.* ***(I)*** *Conidies - spores asexuées produites en chaînes au sommet des conidiophores, responsables de la dissémination et de la reproduction du champignon. La barre horizontale sous les photos D à I représente 10 micromètres. Figure et légende issues et traduite des travaux de (Houbraken et al. 2011).*

3.3. *Penicillium rubens* Wisconsin 54-1255

3.3.1. Historique de la production de pénicilline : de la découverte à l'industrialisation

La découverte fortuite de la pénicilline par le bactériologiste britannique Alexander Fleming a marqué un tournant décisif dans l'histoire de la médecine. En observant par hasard que des colonies de *Staphylococcus aureus* sur une boîte de Pétri avaient été « détruites » par un contaminant fongique, Fleming établit formellement que la production de cette substance antimicrobienne n'est pas commune à l'ensemble des *Penicillium*, et l'attribue à *P. rubrum* en raison de caractéristiques morphologiques communes. La description qu'il faisait de ces colonies était alors en ces termes : « *The colony appears as a white fluffy mass which rapidly increases in size and after a few days sporulates, the centre becoming dark green and later in old cultures darkens to almost black. In four or five days a bright yellow colour is produced which diffuses into the medium. In certain conditions a reddish colour can be observed in the growth.* » *(Fleming 1929)*.

Cependant, bien que Fleming ait observé que la pénicilline inhibait la croissance bactérienne, il rencontre des difficultés pour purifier le composé en raison de son instabilité et des faibles quantités produites *(Fierro et al. 2022)*. Ce n'est qu'en 1932 que ce composé actif est correctement identifié comme étant la pénicilline, mais le manque de progrès dans la production de quantités suffisantes freine son développement clinique *(Barreiro et al. 2012)*. L'intérêt pour la pénicilline décline au début des années 1930, en partie en raison du désintérêt croissant pour les thérapies antimicrobiennes innovantes à cette époque. Durant près d'une décennie, les efforts pour isoler et développer la pénicilline en tant que médicament thérapeutique stagnent *(Gaynes 2017)* et malgré son intérêt initial, Fleming se concentre sur d'autres projets et envoie des échantillons de *Penicillium* à d'autres chercheurs dans l'espoir qu'ils poursuivent ses travaux *(Gaynes 2017)*.

Par la suite, de nombreux scientifiques ont contribué à la stabilisation et à la production de masse de la pénicilline. Parmi les plus importants, citons Ernst Chain et Howard Florey, qui ont partagé avec Fleming le prix Nobel de médecine en 1945 *(Ligon 2004)*. Dés 1940 un groupe de chercheurs de l'université d'Oxford sous la supervision de Florey a pu entreprendre des études approfondies sur la pénicilline *(Barreiro et al. 2012)*. En réalisant des expériences sur des souris auxquelles ils injectèrent une souche virulente de streptocoque et en observant que seules celles traitées par la pénicilline survivaient, ils ont pu mettre en évidence le rôle d'agent chimiothérapeutique de cette molécule *(Chain et al. 1940)*. Ils décrivirent alors la production, la purification et l'utilisation expérimentale de pénicilline suffisamment puissante pour protéger les animaux infectés par *Streptococcus pyogenes*, *S. aureus* et *Clostridium septique* *(Gaynes 2017)*.

Suite à ces éléments, les premiers tests de l'efficacité cliniques de la pénicilline sur la santé humaine ont été effectués en février 1941 à Oxford. Ces essais rencontrèrent un succès relatif en raison du faible stock de pénicilline à leur disposition *(Gaynes 2017)*. Les sociétés pharmaceutiques britanniques, contraintes par leurs engagements pendant la Seconde Guerre mondiale, ne pouvaient produire de la pénicilline en masse. L'équipe de Florey s'est donc tournée vers les États-Unis pour obtenir du soutien et après leur entrée en guerre, le gouvernement américain a pris en charge la production de pénicilline *(Ligon 2004)*. La coopération entre ces deux nations aura permis des améliorations majeures dans la production de grande quantité

d'organismes producteurs et donc de quantité de pénicillines produites *(Barreiro et al. 2012)*. À la fin de l'année 1941, le stock de pénicilline était insuffisant pour le traitement d'un seul patient, l'année suivante, les États-Unis disposaient d'un stock pour soigner au moins 100 patients et en septembre 1943 les stocks répondaient aux besoins des forces armées alliées *(Ligon 2004; Gaynes 2017)*. La coopération entre le Royaume-Uni et les États-Unis n'a pas été la seule initiative internationale visant à produire cet antibiotique durant la Seconde Guerre mondiale. Le Canada a également émergé comme un producteur prospère, jouant un rôle clé dans l'approvisionnement des Alliés en pénicilline *(Burns 2006)*. D'autres nations, telles que la France, l'Allemagne et le Japon, ont mené des recherches pour développer leurs propres capacités de production, mais en vain *(Burns 2006)*. En revanche, aux Pays-Bas, des recherches secrètes ont été conduites dans la ville de Delft dès le début de l'occupation nazie en mai 1940. Bien que les chercheurs néerlandais aient été coupés du monde extérieur et que les publications sur la pénicilline aient été soumises à un embargo par les Alliés à partir de 1943, ces efforts clandestins ont permis, après-guerre, de progresser dans la compréhension du processus de production de cet antibiotique *(Burns 2006)*.

La production de pénicilline n'a depuis cessé de croître puisque en 1949, la production commerciale aux États-Unis, alors principal producteur mondial, dépassait les 83 tonnes, en 1982, la production mondiale atteignait plus de 12 000 tonnes, l'Europe devenant le principal producteur et en 1995, la production mondiale totale était estimée à environ 33 000 tonnes *(Elander 2003)*. Toutefois, comme en témoignent les titres de pénicillines des premières souches utilisées – les travaux de *(Moyer Andrew J. et Coghill Robert D. 1946)* rapportaient une production de pénicilline comprise entre $8{,}4.10^{-3}$ et $1{,}8.10^{-2}$ g/L pour la souche originelle de Fleming (NRRL 824, également appelée NRRL 1249), contre des valeurs allant de $8{,}4.10^{-3}$ à $3{,}2.10^{-2}$ g/L pour la première souche commerciale utilisée en culture submergée (NRRL 832, isolée en 1936 et indépendante de la souche de Fleming) *(Raper et al. 1944)* – le principal obstacle de l'époque demeurait la difficulté d'obtenir suffisamment de pénicilline purifiée *(Fierro et al. 2022)*. Ainsi la recherche de nouvelles souches naturelles plus productives a été engagée *(Barreiro et al. 2012)*.

En 1943, ces efforts aboutissent avec la découverte de la souche naturelle NRRL 1951, isolée d'un cantaloup moisi sur un marché local à Peoria, dans l'Illinois *(Barreiro et al. 2012)* et dont les capacités de. production de pénicilline surpassaient toutes les souches étudiées jusqu'alors *(Fierro et al. 2022)*. Dès lors, cette souche a été soumise à un intense programme d'amélioration classique, impliquant des phases de sélection de variants spontanés et des cycles répétés de mutagenèse aléatoire, soit par exposition aux rayons X, soit par irradiation ultraviolette, soit par l'ajout d'agents chimiques comme la moutarde azotée. Les premières étapes de ce processus, menées notamment au NRRL (*Northern Regional Research Laboratory*) et à l'Université du Wisconsin, ont permis d'obtenir des souches telles que NRRL 1951.B25, X-1612 et Wisconsin Q176 *(Salo et al. 2015; Fierro et al. 2022)*. Il est à remarquer également que la production de pénicilline n'est pas une caractéristique exclusive des souches de *Penicillium* et que des traces de cette molécule ont été détectées chez diverses espèces d'*Aspergillus* ou chez *Saccharomyces cerevisiae* *(Van den Berg 2011)*. Néanmoins, le développement de la croissance submergée et les cycles de mutagénèse répétitive ont eu un impact bien plus significatif sur les souches citées précédemment qui sont alors devenues des références industrielles *(Van den Berg 2011)*.

Notons à cet instant que les questions relatives à l'identification taxonomiques de l'espèce de ces souches de *Penicillium* ont fait l'objet d'une attention particulière non seulement en raison de leur rôle dans la production de pénicilline, mais aussi à cause de leur capacité à contaminer les environnements intérieurs et les denrées alimentaires *(Houbraken et al. 2012)*. Initialement, la souche productrice de pénicilline de Fleming a été identifiée comme *P. rubrum (Fleming 1929)*, mais l'évolution des schémas taxonomiques a conduit à des reclassifications diverses au fil des ans et cette souche s'est retrouvée successivement sous les noms, parfois synonymes, de *P. notatum (Thom 1945)*, *P. chrysogenum (Samson et al. 1977)*, *P. griseoroseum (Pitt 1981)* et à nouveau *P. chrysogenum (Kozakiewicz et al. 1992)*. En revanche, la souche NRLL1951, utilisée pour la production industrielle de pénicilline, a rapidement été identifiée comme un *P. chrysogenum (Stauffer et Backus 1954)*. Cette appellation est demeurée la plus pertinente jusqu'aux travaux d'Houbraken *et al. (2010)* au début des années 2010. En effet, comme les examens phénotypiques, tant macroscopiques et microscopiques ne permettent pas de détecter des différences entre les espèces *chrysogenum* et *rubens*, seule l'analyse de leurs extrolites permettent de les différencier (*e.g.* seuls les *P. chrysogenum* ont la capacité à produire l'acide sécalonique) *(Houbraken et al. 2011)*. Dès lors, les souches des lignées dites de Fleming et de producteurs de pénicilline ont été reclassées en *P. rubens*. Ainsi, et afin d'éviter toute confusion, nous ne mentionnons uniquement que le nom des souches dans cette section.

À partir de 1943, avec l'isolement de la souche NRLL 1951, plusieurs entreprises pharmaceutiques se sont rapidement intéressées à la production et à la commercialisation de cet inhibiteur antibactérien, développant chacune leur propre programme d'amélioration des souches (*i.e. « classical strain improvment »*) de *Penicillium* en vue d'une production industrielle de pénicilline. Selon l'origine des laboratoires industriels, diverses lignées ont émergé. Par exemple, les laboratoires Panlabs Inc. (Taiwan) ont développé les souches P1 et P2, les ancêtres de P2niaD18 *(Lein 1986; Fierro et al. 1995; Specht et al. 2014)*; DSM Biotechnology (Pays-Bas) a produit des souches dérivées de DS04825 à l'origine des mutants surproducteurs actuellement utilisés *(Gouka et al. 1991; Kiel et al. 2005; Nijland et al. 2010; Viggiano et al. 2018)* ; Antibióticos S.A. (Espagne) a mis au point les souches AS-P-78, AS-P-99 et E1 *(Barredo et al. 1989; Fierro et al. 1995)* ; Smith Kline Beecham (Royaume-Uni) a introduit la famille BW 1890 *(Newbert et al. 1997)* ; et North China Pharmaceutical Group Corporation (Chine) a cultivé la souche NCPC10086 *(Wang et al. 2004, 2014)*.

Le point commun de toutes ces souches industrielles résulte dans leur qualificatif à un moment de l'Histoire de « souches industrielles hautement productrices de pénicilline ». Toutefois, définir quantitativement ce haut degré de variation de production entre ces souches s'avère complexe, et à notre connaissance, de telles comparaisons sont rares et difficilement accessibles, d'une part par le manque d'homogénéité des protocoles de cultures qui ont évolué au fil du temps *(Elander 2003),* et d'autre part, en raison du statut même de ces souches. Leur nature industrielle limite l'accès aux données sur les programmes d'amélioration des souches *(Barreiro et al. 2012)*, rendant la traçabilité exhaustive difficile *(Salo et al. 2015)*. Nous présentons néanmoins en **Figure 1-8** une synthèse des éléments accessibles dans la littérature scientifique et quelques données chiffrées sur les productions de pénicilline par souche. À titre d'exemple, les origines exactes de la majorité des souches industrielles demeurent obscures et sont généralement

attribuées à la souche sauvage NRLL 1951 *(Fierro et al. 2022)*. Toutefois, et même si à ce jour la séquence du génome de la souche sauvage NRRL 1951 n'a toujours pas été intégralement publiée *(Martín 2020)*, les connaissances se déverrouillent progressivement depuis la fin des années 1990 avec la mise à disposition pour les laboratoires de recherche de données génomiques partielles ou complètes sur certaines souches, conduisant à diverses publications scientifiques de génomique comparative *(García-Estrada et al. 2020)*.

En plus de 70 ans de recherche, les divers cycles de mutagenèse aléatoire appliqués aux souches de *P. rubens* ont conduit à l'accumulation de mutations ponctuelles *(Salo et al. 2015)* qui ont permis la sélection de souches optimisées capables de produire des titres toujours plus élevés de pénicilline dans des fermenteurs à grande échelle *(Van den Berg 2011)*. Cependant, cette sélection s'est faite au détriment de la production d'autres métabolites spécialisés produits originellement par la souche sauvage NRRL 1951 *(García-Estrada et al. 2020)*. Plus précisément l'analyse du protéome *(Jami et al. 2010)* et l'étude de ces mutations *(Salo et al. 2015)* ont révélé la formation de protéines non fonctionnelles ainsi que des niveaux d'expression réduits pour certains gènes liés au métabolisme spécialisé, notamment divers groupes de gènes de polycétides (PK) et de peptides non ribosomiques (NRP), partiellement ou complètement réduits au silence *(Martín 2020)*. L'exemple le plus représentatif de ce phénomène est la perte progressive et volontaire de production de pigments, fortement abondants dans la souche sauvage *(Barreiro et al. 2012)*. La souche BL3-D10 isolée en 1947 a été la première qui ne présentait pas la couleur jaune caractéristique de ces pigments, qui ont récemment été identifiés comme étant un mélange de sorbécillinoïdes *(Martín 2020)*. Elle a alors été sélectionnée à dessein pour des raisons commerciales car les processus d'extraction pigmentaires réduisaient les rendements de pénicilline *(Barreiro et al. 2012)*. Aujourd'hui, en raison du large spectre de bio-activité associé à ces molécules et de leur potentiel pharmaceutique, la capacité de production de ces composés a été réhabilitée chez les nouvelles souches industrielles *(Salo et al. 2016; Guzmán-Chávez et al. 2017)*.

D'autres ré-équilibrages métaboliques ont également été observés puisque les souches sauvages présentent un nombre plus élevé de gènes associés à la pathogénicité et à la virulence que ceux des souches industrielles *(Peng et al. 2014)* et, de manière plus générale, comme la synthèse de pénicilline s'accompagne d'une forte consommation d'ATP, de nouveaux équilibres métaboliques ont été observés *(Van den Berg 2011; Barreiro et al. 2012; Pohl et al. 2020)*. Enfin, un des résultats les plus remarquables des cycles de mutagenèse est l'amplification indépendante du groupe de gènes biosynthétiques de la pénicilline dans toutes les lignées de souches commerciales *(Martín 2020)*. Cette observation a conduit à l'hypothèse largement acceptée selon laquelle les titres élevés de pénicilline résultent de la présence de multiples copies des gènes de la pénicilline. Toutefois, l'augmentation du titre de pénicilline n'est pas strictement corrélée au nombre de copies du cluster *(Nijland et al. 2010; Ziemons et al. 2017)* et une relation non linéaire ayant été décrite entre le nombre de copies de gènes, les niveaux de transcrits, les niveaux d'enzymes et la productivité en pénicilline, suggère la présence de mutations au rôle régulateur *(Van den Berg 2011; García-Estrada et al. 2020)*.

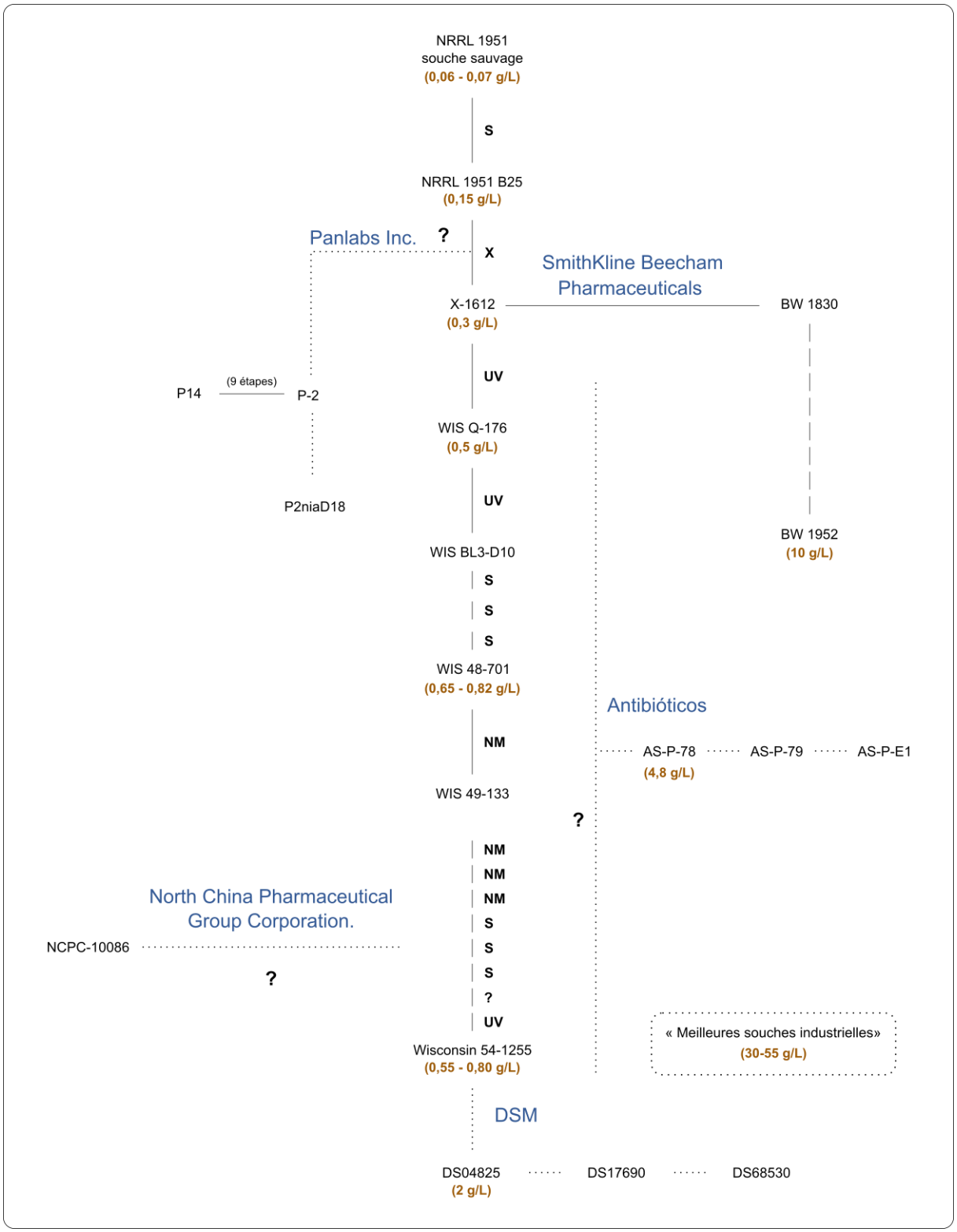


Figure 1-8 : Évolution historique des souches productrices de pénicilline et des quantités produites (en g/L). *Depuis l'identification de la souche sauvage NRLL 1951 en 1943, prometteuse en termes de production de pénicilline, diverses entreprises pharmaceutiques ont développé leurs programmes d'amélioration de souches, dont cinq lignées indiquées en bleu sont présentées sur cette figure. En raison de la confidentialité des programmes industriels l'origine et le développement de certaines lignées restent cependant peu documentés et cette absence d'information est matérialisée par les lignes pointillées. À l'inverse, les lignes pleines correspondent à des étapes connues et référencées dans la littérature. Lorsqu'ils ont été communiqués, les traitements appliqués aux souches sont précisés : S : sélection spontanée sans mutagenèse ; X : traitement par rayons X ; UV : irradiation par UV ; NM : traitement à la moutarde azotée, un agent mutagène chimique. Les informations de cette figure sont tirées et adaptées des articles suivants (Newbert et al. 1997; Rodríguez-Sáiz et al. 2001; Barreiro et al. 2012; Salo et al. 2015; Martín 2020). Les rendements en pénicilline sont fournis en g/L, obtenus par conversion des unités internationales (UI/mL) en g/L, selon la relation suivante : 1 UI/mL = 6 × 10^{-4} g/L. Les données chiffrées sont issues d'un consensus des publications clés (Van den Berg 2011; Barreiro et al. 2012; Reddy et al. 2012).*

En résumé, les souches terrestres de *Penicillium rubens*, qu'elles soient industrielles *(Bhadury et al. 2006; Van den Berg et al. 2008; Specht et al. 2014; Wang et al. 2014; Pohl et al. 2020)* ou sauvages *(Peng et al. 2014; Gujar et al. 2018)*, ont été largement étudiées au cours du dernier siècle, confirmant ainsi le statut de *P. rubens* comme espèce modèle parmi les champignons filamenteux. Aujourd'hui encore, les antibiotiques dérivés des *β*-lactames, dont les pénicillines sont les sont les plus emblématiques, demeurent parmi les médicaments les plus exploités et administrés, occupant une place prépondérante sur les marchés mondiaux des antibiotiques (cf. encart : *Influence économique et marchés mondiaux : le poids croissant des champignons filamenteux*, page 67). La production commerciale de pénicilline a connu un développement industriel considérable en plus de 70 ans d'exploitation. Pour obtenir de tels résultats, de nombreuses améliorations ont été apportées à la technologie de fabrication, tant au niveau des procédés (*e.g.* culture en fed-batch, stérilisation continue des milieux, réduction des étapes d'extraction, *etc.*) qu'au niveau des substrats de croissance (*e.g.* remplacement du lactose par le glucose ou le sucrose) *(Elander 2003)*, avec comme innovation majeure la sélection de souches de production toujours plus performantes. Aujourd'hui, la majorité des souches industrielles à haut rendement en pénicilline proviennent de la lignée Wisconsin 54-1255. Utilisée comme norme industrielle, cette souche améliorée est la référence mondiale pour les recherches sur la biosynthèse de la pénicilline et, plus largement, sur les processus biochimiques, génétiques et cellulaires de *P. rubens* *(Salo et al. 2015; Fierro et al. 2022)*. Historiquement, la découverte de la pénicilline et le développement des méthodes de production industrielle en masse ont marqué l'un des plus grands progrès médicaux, caractérisé par des découvertes fortuites, l'influence du contexte de la Seconde Guerre mondiale, l'attribution du prix Nobel de médecine, des enjeux industriels, et surtout, une collaboration sans précédent entre chercheurs académiques et industriels *(García-Estrada et al. 2020)*.

3.3.2. Usine de production de produits naturels et modélisation du métabolisme

Les *Penicillium* sont des organismes pourvoyeurs d'une grande diversité et quantité de médicaments qui incluent, certes, les antibiotiques de la classe des pénicillines permettant de traiter diverses infections bactériennes (*e.g.* ampicilline, amoxicilline, dicloxacilline, flucloxacilline, ticarcilline, pipéracilline, benzylpénicilline, phénoxyméthylpénicilline) mais également les antifongiques (*e.g.* griséofulvine), les immunosuppresseurs (*e.g.* mycophénolate mofétil administrés pour prévenir le rejet d'organes après une transplantation), les anticoagulants (*e.g.* warfarine) ou encore les statines (*e.g.* simvastatine, pravastatine, fluvastatine, atorvastatine, rosuvastatine) utilisées pour abaisser les niveaux de cholestérol dans le sang et réduire le risque de maladies cardiovasculaires *(Ashtekar et al. 2021)*. Les espèces de *Penicillium* sont ainsi perçues comme de véritables usines de production de métabolites spécialisés aux applications médicales de premier plan.

Ces développements trouvent leur origine dans l'intérêt scientifique pour le genre *Penicillium* qui s'est accru au cours des deux derniers siècles, comme en témoignent les volumes croissants de publications dans les bases de données Scopus et PubMed (**Figure 1-9**). Avec plus de 103 000 documents référencés dans Scopus sous le mot-clé « *Penicillium* » et 18 000 dans PubMed, l'étude de ces champignons a joué un rôle central dans la recherche scientifique, particulièrement dans les sciences de la vie et les disciplines biomédicales. Ce volume notable de publications reflète une fois de plus, l'adoption de ces espèces comme organismes modèles par la communauté scientifique, intérêt qui n'a cessé de croître au fil du temps.

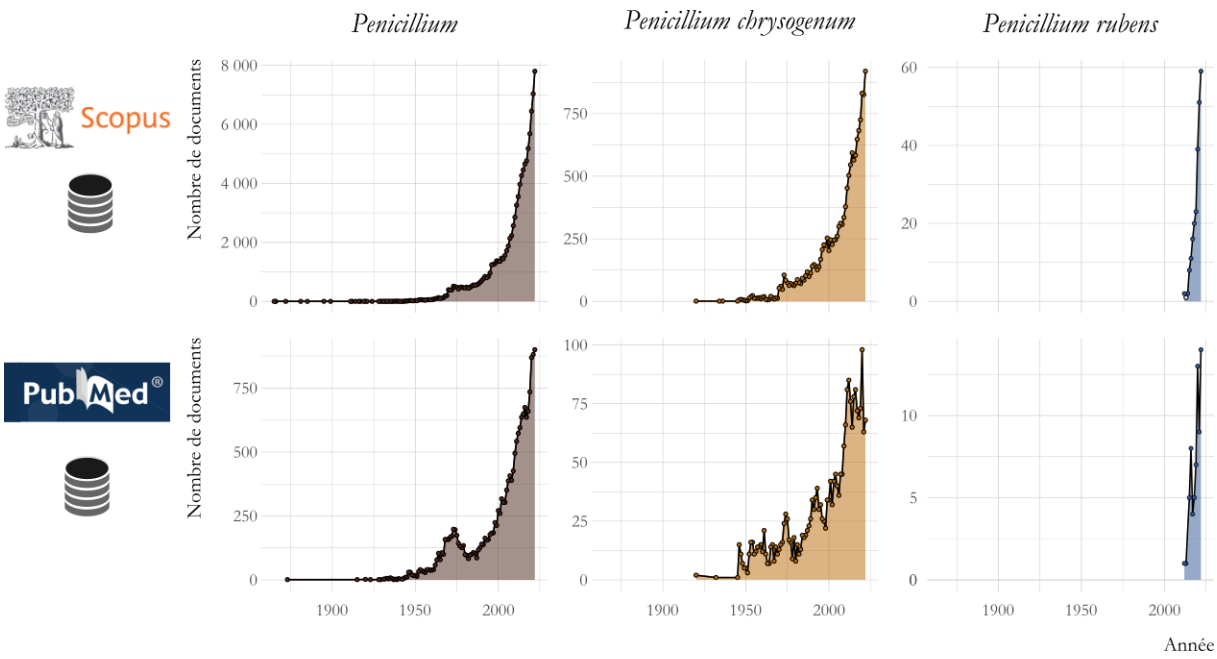


Figure 1-9 : Nombre de documents publiés annuellement et répertoriés dans les bases de données bibliographiques Scopus et PubMed, de 1865 à 2022, à partir de la recherche des mots clés « Penicillium », « Penicillium chrysogenum » et « Penicillium rubens ». *Ces deux bases de données diffèrent notamment en termes de couverture puisque PubMed est axée principalement sur les sciences de la vie et les disciplines biomédicales, alors que Scopus est multidisciplinaires. Ainsi, un total de 103 291 documents couvrant la période de 1865 à 2022 pour Scopus et 18 015 documents pour la période de 1873 à 2022 pour PubMed sont détectés à l'aide du mot-clé « Penicillium ». Les expressions « Penicillium chrysogenum » et « Penicillium rubens » font, quant à elles, leur apparition respectivement en 1920 et 2012. À ce jour, 14 581 et 2 334 documents sont référencés sous le terme « Penicillium chrysogenum » dans les bases de données Scopus et PubMed interrogées. Concernant les termes « Penicillium rubens le nombre de documents accessibles chute drastiquement. Scopus répertorie 232 documents, contre 67 pour PubMed avec le mot-clé « Penicillium rubens ». (Données collectées en avril 2023).*

Soulignons à cet instant que les changements taxonomiques au sein du genre *Penicillium* peuvent compliquer les recherches bibliographiques. Lorsqu'une révision nomenclaturale est adoptée, outre l'existence d'une période de latence entre le changement de nom et son adoption généralisée, les reclassements successifs induisent une fragmentation et une multiplicité des données, pouvant être source d'une certaine confusion. Cela peut générer des ambiguïtés dans l'accès à l'information, nécessitant une vigilance accrue dans la recherche des informations pour garantir l'exactitude et la complétude des données exploitées. Cet aspect, intrinséquement lié aux problématiques de l'accessibilité et de la traçabilité des données, sera discuté au cours du Chapitre 3, notamment dans les sections *2.1.1* *Description des données utilisées*, page 196 et *2.2.1* *Un gap-filling « raisonné »*, page 239.

Malgré son intense utilisation dans la recherche, *P. rubens* Wisconsin 54-1255 n'a pas encore révélé tous ces secrets. Le séquençage de son génome en 2008 *(Van den Berg et al. 2008)* a marqué un tournant dans la compréhension de son métabolisme. Les gènes impliqués dans la biosynthèse, la régulation et le transport des métabolites spécialisés sont généralement regroupés en clusters et codent principalement pour des peptides non ribosomiques, synthétisés par des *Nonribosomal Peptide Synthetases* (NRPS), des polycétides, catalysés par des *Polyketide Synthases* (PKS), des terpènes produits par des *Terpene Synthases* ou alors des molécules hybrides (cf. encart : *Classification des métabolites spécialisés*, page 79). Les analyses d'exploration du génome (*i.e. « genome mining »*) ont permis à ce jour d'identifier 33 gènes principaux du métabolisme spécialisé, dont 10 NRPS, 20 PKS, 2 hybrides NRPS-PKS et 1 diméthyl-allyl-tryptophane synthase (**Figure 1-10**).

Cependant, la majorité de ces gènes n'a été que partiellement étudiée, et la caractérisation de leurs produits demeure incomplète. Il est donc probable que *P. rubens* Wisconsin 54-1255 soit capable de synthétiser des composés encore non détectés, laissant entrevoir des perspectives inédites pour la découverte de nouveaux métabolites *(Guzmán-Chávez et al. 2018)*.

Classification des métabolites spécialisés

Les métabolites spécialisés des champignons filamenteux sont classés selon leur structure chimique et leurs fonctions biologiques et, en conséquence, les enzymes impliquées sont associées aux voies de biosynthèse correspondantes. Cette double classification permet une meilleure compréhension de l'importance et de la diversité de ces composés dans les stratégies adaptatives de ces organismes puisque chaque groupe de métabolites spécialisés joue un rôle distinct dans l'écologie et la survie des champignons, mettant en lumière leur impact crucial dans leurs interactions avec l'environnement.

CLASSIFICATION STRUCTURELLE (ET ENZYMES ASSOCIÉES)

- **Nonribosomal Peptide Synthetases (NRPS)** : Enzymes multimodulaires catalysant la biosynthèse des peptides non ribosomaux en assemblant directement des acides aminés avec des monomères, sans dépendance de l'ARNm. Les produits résultants peuvent inclure des acides aminés non protéinogènes.
- **Dimethylallyl Tryptophan Synthase : (DMATS)** : Enzyme impliquée dans la biosynthèse de divers alcaloïdes en catalysant la réaction de diméthylallylation du tryptophane.
- **Polyketide Synthases (PKS) :** Enzymes responsables de la biosynthèse des polycétides à partir d'unités d'acétyl-CoA ou de malonyl-CoA, fonctionnant de manière similaire aux synthétases d'acides gras. Il en existent divers types, principallement la PKS de type I (T1PKS) qui synthétisent des polycétides dans une configuration modulaire où chaque module ajoute une unité de chaîne carbonée à la molécule croissante. Chaque module contient les domaines nécessaires pour l'incorporation et la modification des unités de base, permettant une grande diversité structurale des polycétides produits.
- **Terpène Synthase (TS)** : Enzymes catalysant la cyclisation des précurseurs terpéniques, convertissant les isoprènes linéaires en structures cycliques complexes. Les terpènes, en servant de signaux chimiques, jouent des rôles prépondérants dans les mécanismes de défense et de communication entre organismes. Ils sont également utilisés comme précurseurs de nombreuses molécules bioactives.
- **Voies Hybrides** : D'autres mécanismes enzymatiques hybrides combinant des modules de diverses voies citées précédemment (*e.g.* hybrides NRPS-PKS, NRPS-TS, PKS-TS) existent. Ces voies permettent de générer une plus grande diversité structurale à partir de précurseurs variés.

CLASSIFICATION FONCTIONNELLE

- **Sidérophores :** Composés chélatant le fer, facilitant son acquisition dans des environnements pauvres en cet élément. Ils sont cruciaux pour la survie des champignons dans des conditions de compétition intense pour les nutriments.
- **Toxines :** Composés qui inhibent la croissance de bactéries, d'autres champignons ou qui peuvent intoxiquer des organismes plus complexes. Elles jouent un rôle défensif en empêchant la croissance ou en causant des dommages à d'autres organismes.
- **Pigments :** Composés colorés qui protègent contre les dommages causés par les rayons UV et les stress oxydatifs. Les pigments sont produits par les champignons pour leur permettre de survivre dans des environnements hostiles.
- **Antibiotiques :** Agents antimicrobiens qui inhibent la croissance des compétiteurs microbiens, jouant un rôle primordial dans les interactions écologiques en limitant la concurrence pour les ressources.

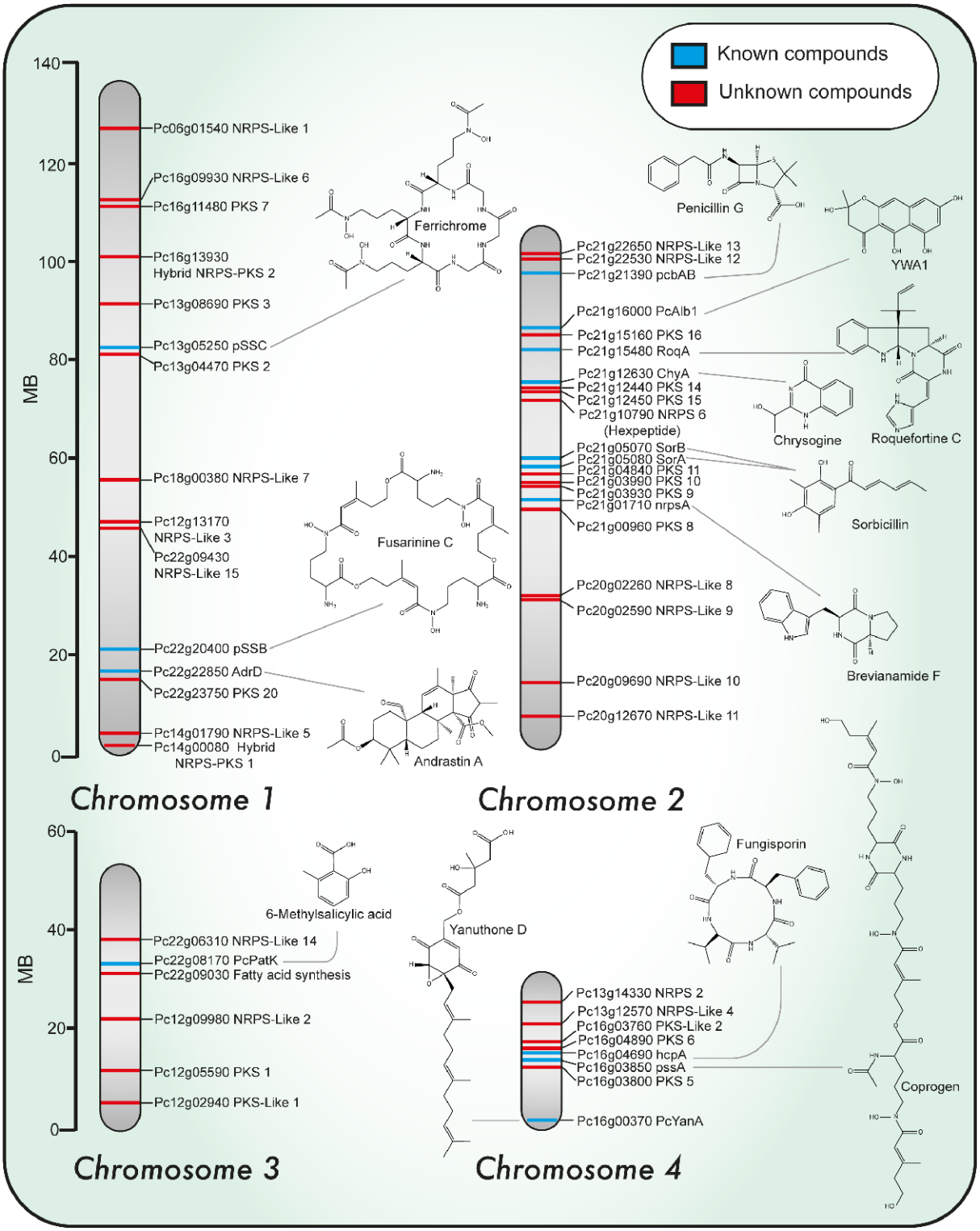


Figure 1-10 : « Localisation chromosomique des gènes PKS et NRPS connus et prédits, ainsi que des structures représentatives des métabolites spécialisés associés identifiés chez Penicillium [rubens]. *Les lignes bleues indiquent les produits associés connus, tandis que les lignes rouges indiquent les produits inconnus jusqu'à présent. ». Figure et légende issues et traduite des travaux de (Guzmán-Chávez et al. 2018).*

Au cours des 15 dernières années, plusieurs modèles métaboliques ont été développés pour *P. rubens* Wisconsin 54-1255 (**Tableau 1-2**). Cependant, il est crucial de noter que ces modèles ne sont pas équivalents en termes d'applicabilité. L'objectif principal lié à la génération d'un nouveau modèle (*i.e.* analyse de processus évolutifs, étude de la distribution des flux, *etc.*) influence fortement la réutilisabilité et l'universalité des modèles proposés. Ainsi, chaque modèle est adapté à un contexte spécifique, limitant parfois son usage dans d'autres études.

En 2013, cinq ans après le séquençage de son génome, le premier **GSMN** pour *P. rubens*, nommé *i*AL1006, a été publié *(Agren et al. 2013)*. Ce modèle, composé de 1 471 réactions, fut reconstruit à l'aide de RAVEN Toolbox et de la base de données KEGG. Notons que le **GSMN** *i*AL1006 constitue depuis un point de référence pour l'étude du métabolisme de ce champignon filamenteux et la génération des modèles suivants. En 2018, avec l'amélioration et l'automatisation des techniques de reconstruction, un nouveau modèle, Prubens, composé de 2 533 réactions avec identifiants uniques, a vu le jour *(Prigent et al. 2018)*. Ce modèle repose sur une révision à grande échelle d'*i*AL1006, en intégrant des données issues de MetaCyc. Au grés de l'évolution des connaissances intégrées dans les bases de données, chaque nouvelle version de **GSMN** évolue. Par exemple, entre les **GSMNs** *i*AL1006 et Prubens, publiés à cinq ans d'intervalles 986 nouvelles séquences génomiques ont été incluses (*i.e.* liées essentiellement soit à la dégradation de composés aromatiques, au métabolisme des lipides et plus largement au métabolisme spécialisé) et 199 de ces séquences n'ont pas été retenues dans la version Prubens. Parallèlement, un modèle non curé généré en 2016 à l'aide du pipeline de reconstruction automatique CoReCo a également été publié. Cependant, ce réseau, issu lui aussi d'un processus global de reconstruction, se base sur un ensemble de souches de *P. chrysogenum (Castillo et al. 2016)*. Entre temps, un modèle métabolique central spécifique à la production de pénicilline dérivé essentiellement d'*i*AL1006 et intégrant des données issues de KEGG, BioCyc et d'études biochimiques spécifiques à la pénicilline a été proposé *(Prauße et al. 2016)*. Ce modèle a permis d'examiner les flux élémentaires pour mieux comprendre les mécanismes de production de cet antibiotique. Enfin, outre ces modèles de réseaux, les bases de données spécialisées KEGG *(Kanehisa et al. 2022)* et BioCyc *(Karp et al. 2019)* disposent également chacune de « sous-bases » de données relatives aux données métaboliques de *P. chrysogenum/rubens*.

Malgré ces nombreuses ressources, la nécessité de réaliser une nouvelle reconstruction s'est rapidement imposée, motivée à la fois par l'évolution continue des connaissances intégrées dans les bases de données, en particulier celles liées au métabolisme spécialisé, et par des questions d'adéquation de format avec les outils utilisés. Les difficultés rencontrées au cours des travaux, notamment en raison de problèmes d'interopérabilité (*e.g.* chargement des modèles sur divers outils, absence de paramétrisation, impossibilité de reproduire certains résultats de simulations) et de prise en main des modèles souvent complexes, nécessitant un temps incompressible pour leur compréhension et leur appropriation, font que ce choix se sera avéré judicieux. Ces aspects seront approfondis dans le Chapitre 3 : *Reconstruction et Réconciliation d'un Réseau Métabolique à l'Échelle du Génome de Haute Qualité pour Penicillium rubens Wisconsin 54-1255*, page 125.

Tableau 1-2 : Nombre d'occurrences des entités dans les différents modèles de réseaux métaboliques de *Penicillium rubens* Wisconsin 54-1255. Nous présentons dans ce tableau un résumé synthétique des éléments contenus dans les divers réseaux métaboliques disponibles liés à notre organisme d'étude. Les nombres d'entités affichés proviennent de l'exploration des fichiers des modèles. Il est à noter toutefois que de légères différences peuvent exister entre les valeurs affichées ici et celles rapportées dans les publications, à l'instar de l'article sur *i*AL1006, qui mentionne un réseau avec 1 235 métabolites, nombre que nous n'avons pas retrouvé dans les fichiers du modèle (*i.e.* **SBML** et fichier Excel). Pour les **GSMNs** *i*AL1006 et Prubens, nous faisons la distinction entre le nombre d'entités total et unique, cette variation provenant de la compartimentation intracellulaire de ces modèles qui occasionne une multiplicité de réactions et métabolites. Les fichiers relatifs au contenu du PGDB (« *Pathway/Genome Database Collection* ») de *P. rubens* Wisconsin 54-1255 ont été téléchargés en décembre 2019 et, hormis le fait qu'ils ont été générés par une équipe de l'Université de technologie de Delft (*i.e.* Pays-Bas), nous n'avons pas, à notre connaissance, de visibilité sur l'évolution ou l'historique de ces données (*e.g.* date de dépôt, nombre de versions, fréquence des mises à jour, *etc.*). Ce PGDB, qui intègre le réseau de réactions biochimiques et les voies métaboliques de l'organisme avec son génome, a été améliorée par une curation manuelle limitée et a été créé à partir du génome annoté en utilisant le composant PathoLogic du logiciel Pathway Tools et la base de données de référence MetaCyc. Le dernier modèle présenté est issu du pipeline de reconstruction automatique CoReCo, disponible sur BioModel. Ce réseau a été utilisé lors des premières étapes de reconstruction de notre réseau ***i*Prub22** (cf. *Annexe 2 : Reconstruction d'iPrub22 - sous-réseau issu des recherches d'orthologie*, page 433) et pour ce faire, des adaptations du **SBML** ont du être réalisées et sont mentionnées dans ce tableau sous les étiquettes avant et après nettoyage. La réduction drastique du nombre de réactions est due à la conception même du modèle avec une absence de consensus et d'uniformisation des données. Par exemple, dans le modèle d'origine, nous détectons des réactions multiples en raison soit de leur source de détection (*e.g.* une réaction détectée par recherche d'homologie chez divers organismes figurera autant de fois dans le modèle) soit en raison de la base de données utilisée (*e.g.* MetaCyc, KEGG).

Réseau métabolique	Nombre de					
	Réactions		Métabolites		Séquences génomiques	
***i*AL1006** *(Agren et al. 2013)*	Biochimiques	1 471	Totaux	1 395	1 006	
	De modélisation	160	Uniques	849		
Modèle de synthèse de la pénicilline *(Praußе et al. 2016)*	65		81		-	
Prubens *(Prigent et al. 2018)*	Biochimiques (Totales)	2 556	Totaux	3 058	1 786	
	Biochimiques (Uniques)	2 515				
	De modélisation	18	Uniques	2 594		
PGDB (BioCyc)	Métaboliques	1 291	1 111		13 933	
	De transport	92				
MODEL1604280030 (CoReCo) *(Castillo et al. 2016)*	Avant nettoyage	6 876	Avant nettoyage	3 329	1 352	
	Après nettoyage	3 666	Après nettoyage	3175		

CHAPITRE 2 :

Glossaire des Ressources, Outils et Concepts Employés

Ce Chapitre a pour objectif de fournir un « glossaire » des outils clés abordés dans ce document, afin de faciliter l'assimilation des informations présentées. Il décrit de manière concise certaines ressources essentielles et les principaux outils utilisés dans ce manuscrit, tout en expliquant les raisons et les modalités de leur utilisation. De plus, il met en lumière les avantages et les inconvénients rencontrés lors des travaux de recherche.

Face à la multitude de possibilités pour la reconstruction de **GSMNs**, notre sélection d'outils a été principalement guidée par notre volonté d'utiliser la base de données MetaCyc et la boîte à outils **AuReMe**. Nous étions déjà familiarisés avec leur mode de fonctionnement et convaincus de leurs performances et caractéristiques, des aspects qui sont détaillés dans les sections correspondantes. Outre les besoins spécifiques liés à l'utilisation de ces ressources, notre choix d'outils s'est également appuyé sur leur notoriété et utilisation répandue dans les domaines concernés, ainsi que sur leur facilité d'installation et d'utilisation. La majorité d'entre eux ont été employés localement en ligne de commande – des détails sur l'environnement de travail sont fournis ci-dessous – et le système de gestion des packages et des environnements conda (v4.12.0) a été utilisé pour certaines installations. Un extrait de leur documentation, lorsqu'elle était disponible, ainsi que les commandes employées sont présentés dans ce manuscrit pour attester de leur simplicité d'utilisation.

- **Système d'exploitation** : Ubuntu 18.04.6 LTS
- **Kernel Linux** : 5.4.0-150-generic, x86_64
- **Machine** : Dell OptiPlex 7470 AIO
- **Desktop**: Gnome 3.28.4
- **Processeur (CPU)** : Intel Core i7-8700, 6 cœurs (12 threads), fréquence min/max : 800/4600 MHz
- **RAM** : 16 GB
- **Carte graphique** : Intel UHD Graphics 630, OpenGL version 4.6, pilote i915
- **Stockage** : Disque NVMe 512 GB
- **Environnement graphique** : Gnome 3.28.4, X.Org 1.20.8

Afin de faciliter la navigation, les termes sont organisés par thématique et présentés dans l'ordre chronologique de leur utilisation. La **Figure 2-1** propose un récapitulatif graphique des éléments présentés dans ce glossaire ainsi que de l'ensemble des éléments employés.

1 - Synthèse

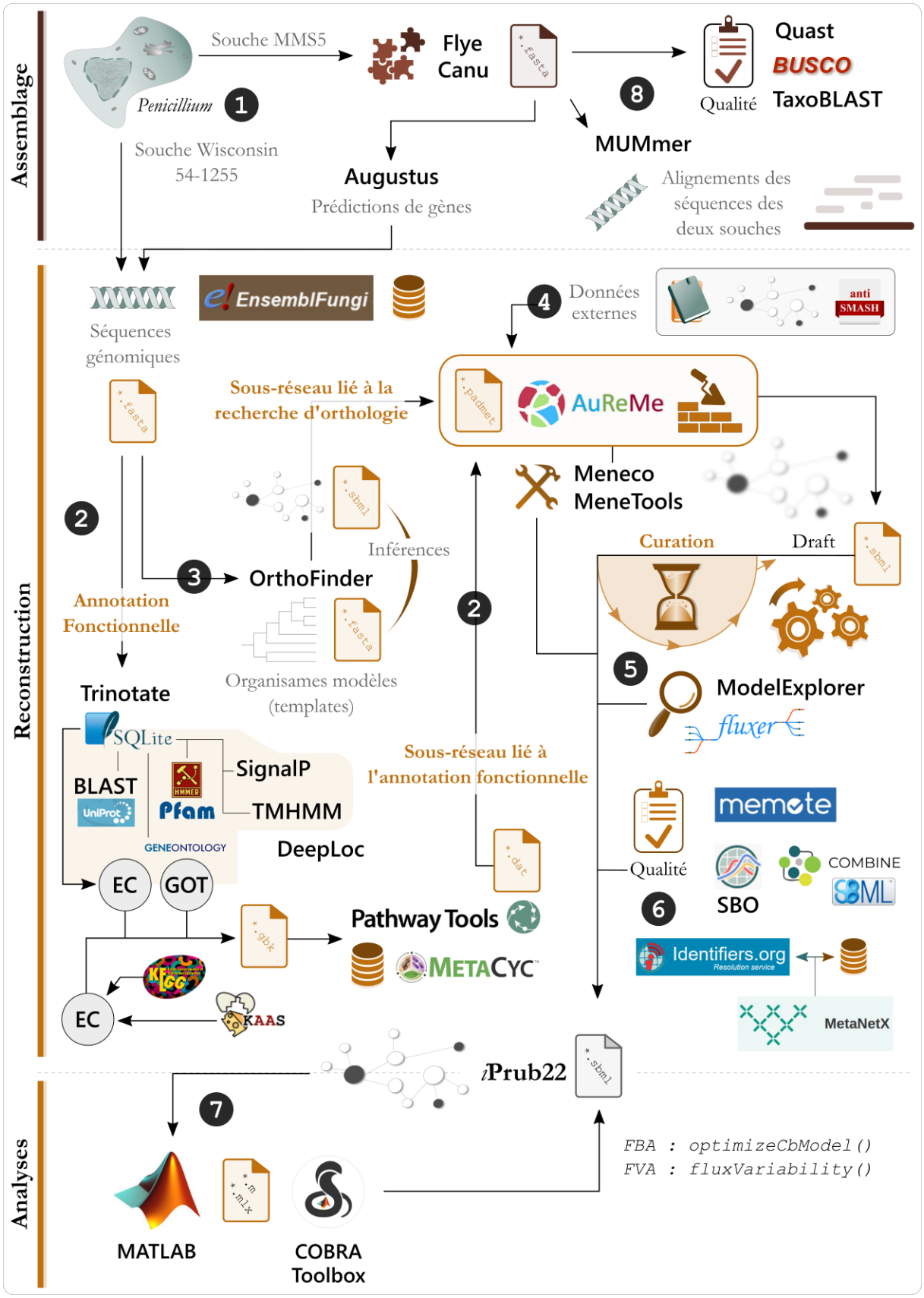


Figure 2-1 : Résumé graphique des outils et ressources utilisés dans le cadre des travaux présentés dans ce manuscrit.

La **Figure 2-1** présente l'ensemble des outils utilisés qui seront cités dans ce manuscrit. Les principales bases de données interrogées ainsi que les types de fichiers associés aux outils y figurent également. L'objectif de ce schéma est d'offrir une visualisation globale des protocoles suivis pour chaque thématique, facilitant ainsi la compréhension. Les paragraphes suivants fournissent des explications succinctes, chaque item **mis en relief en gras** étant traité plus en profondeur ultérieurement. À cet effet, nous utilisons les abréviations courantes des termes qui seront expliquées par la suite.

❶ SÉLECTION DES DONNÉES

Les travaux présentés dans ce manuscrit reposent sur des données génomiques de champignons filamenteux du genre *Penicillium*. Nous avons concentré nos efforts sur la modélisation du métabolisme de l'organisme modèle *P. rubens* Wisconsin 54-1255, dont les données sont extraites d'**EnsemblFungi**.

❷ GÉNÉRATION DU DRAFT : SOUS-RÉSEAU LIÉ À L'ANNOTATION FONCTIONNELLE

Un sous-réseau lié à l'annotation fonctionnelle, reposant sur la base de données **MetaCyc**, a été généré avec **AuReMe** en utilisant les résultats de l'outil **Pathway Tools**. Les annotations, principalement les numéros EC et les GOT, ont été obtenues à l'aide du pipeline **Trinotate**. Ce dernier permet, entre autres, d'effectuer des recherches d'homologie *via* Blast contre UniProt, d'identifier la présence de domaines protéiques avec HMMER contre PFAM et d'extraire les termes de la Gene Ontology associés aux séquences génomiques interrogées. Les résultats sont ensuite intégrés dans une base de données SQLite afin de faciliter leur visualisation et leur utilisation.

L'annotation en numéros EC a été enrichie avec les résultats issus de **KAAS**, ainsi que des données préexistantes pour *P. rubens* Wisconsin 54-1255, hébergées dans KEGG.

Enfin, lors du processus d'annotation, nous avons également utilisé **SignalP**, **TMHMM**, et **DeepLoc** pour générer des informations sur la localisation subcellulaire des protéines, en vue de guider certaines étapes de la curation manuelle ultérieure.

❸ GÉNÉRATION DU DRAFT : SOUS-RÉSEAU LIÉ À LA RECHERCHE D'ORTHOLOGIE

Un sous-réseau lié à la recherche d'orthologie a été reconstruit en s'appuyant sur les résultats issus d'**OrthoFinder**. Pour ce faire, sept reconstructions provenant d'organismes phylogénétiquement plus ou moins distants ont été utilisées. La génération du sous-réseau s'effectue dans **AuReMe**.

❹ GÉNÉRATION DU DRAFT : SOURCES EXTERNES

Les sous-réseaux intermédiaires ont été fusionnés sous **AuReMe** et enrichis avec des données issues de réseaux et de sources antérieures. Une étude bibliographique, combinée aux résultats d'**AntiSMASH**, ont permis de générer une liste de métabolites appartenant à *P. rubens* Wisconsin 54-1255, comprenant, à la fois, des métabolites de base et des métabolites spécialisés que nous nous attendons à retrouver dans le réseau de ce champignon filamenteux.

❺ CURATION MANUELLE ET RAFFINEMENT DE LA RECONSTRUCTION

Le draft a ensuite été soumis à plusieurs étapes itératives de curation, principalement à l'aide de l'outil de *gap-filling* **Meneco**. De plus, des analyses topologiques ont été effectuées en utilisant la suite d'outils **Mene Tools**. La topologie du réseau et la distribution des flux ont été visualisées à l'aide de **ModelExplorer** et **Fluxer**.

❻ ÉVALUATION DE LA QUALITÉ D'UN GSMN

L'interopérabilité des données présentes dans notre réseau avec les autres bases de données a été assurée par l'utilisation de la base de données MetaNetX, et nous avons respecté les recommandations d'écriture fournies sur le site identifiers.org. Le réseau a également été enrichi avec des termes issus de l'ontologie SBO afin de faciliter l'intégration et l'échange de données. Enfin, notre reconstruction ***i*Prub22**, distribuée au format **SBML**, a été testée et validée à l'aide de la suite d'outils **MEMOTE** et est en adéquation avec les directives établies par la communauté scientifique pour la modélisation (*e.g.* documentation **SBML** issue de COMBINE).

❼ ANALYSES QUANTITATIVES ET MODÉLISATION DES FLUX

La paramétrisation d'une reconstruction permet de définir des modèles qui sont analysés quantitativement par l'étude de la distribution des flux à l'aide de la modélisation par contraintes. Pour ce faire, nous avons utilisé le logiciel **MATLAB**, un environnement de calcul numérique avancé, associé à la bibliothèque **COBRA Toolbox**, spécifiquement dédiée à l'analyse des **GSMNs**.

❽ PRÉPARATION DES DONNÉES DE LA SOUCHE MARINE MMS5

Suite au séquençage de la souche MMS5 (*Marine Mycological Strains*), nous avons reçu deux assemblages générés par Canu et Flye, ainsi que les prédictions des gènes codants fournies par Augustus. En l'absence de données RNASeq et afin de classifier ces deux jeux de données, nous avons utilisé BUSCO et Quast, deux outils permettant d'évaluer la qualité des assemblages. Nous avons également employé TaxoBLAST pour vérifier si l'ensemble des contigs mis à notre disposition était bien d'origine eucaryote (*i.e.* pour confirmer l'absence d'une potentielle contamination). Par la suite, nous avons utilisé MUMmer, un outil d'alignement de génomes, pour réaliser une comparaison génomique préliminaire entre la souche marine MMS5 et la souche terrestre Wisconsin 54-1255. Enfin, nous avons réalisé un draft de reconstruction en nous appuyant sur le protocole de reconstruction d'***i*Prub22**.

2 - Types de fichiers employés

Les formats de fichiers utilisés dans ces travaux sont des standards éprouvés en biologie des systèmes, caractérisés par leur simplicité et leur nature textuelle. Qu'il s'agisse de formats spécifiques (*e.g.* `*.dat` de **Pathways Tools**, `*.padmet` pour **AuReMe**), propriétaires (*e.g.* `*.mlx` de **MATLAB**) ou génériques (*e.g.* `*.fasta`), ces fichiers sont pour la plupart directement lisibles par l'humain, ce qui facilite une prise en main rapide et une compréhension efficace des données. Ces formats reposent sur des standards assurant une interopérabilité optimale entre divers outils, indispensables pour la gestion fluide des flux de travail. Leur sélection est principalement fondée sur des critères de simplicité, de compatibilité, et de capacité à intégrer des annotations. Lorsque des informations supplémentaires sont requises, des scripts personnalisés nous ont permis d'ajouter directement ces annotations par le biais de fichiers texte, généralement tabulés, optimisant ainsi la gestion des données et leur lisibilité.

Séquences génomiques :

- `*.fasta` : fichiers textuels standards servant à stocker des séquences de nucléotides (`*.fna`) ou de protéines (`*.faa`).

Fichiers d'annotations :

- `*.gff` : format standard et concis pour le stockage des annotations génomiques. Un GFF (General Feature Format) est un fichier tabulé composé de neuf champs comprenant l'identifiant de la séquence, sa source, la nature de l'élément considéré, les coordonnées génomiques de début et de fin de séquences, un score de confiance, le brin codant de l'élément, la phase du cadre de lecture et un dernier champ non standardisé contenant d'autres caractéristiques. Ce format textuel est largement adopté pour sa simplicité, sa légèreté et sa compatibilité avec divers outils de bio-informatique.

- `*.gbk` : format standard pour le stockage des séquences d'acides nucléiques ou de protéines avec leur annotation. Un fichier GBK (**GenBank** format) est un fichier textuel qui contient des informations détaillées sur les séquences d'ADN, y compris des éléments tels que les gènes, les exons, les introns et les caractéristiques associées. Ce format présente des sections bien définies, incluant l'identifiant de la séquence, des annotations fonctionnelles, ainsi que des informations sur les publications associées et les accès aux séquences. Les annotations sont structurées à l'aide de lignes spécifiques, permettant une interprétation facile et un accès direct aux données. Le format GBK est largement utilisé dans la recherche biologique pour son exhaustivité et sa compatibilité avec de nombreux outils de bio-informatique, facilitant ainsi le partage et l'échange d'informations génomiques.

Fichiers de réseaux métaboliques

- `*.dat` (générés par **Pathways Tools**) : fichiers de données polyvalents dépendants généralement du logiciel qui les produit. Les PGBDs générés à l'aide de **Pathways Tools** peuvent être exportés sous forme de fichiers de données attribut-valeur (*i.e.* ensemble de paires attribut-valeur qui décrivent les propriétés de l'objet et les relations de l'objet avec d'autres objets) afin de faciliter l'utilisation de ces données par d'autres programmes et systèmes de gestion de bases de données. Chaque fichier attribut-valeur contient des données relatives à une classe d'objets, tels que les gènes, les enzymes, les réactions, *etc.* Ces fichiers structurent les données de manière simple et flexible, permettant d'organiser les informations textuelles. Ce sont eux qui sont lus et interprétés par **AuReMe** pour la génération du sous-réseau d'annotations fonctionnelle.

- `*.padmet` (spécifique à **AuReMe**) : fichier représentant le réseau métabolique d'une espèce en s'appuyant sur une base de données de référence, généré au sein d'**AuReMe** *via* le package Python PADMet. Ce dernier fournit un ensemble de méthodes pour la gestion, l'analyse, et la visualisation des données lors de la reconstruction des **GSMNs**. Le format PADMet se distingue par une structure de données claire, composée d'entités distinctes (*Node, Policy, Relation*) organisées pour capturer les interactions au sein du réseau métabolique. Cette organisation modulaire et relationnelle permet une annotation flexible et une hiérarchisation des informations, facilitant l'intégration de métadonnées additionnelles. La nature textuelle et relativement compacte du format le rend particulièrement efficace pour le *parsing* et les étapes de curation manuelle.

- `*.sbml` : format standard de stockage et de diffusion des **GSMNs**. Le format **SBML** (***S****ystems* ***B****iology* ***M****arkup* ***L****anguage*) est un standard XML (E***x***tensible ***M****arkup* ***L****anguage* - *langage de balisage extensible*) pour la description de modèles biologiques. Il inclut, entre autres, des balises spécifiques pour représenter les produits géniques, les réactions enzymatiques et les métabolites. Ce format facilite l'échange automatisé de contenus complexes entre de nombreux logiciels de biologie des systèmes. Toutefois, sa complexité relative et sa taille nécessitent généralement des outils spécialisés pour le *parsing* et l'analyse des fichiers. À titre d'illustration, le fichier `*.padmet` ayant servi à la génération d'***i*Prub22** contient 180 086 lignes et a une taille de 19 Mo. Le fichier `*.sbml` généré à partir de ce fichier sous **AuReMe** comprend 426 745 lignes pour une taille de 27 Mo. Après modifications et ajustement du format (*i.e.* conversion en FBC2 et peuplement des annotations des produits géniques), la version finale d'***i*Prub22** comporte 606 685 lignes pour un poids de 32 Mo.

Fichiers de diffusion :

- `*.mlx` : format dans lequel **MATLAB** propose ses Live scripts, des documents interactifs combinant texte, cellules de code et résultats. Ces documents sont structurés pour une exécution interactive et itérative des analyses, assurant ainsi leur reproductibilité. Cependant, l'exécution de ce format propriétaire requiert une licence **MATLAB**.

- `*.html` : HTML (*HyperText Markup Language*) est un langage de balisage destiné à la création de pages web. Utilisé comme alternative pour la diffusion du processus de reconstruction d'***i*Prub22**, présenté dans le fichier `*.mlx` rendu disponible avec la publication de notre **GSMN**.

3 - Outils, logiciels et algorithmes

3.1. Annotation fonctionnelle – Reconstruction du draft

Les annotations structurales et fonctionnelles des séquences génomiques sont la clé de voûte de la modélisation du métabolisme d'un organisme. Ainsi, afin de créer un sous-réseau d'annotations, nous avons utilisé l'outil **Pathways Tools**, qui produit un PGBD (*Pathway/Genome Databases*) sous forme de fichiers plats qui seront lus et interprétés par **AuReMe**. Pour ce faire, nous avons fourni à **Pathways Tools** un fichier GenBank (*i.e.* *.gbk) généré à partir de divers fichiers d'annotations (*.GFF/GFF3) et d'un fichier tabulé enrichi en annotations.

Concernant les trois fichiers GFF que nous avons utilisés, deux d'entre eux proviennent de la recherche d'ARN effectuée à l'aide de **barrnap** et **tRNAScan-SE**, tandis que le troisième correspond au fichier d'annotations de référence téléchargé depuis **EnsemblFungi**.

Le fichier d'annotations tabulé contient, quant à lui, les numéros EC (EC), les termes issus de la Gene ontology (GOT) et les domaines protéiques (PFAM). Ces trois types d'annotations ont été principalement déterminés à l'aide du pipeline **Trinotate**. Plus précisément, les numéros EC sont extraits en interrogeant la base de données KEGG à partir de ses propres identifiants KEGG Orthology (KO). Ces identifiants KO proviennent, soit des enregistrements des gènes de *P. rubens* Wisconsin 54-1255 stockés dans la base de données KEGG GENOME sous le numéro d'accession T01091, soit des résultats issus de **KAAS** (*KEGG Automatic Annotation Server*), soit des résultats fournis par **Trinotate**.

Un résumé de ces résultats est présenté en Annexe 1 : *Reconstruction d'iPrub22 – sous-réseau issu de l'annotation fonctionnelle*, page 425.

KAAS (version 2.1) *(Moriya et al. 2007)*

KEGG *Automatic Annotation Server* (**KAAS**) est un pipeline dédié à l'annotation automatique de séquences génomiques. Lancé en 2005, ce serveur en libre accès utilise des algorithmes tels que BLAST ou GHOST pour réaliser des alignements et rechercher des similitudes entre les gènes d'un organisme cible et l'ensemble ou des sous-ensembles de la base de données KEGG GENES.

Ce système d'annotation repose sur la propagation des informations des séquences de gènes orthologues. Les gènes orthologues, résultant du processus de spéciation, partagent une fonction biologique similaire, conservée à travers l'évolution au sein de différentes espèces. Ainsi, l'identification de gènes orthologues constitue une approche courante et rapide pour, par exemple, caractériser fonctionnellement des génomes nouvellement séquencés. Concrètement, la détermination des ensembles de gènes orthologues repose sur le calcul de la relation de la meilleure correspondance bidirectionnelle (*Best Bidirectional Hit* – BBH) lors de la comparaison des génomes par paire. Cette méthode implique que chaque gène du génome requête est comparé à l'ensemble des gènes du génome de référence, et inversement.

Au sein de la base de données KEGG, les gènes sont annotés avec des identifiants KEGG Orthology (KO) qui permettent la classification hiérarchique et fonctionnelle des gènes ainsi que de leurs produits géniques. Chaque groupe de gènes orthologues est caractérisé par un identifiant unique KO qui définit une fonction biologique ou métabolique spécifique. Ces identifiants KO, directement liés aux entrées de la base de données KEGG PATHWAY, sont ensuite utilisés pour générer des cartes des voies métaboliques spécifiques de l'organisme étudié. Enfin, l'attribution d'un identifiant KO à un gène repose sur la détermination du score le plus élevé parmi tous les candidats orthologues appartenant au même groupe d'orthologues. Ce score est normalisé en fonction de la taille des séquences, puis pondéré selon le nombre de candidats orthologues associés au KO considéré.

L'assignation des KO par le serveur **KAAS** a été validée par les travaux de Moriya *et al. (2007)* avec respectivement des taux de sensibilité et de spécificité supérieurs à 70 % et 90 %. Leurs résultats ont été obtenus en interrogeant les génomes d'*Homo sapiens*, de *Saccharomyces cerevisiae*, d'*Escherichia coli* et d'*Arabidopsis thaliana* contre la base de données KEGG GENES.

En conséquence, comme **KAAS** offre une opportunité supplémentaire d'interroger et d'explorer les données présentes dans KEGG, nous avons exploité ce pipeline dans le cadre de notre recherche d'enrichissement en numéros EC. Nous avons ainsi effectué une analyse basée sur la méthode de recherche des BBH confrontant le protéome de *P. rubens* Wisconsin 54-1255 à un ensemble de 367 616 séquences réparties sur 37 organismes appartenant tous au règne fongique. Par défaut, les hits dont le score est inférieur à 60 ne sont pas retenus comme étant des orthologues potentiels à la séquence interrogée. Les KO résultants permettent alors de faire un lien direct avec les numéros EC recherchés et fournissent un point d'ancrage pour l'utilisation des diverses ressources disponibles et spécifiques à la base de données KEGG.

Trinotate (version 3.2.1)  *(Bryant et al. 2017)* Trinotate

Trinotate est une suite d'annotations fonctionnelle open-source conçue pour l'annotation automatique des transcriptomes, particulièrement ceux assemblés *de novo*. Ce *pipeline* combine et intègre plusieurs analyses pour fournir des informations exhaustives sur les transcrits, incluant l'identification des protéines homologues, la prédiction des domaines fonctionnels et l'adressage cellulaire. Les résultats sont ensuite intégrés dans une base SQLite, initialisée et pré-remplie avec des données génériques sur les enregistrements SWISSPROT et les domaines Pfam. Un rapport d'annotations (*i.e.* format textuel tabulé) consolidé de toutes les informations collectées de l'organisme étudié est ensuite généré, incluant les résultats de chaque analyse et les informations des bases de données croisées (*e.g.* eggNOG, GO, KEGG). Une vue d'ensemble des caractéristiques fonctionnelles potentielles de chaque transcrit, comme les familles de protéines, les domaines fonctionnels, et les informations sur la localisation subcellulaire sont également disponibles sous forme de graphiques interactifs. Enfin, il est à noter que **Trinotate** ne fait plus l'objet d'un développement ou d'un soutien actif depuis mars 2024.

Bien que nous ne travaillons pas sur des données transcptomiques (*i.e. stricto sensu* nous avons utilisé les séquences codantes nucléiques de *P. rubens* et leur version transcrite) nous avons choisi **Trinotate** pour bénéficier d'un cadre unifié permettant le stockage, l'analyse et l'extraction facilitée des annotations. Ces dernières sont ensuite utilisées pour enrichir le fichier **GenBank**, qui sert d'entrée dans **Pathway Tools**. Afin d'adapter **Trinotate** à ce contexte, seuls des ajustements d'identifiants étaient nécessaires lors du chargement des séquences génomiques dans la base SQLite. Ainsi, le pipeline d'annotations fonctionnelles **Trinotate** (v3.1.1) a été utilisé pour réaliser des recherches d'homologie en utilisant blastx et blastp (v2.9.0+) contre UniProt Swiss-Prot, des recherches de domaines protéiques à l'aide de hmmscan (3.2.1) contre PFAM et des recherches de peptides signaux et de domaines transmembranaires avec **SignalP** (v4.1) et **TMHMM** (v2.0c). Les bases de données associées ont été téléchargées le 13 janvier 2020. Un script interne à **Trinotate** permet ensuite d'extraire les termes GO associés à chaque séquence présente dans le rapport d'annotations généré par **Trinotate**. Pour les recherches d'homologie réalisées avec BLAST, seul l'alignement ayant obtenu le meilleur score a été sélectionné et constitue son annotation. Les seuils de détection par défaut, *e*-value fixée à 10, ont été appliqués pour BLAST et hmmscan.

```
# Recherche de similarités
blastx -query *.fna -db uniprot_sprot.pep -outfmt 6 -evalue 0.001 > blastx_output
blastp -query *.faa -db uniprot_sprot.pep -outfmt 6 -evalue 0.001 > blastp_output

# Recherche des domaines protéiques
hmmscan -E 0.001 --domE 0.01 --domtblout pfam.output Pfam-A.hmm *.faa > pfam_log
```

```
# Mise en place de Trinotate
Build_Trinotate_Boilerplate_SQLite_db.pl data_Prubens

# Initialisation des bases de données uniprot pour blast et Pfam pour Hmmscan
makeblastdb -in uniprot_sprot.pep -dbtype prot
gunzip Pfam-A.hmm.gz
hmmpress Pfam-A.hmm

# Chargement des séquences dans la base de données
[…]

# Peuplement de la base de données
Trinotate data_Prubens.sqlite LOAD_swissprot_blastx blastx.output
Trinotate data_Prubens.sqlite LOAD_swissprot_blastp blastp.output
Trinotate data_Prubens.sqlite LOAD_pfam pfam_output.output
Trinotate data_Prubens.sqlite LOAD_signalp signalp.output
Trinotate data_Prubens.sqlite LOAD_tmhmm tmhmm_output

# Génération du rapport
Trinotate data_Prubens.sqlite report > annotation_report.xls

# Extraction des GO Termes
extract_GO_assignments_from_Trinotate_xls.pl --Trinotate_xls
annotation_report.xls -G --include_ancestral_terms > got_annotations.txt
```

Barrnap (version 0.9)  *https://github.com/tseemann/barrnap* Barrnap

Barrnap est un outil bio-informatique développé au début des années 2010, utilisé pour détecter les ARN ribosomiques (ARNr) au sein des séquences génomiques. Initialement conçu pour identifier les ARNr bactériens et archéens, **Barrnap** était l'acronyme de *Bacterial/Archaeal Ribosomal RNA*. Avec le temps, **Barrnap** a élargi sa portée pour inclure un éventail plus large de taxonomies, et son acronyme a été rétroactivement repensé en tant que ***BA****sic* ***R****apid* ***R****ibosomal* ***RNA*** ***P****redictor*. Cet outil utilise des modèles de Markov cachés (HMM) spécifiques pour détecter divers types d'ARNr. Ainsi, ces modèles HMM, issus de Rfam, Silva et RefSeq, ont été construits en se basant sur les caractéristiques de séquence et de structure typiques des ARNr des bactéries (5S, 23S, 16S), des archées (5S, 5.8S, 23S, 16S), des mitochondries de métazoaires (12S, 16S) et des eucaryotes (5S, 5.8S, 28S, 18S).

Les ARNr jouent un rôle essentiel et structural au sein des ribosomes en contribuant, entre autres, au positionnement précis de ces complexes au niveau des codons initiateurs lors de la traduction des ARN messagers. Typiquement intégrée dans des pipelines d'annotation génomique et de séquençage, la détection des ARNr dans un génome offre une diversité d'applications dans des domaines majeurs de la biologie, notamment en phylogénie moléculaire, en biologie évolutive ou bien en métagénomique.

En conséquence, et en raison de sa simplicité en termes d'utilisation, nous avons choisi d'utiliser la dernière version de **Barrnap**, la version 0.9, disponible depuis 2018, pour annoter les génomes des deux souches de *Penicillium* dont nous disposions. **Barrnap**, utilisé en version locale, génère un fichier d'annotation au format GFF3 à partir d'un fichier FASTA contenant les séquences d'ADN de l'organisme étudié.

```
Synopsis: barrnap 0.9 - rapid ribosomal RNA prediction
Author: Torsten Seemann
Usage: barrnap [options] chr.fa

Options:
  --kingdom bac arc mito euk (default 'bac')
  --quiet No screen output (default OFF)
  --lencutoff Proportional length threshold to label as partial (default '0.8')
  --reject Proportional length threshold to reject prediction (default '0.25')
  --evalue Similarity e-value cut-off (default '1e-06')
  --outseq Save rRNA hit seqs to this FASTA file (default '')
```

```
barrnap --kingdom euk PenchWisc1_1_AssemblyScaffolds.fasta --quiet
--outseq rRNA_hits.fasta > output_barrnap.gff
```

tRNAscan-SE (version 2.0.5) *(Chan et al. 2021)*

L'outil **tRNAscan-SE**, développé à la fin des années 90 à la suite du décryptage des premiers génomes, est spécifiquement conçu pour la détection des gènes codant pour les ARN de transfert (ARNt) à partir de séquences génomiques. Les ARNt jouent un rôle central en tant qu'intermédiaires entre l'ARN messager et les acides aminés dans le processus de traduction des protéines. Les modèles statistiques de covariance sur lesquels repose **tRNAscan-SE** intègrent des informations liées à la séquence primaire et à la structure secondaire des ARNt. Ces modèles ont été entraînés et construits à partir de génomes représentatifs de procaryotes, d'eucaryotes et d'archées.

Cet outil, largement reconnu comme une référence en matière de détection des ARNt, est fréquemment utilisé dans les domaines de la biologie moléculaire, de la phylogénie et de la compréhension des processus d'évolution. La caractérisation des ARN contribue à une meilleure annotation des génomes et à une compréhension plus approfondie de la machinerie cellulaire. De plus, les ARNt et ARNr étant bien caractérisés et relativement faciles à identifier expérimentalement, leur détection dans un génome séquencé permet de valider la qualité de l'annotation génomique. Enfin, la recherche d'ARNt, associée à celle des ARNr, permet, dans le cadre de la reconstruction de réseaux métaboliques, d'appréhender les relations entre métabolisme et régulation de la traduction.

À l'instar de **Barrnap**, **tRNAscan-SE** fournit ses résultats sous forme de ficher d'annotations `*.gff`, fichier qui est ensuite directement interprétés par **Pathways Tools**. Par conséquent, nous avons appliqué **tRNAscan-SE** aux séquences de *P. rubens* Wisconsin 54-1255 pour enrichir les données utilisées dans la génération du sous-réseau lié à l'annotation fonctionnelle, ainsi qu'aux séquences de *P. chrysogenum* MMS5 pour affiner l'annotation de son génome.

```
tRNAscan-SE 2.0.5 (October 2019)
Copyright (C) 2019 Patricia Chan and Todd Lowe
University of California Santa Cruz
Freely distributed under the GNU General Public License (GPLv3)

Usage: tRNAscan-SE [-options] <FASTA file(s)>

Scan a sequence file for tRNAs
-- default: use Infernal & tRNA covariance models with eukaryotic sequences […]

Output options:
 -o --output <file> : save final results in <file>
 -f --struct <file> : save tRNA secondary structures to <file>
 -s --isospecific <file> : save results using isotype-specific models in <file>
 -m --stats <file> : save statistics summary for run in <file>
 -a --fasta <file> : save predicted tRNA sequences in FASTA file format of <file>
 --detail : display prediction outputs in detailed view
[…]

# Exécuction sur l'assemblage
tRNAscan-SE Assemblage_Scaffolds.fasta -o tRNAScan.output -f struct
-s isospecific -m stats -a fasta --detail

# Conversion des résultats en gff
./convert_tRNAScanSE_to_gff3.pl -i tRNAScan.output > tRNAScan.output.gff
```

Pathway Tools (v 23.0) *(Karp et al. 2021)* Pathway Tools

Pathway Tools est un logiciel propriétaire complet de biologie des systèmes qui est associé à la collection de bases de données **BioCyc** et développé depuis les années 1990 par le *Bioinformatics Research Group* du SRI International. Cette plateforme est conçue pour l'annotation de génomes et la construction de bases de données des voies métaboliques, appelées *Pathway/Genome Databases* (PGDBs). **Pathway Tools** permet l'analyse, la visualisation, et l'interprétation de réseaux métaboliques, d'annotations génomiques et des voies biologiques d'organismes variés.

Seule la composante PathoLogic qui crée un PGDB sous forme de fichiers plats contenant la prédiction des voies métaboliques issues de **MetaCyc** a été utilisée. À partir des séquences génomiques d'un organisme et de fichiers d'annotations (`*.gbk` ou `*.gff`), ce module permet de réaliser une déduction du réactome de

l'organisme étudié. Les PGDBs permettent de représenter et d'organiser les voies métaboliques en présentant le réseau réactionnel d'un organisme dans un format relationnel. Ces bases de données peuvent être modifiées manuellement pour affiner les annotations et ajuster les prédictions automatiques. L'utilisateur peut ainsi ajuster la structure et le contenu des réseaux métaboliques, ajouter des annotations manquantes, ou intégrer de nouvelles données expérimentales.

Bien que **Pathway Tools** soit un logiciel puissant, il dépend de la qualité des données d'entrée. Les prédictions de voies peuvent nécessiter une validation expérimentale, et les annotations automatiques sont parfois incomplètes, d'où l'importance d'une révision manuelle. De plus, la qualité et l'exhaustivité des PGDBs varient selon l'espèce et la couverture des voies dans les bases de référence (*e.g.* **MetaCyc**).

Pathway Tools est compatible avec des formats standards comme **SBML** pour faciliter l'exportation et l'intégration avec d'autres outils de modélisation métabolique. Il peut être utilisé en complément d'autres logiciels de bio-informatique, notamment dans des pipelines de reconstruction métabolique tel qu'**AuReMe**, pour affiner les modèles avant simulations quantitatives. Le **PGBD** de P. *rubens* a été obtenu en utilisant les paramètres par défaut de Pathologic, avec une faible sensibilité relative à l'élagage (*i.e. pruning*) taxonomique des pathways, visant à réduire le nombre de faux positifs (seuil de 0,15). Les fichiers résultants sont nécessaires à la création du sous-réseau d'annotations au sein d'**AuReMe**.

3.2. Annotation fonctionnelle et compartimentation subcellulaire

Outre les annotations nécessaires à la génération du draft, la connaissance de la localisation subcellulaire des enzymes est une aide précieuse pour déterminer la composition de l'environnement dans lequel l'organisme évolue et pour établir, de ce fait, la liste des réactions de transport et d'échange nécessaires à la simulation de différents milieux de culture.

Notre approche consiste à recueillir l'ensemble des produits et réactants des réactions dont la localisation est assignée au milieu extracellulaire. Ensuite, nous avons attribué pour chacun de ces composés une réaction d'absorption (*i.e. uptake*) et nous avons vérifié l'existence d'une réaction de transport entre le milieu intracellulaire et le milieu extracellulaire. Lorsqu'une telle réaction existait, la localisation des gènes associés a été vérifiée en utilisant les résultats fournis par **DeepLoc** combinés à ceux de **TMHMM** et **SignalP**. Seuls les gènes pour lesquels nous avions l'annotation membrane cellulaire ou la présence d'un domaine transmembranaire ont été conservés dans les associations GPR des réactions de transport. Dans les cas où nous n'avions pas de réaction de transport, et compte tenu de la capacité particulière des champignons filamenteux à effectuer une nutrition extracellulaire, nous avons créé des réactions de transport réversibles.

DeepLoc (version 1.0)  *(Almagro Armenteros et al. 2017)* **DeepLoc - 1.0**

DeepLoc est un outil bio-informatique récent qui, à l'aide de l'apprentissage profond (*i.e.* deep learning), prédit la localisation subcellulaire des protéines en utilisant exclusivement les informations de séquence. Cet outil repose sur des réseaux de neurones convolutifs profonds et son modèle a été entraîné sur des données

protéiques annotées expérimentalement et de haute qualité, extraites d'UniProt. **DeepLoc** se distingue ainsi de la plupart des méthodes utilisées jusqu'à présent, qui dépendent essentiellement de l'annotation d'homologues. En s'appuyant sur des séquences protéiques annotées et des caractéristiques structurelles, cet outil peut différencier dix localisations différentes : le noyau, le cytoplasme, le milieu extracellulaire, les mitochondries, la membrane cellulaire, le réticulum endoplasmique, les chloroplastes, les appareils de Golgi, les lysosomes/vacuoles et les peroxysomes.

DeepLoc v.1.0 a été téléchargé *via* https://services.healthtech.dtu.dk/. L'outil a été utilisé en version locale sur les 12 556 séquences protéiques de *P. rubens* Wisconsin 54-1255 téléchargées sur **EnsemblFungi** ainsi que sur les 31 séquences présentes dans ***i*Prub22** mais absentes du fichier protéique d'origine.

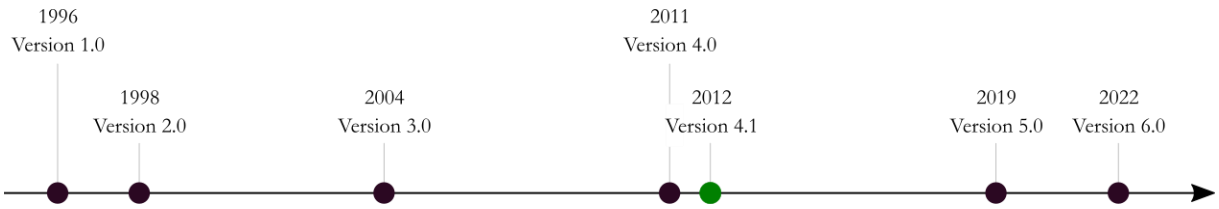
```
optional arguments:
  -h, --help                    show this help message and exit
  -f FASTA, --fasta FASTA
                                Input proteins in fasta file.
  -o OUTPUT, --output OUTPUT
                                Output prefix.
  -a, --attention               Generate file with attention values for each
                                protein.
deeploc -f sequences_fasta.faa -o DeepLoc_results.tsv -a
```

Nous avons sélectionné **DeepLoc** en raison de ses performances qui surpassent celles des algorithmes historiquement utilisés, y compris ceux basés sur l'homologie. **DeepLoc** parvient à assigner avec une précision globale de 78 % les produits géniques aux dix compartiments susmentionnés, et cette précision atteint 92 % pour les protéines membranaires ou solubles. Il est toutefois important de noter que, dans le cadre de l'annotation de séquences protéiques de champignons filamenteux, la catégorie « plastides » devrait rester vide. De plus, compte tenu de l'absence de lysosomes chez ces organismes, la catégorie « lysosomes/vacuoles » est exclusivement associée aux vacuoles.

SignalP (version 4.1g)  *(Petersen et al. 2011)* **SignalP - 4.1**

SignalP est un outil bio-informatique largement utilisé au sein de la communauté scientifique pour prédire l'existence des peptides signaux de sécrétion (PS) et des sites de clivage au sein des séquences protéiques. Historiquement, la détection de PS est l'un des premiers motifs à avoir été recherchés et étudiés sous l'angle de la bio-informatique.

Développé en 1996, **SignalP** est une méthode d'apprentissage automatique qui a reçu un accueil favorable au sein de la communauté scientifique. Il est également le premier outil de détection des PS à avoir été mis à disposition librement *via* un serveur web (*i.e.* version 1.0). Au fil des versions, des modifications structurales de l'algorithme ont été apportées pour améliorer notamment la distinction entre les PS et les hélices transmembranaires. Jusqu'à la version 4.0, cet outil reposait essentiellement sur une méthode basée sur des réseaux neuronaux artificiels. La version 4.1, que nous avons sélectionnée, présente une sensibilité de 68,3 % (*i.e.* pourcentage de sites de clivage réels prédits correctement) et une précision de 65,9 % (*i.e.* pourcentage de sites de clivage prédits qui sont corrects) pour la prédiction des sites de clivage des protéines d'origine eucaryote. La frise ci-dessous représente les dates de mise à jour de l'outil.



Il convient de mentionner que dans la version 2.0 de **SignalP**, un modèle de Markov caché (HMM) a été introduit en complément des réseaux neuronaux artificiels. Cependant, cette méthode a été abandonnée à partir de la version 4.0 en raison de performances similaires entre les deux approches. Pour améliorer la précision de la prédiction, des seuils de taille minimale de site de clivage ont été instaurés dans la version 4.1. Ces ajustements ont permis à la version 4.1 de surpasser les taux de sensibilité observés dans la version 3.0, outils et versions qui sont définis par des benchmarks indépendants comme étant les plus performants pour la prédiction de PS *(Nielsen 2017)*.

La version 5.0 de **SignalP**, quant à elle, repose sur des réseaux neuronaux convolutifs et récurrents, une architecture mieux adaptée à la reconnaissance de motifs de séquences de longueur variable. Dans la version suivante, 6.0, le modèle de langage transformateur utilisé a été entraîné sur un vaste ensemble de données de séquences protéiques non annotées afin d'améliorer les performances de prédiction, notamment pour les types de PS faiblement caractérisés. Les améliorations liées aux mises à jour incluent la capacité à discrétiser les PS en cinq types distincts pour les archées et les procaryotes, tandis que seulement un type de PS est détecté chez les eucaryotes. En outre, il est désormais possible de définir la position des trois régions biochimiques du PS, lesquelles sont responsables de leur fonction biologique. À long terme, il est plausible que ces avancées permettront d'affiner les prédictions d'adressage des protéines et de différencier ainsi celles destinées à être excrétées vers le milieu extracellulaire de celles vouées à circuler à l'intérieur de la cellule.

Enfin, nous tenons à souligner que **SignalP** a fait l'objet de plusieurs mises à jour significatives depuis la réalisation des travaux présentés dans ce manuscrit. Avec l'émergence des algorithmes d'apprentissage automatique, l'architecture de **SignalP** a été modifiée dans les versions 5.0 et 6.0, respectivement publiées en 2019 et 2022. Nous aurions pu utiliser la version 5.0 mais ne disposant alors que de peu de recul sur ses

performances et bien que ces améliorations soient significatives, leur pertinence est principalement axée sur la caractérisation des séquences protéiques appartenant à des règnes différents de celui des eucaryotes. Nous avons donc préféré utiliser la version 4.1, stable et largement plébiscitée jusqu'alors. Enfin, nous avons estimé qu'il n'était pas encore opportun de mettre à jour notre annotation étant donné que les informations nouvellement fournies ne pourraient être exploitables dans le cadre de notre reconstruction.

SignalP v.4.1g a été téléchargé sur le site https://services.healthtech.dtu.dk/ et a été utilisé en version locale avec les données précédemment mentionnées.

```
Description: Predict signal peptide and cleavage site.

Usage: /usr/bin/signalp -f <format> -p <graphics-type> -k -s <networks> -t
<organism-type> -m <fasta-file> -n <gff-file> -v -l <logfile> -u <value> -U
<value> -w -h -c <value> -T <temp dir> -V <fasta-file(s)>

Options:
 -f Setting the output format ('short', 'long', 'summary' or 'all')
    Default: 'short'
 -s Signal peptide networks to use ('best' or 'notm'). Default: 'best'
 -t Organism type> (euk, gram+, gram-). Default: 'euk'
 -n Make gff file of processed sequences. Default: 'Off'
 -M Minimal predicted signal peptide length. Default: [10]
 -c truncate to sequence length - 0 means no truncation. Default '70'
[…]

signalp -f short -n resultats.SignalP.out sequences_fasta.faa>
```

Pour chaque acide aminé de la séquence protéique, **SignalP** fournit trois scores :

- C-score : score brut du site de clivage qui indique le premier résidu de la protéine mature.
- S-score : score du PS.
- Y-score : score combiné du site de clivage, calculé comme moyenne géométrique du C-score et de la pente du S-score. Ce score permet de détecter le véritable site de clivage en présence de plusieurs pics liés au C-score.

La séquence en elle-même est ensuite caractérisée par les deux métriques suivantes :

- $S_{moyenne}$: score moyen du PS, calculé à partir du premier résidu jusqu'au résidu précédent le site de clivage putatif.
- D-score : score de discrimination correspondant à la moyenne pondérée des scores $S_{moyenne}$ et $Y_{max.}$

Dans le cadre des données de *P. rubens* Wisconsin 54-1255, nous avons utilisé uniquement le D-score et nous avons considéré qu'une séquence protéique possédait un PS si ce score était supérieur à 0.45.

Nous savons que le PS est un signal de triage ubiquitaire situé dans la région *N*-terminale des protéines. Néanmoins, la présence d'un PS au sein d'une séquence protéique ne garantit pas automatiquement sa sécrétion vers le milieu extracellulaire. Par conséquent, pour obtenir des informations plus précises et renforcer la fiabilité des résultats, il est essentiel de mettre en regard les résultats de **SignalP** avec ceux de **DeepLoc** et **TMHMM**.

En effet, la présence d'un PS indique que la protéine entre dans les voies de sécrétion. Selon la structure du PS, la protéine peut être dirigée vers différents compartiments tels que le réticulum endoplasmique, les mitochondries, les peroxysomes, les appareils de Golgi, le noyau ou les vacuoles. L'adressage des protéines est notamment conditionné par la composition en acides aminés du PS, généralement structuré en trois parties : la région *N*-terminale de longueur variable et dont la charge est positive, une région centrale hydrophobe et une région C-terminale composée de résidus polaires. En outre, en aval de la présence du PS, toute protéine peut contenir une ou plusieurs hélice α transmembranaire, détectées avec **TMHMM**, et ainsi être retenue dans les membranes. En revanche, en l'absence de PS, la protéine est généralement considérée comme demeurant dans le cytoplasme.

TMHMM (version 2.0c)  *(Krogh et al. 2001)* **TMHMM - 2.0**

TMHMM sert à la prédiction de la structure des protéines, et plus précisément, détecte, à partir de profils de modèles de Markov caché, l'existence d'hélices transmembranaires avec une précision allant jusqu'à 97-98 %. Depuis 2001, cette méthode, l'une des premières et la plus usitée depuis, permet de discriminer les protéines solubles des protéines membranaires avec une spécificité et une sensibilité supérieures à 99 %. Néanmoins, les résultats de **TMHMM** seuls, ne sont, en aucun cas, suffisants pour déterminer la localisation cellulaire des protéines.

Les membranes cellulaires sont formées de quatre types de composants : une bicouche de phospholipides, un réseau de protéines intracellulaires, des marqueurs de surface et des protéines transmembranaires. Ces dernières qualifiées d'intrinsèques en raison de la présence d'au moins une région hydrophobe qui traverse la membrane, représentent 20 à 30 % du protéome des organismes. De surcroît, elles revêtent un intérêt pharmaceutique majeur dans la recherche de médicaments, puisqu'elles constituent plus de 50 % de toutes les cibles médicamenteuses humaines.

En effet, ces protéines jouent un rôle primordial en assurant le transport des molécules à travers la cellule. Elles se répartissent en de nombreuses classes, dont les canaux ioniques, les aquaporines, les pompes ou les transporteurs enzymatiques. Un domaine transmembranaire est un segment hydrophobe qui traverse la membrane. Un seul de ces segments peut suffire à ancrer la protéine à la membrane, même si la plupart des protéines intrinsèques peuvent en posséder plusieurs. Ces segments transmembranaires sont organisés en hélices α cylindriques ou en feuillets β plats (*i.e.* structure secondaire) qui permettent la détection de motifs et de domaines fonctionnels au sein de ces protéines. À titre d'exemple, les porines sont des protéines dont les feuillets plissés β sont empilés les uns sur les autres, en forme de cercle, afin de former des segments apolaires appelés pores.

Il convient de souligner que **TMHMM** ne permet pas de caractériser les feuillets β et qu'une confusion entre PS et domaines transmembranaires situés en régions *N*-terminale existe.

TMHMM v.2.0c a été téléchargé *via* https://services.healthtech.dtu.dk/ et a été utilisé en version locale avec les données précédemment mentionnées.

```
tmhmm --short < sequences_fasta.faa > resultats.TMHMM.out
```

Les séquences dont la topologie comporte plus de deux caractères (*i.e.* une hélice α est composée d'au moins une séquence `[o|i]\d*-\d*[o|i]`- avec `o` pour outside, `i` pour inside et où les nombres correspondent à la position des résidus sur la séquence primaire) ont été annotées comme étant des protéines transmembranaires.Au cours des dernières années, **TMHMM** a été surpassé par les méthodes d'apprentissage profond. Aujourd'hui, DeepTMHMM *(Hallgren et al. 2022)* est la méthode la plus complète et performante pour prédire la topologie des deux classes de protéines transmembranaires au sein de protéomes complets. Comparé à son prédécesseur, ce nouvel outil permet non seulement d'améliorer la prédiction de la topologie des protéines transmembranaires en ce qui concerne la discrétisation des hélices α et des feuillets β, mais également, de détecter la présence de PS lorsqu'ils existent.

Ainsi, à l'avenir, l'utilisation conjointe de **DeepLoc** et de DeepTMHMM serait suffisante pour caractériser la localisation subcellulaire des produits géniques afin d'affiner les anciennes reconstructions ou de produire de nouveaux **GSMNs**.

3.3. Orthologie - Reconstruction du draft

Parallèlement au sous-réseau issu de l'annotation fonctionnelle, un sous-réseau axé sur les relations d'orthologie (*i.e.* séquences homologues résultant d'événements de spéciation) a été reconstruit.

La principale limite que nous pouvons opposer à cette approche réside dans la qualité des données d'entrée utilisées. En d'autres termes, la qualité des réactions inférées dépend intrinsèquement et *a minima* de la spécificité de chaque association GPR (*i.e.* ratio entre le nombre de vrais négatifs et la somme des vrais négatifs et des faux positifs – les faux positifs étant des gènes associés à tord à la réaction). Les questions relatives à la sensibilité des associations GPR (*i.e.* ratio entre le nombre de vrais positifs et la somme des vrais positifs et des faux négatifs) sont dommageables en termes de pertes d'information, mais n'auront pas d'incidence sur la topologie des voies inférées. Toutes les reconstructions publiées ne sont pas égales en termes de qualité et de grandes disparités ont pu être observées entre certains modèles. Ainsi, avec le recul, le choix des *templates* que nous avons utilisés pourrait évidemment être remis en question, mais nous espérons que la multiplicité des approches employées permet de contrebalancer ce biais, s'il existe.

Un résumé des résultats est présenté en *Annexe 2 : Reconstruction d'iPrub22 - sous-réseau issu des recherches d'orthologie*, page 433)

OrthoFinder (version 2.3.12)  *(Emms et Kelly 2019)*

OrthoFinder est un outil dédié à l'inférence des relations homologues entre les séquences de protéomes multiples. Il identifie les séquences partageant un ancêtre commun dans différentes espèces, facilitant ainsi les études de génomique comparative et la classification fonctionnelle des gènes. Cet outil offre une plateforme simple, rapide et précise pour la génomique comparative, en utilisant le principe du *Reciprocal Best Normalized Hit* (RBNH) pour détecter les relations d'orthologie et de paralogie *via* un alignement local des séquences protéiques.

Les résultats fournis par **OrthoFinder** permettent, entre autres, d'identifier des gènes orthologues pour une analyse comparative des génomes, d'explorer l'évolution des familles de gènes, de comprendre les événements de duplication, et de réaliser l'annotation fonctionnelle des gènes entre espèces. La liste des orthogroupes générée (*i.e.* clusters d'orthologues) est ensuite intégrée et interprétée dans **AuReMe** pour construire le sous-réseau d'orthologie.

Bien qu'**OrthoFinder** soit intégré dans **AuReMe**, l'analyse d'homologie ne peut s'effectuer qu'avec Diamond. La documentation d'**OrthoFinder** indique cependant qu'un alignement avec Blastp pourrait améliorer la précision des résultats d'environ 1 à 2 %. Pour cette raison, nous avons opté pour une utilisation locale d'**OrthoFinder**, puis intégré ses résultats dans **AuReMe**. La différence majeure entre ces deux algorithmes réside dans leur temps de calcul : Diamond est plus rapide, tandis que Blastp est plus gourmand en ressources. Afin d'optimiser l'analyse, un filtrage des protéomes des espèces *templates* a été réalisé, ne retenant que les séquences associées à au moins une réaction dans leur reconstruction respective.

```
OrthoFinder version 2.3.12 Copyright (C) 2014 David Emms

SIMPLE USAGE:
Run full OrthoFinder analysis on FASTA format proteomes in <dir>
   orthofinder [options] -f <dir>

OPTIONS:
 […]
 -S <txt>     Sequence search program [Default = diamond]
              Options: blast, diamond, blast_gz, mmseqs
[…]
```

`orthofinder -S blast* -f Dossier_contenant_les_séquences_protéiques > log.err`

***** (blastp -outfmt 6 -evalue 0.001)

3.4. Reconstruction

AuReMe (version 2.4)  *(Aite et al. 2018)*

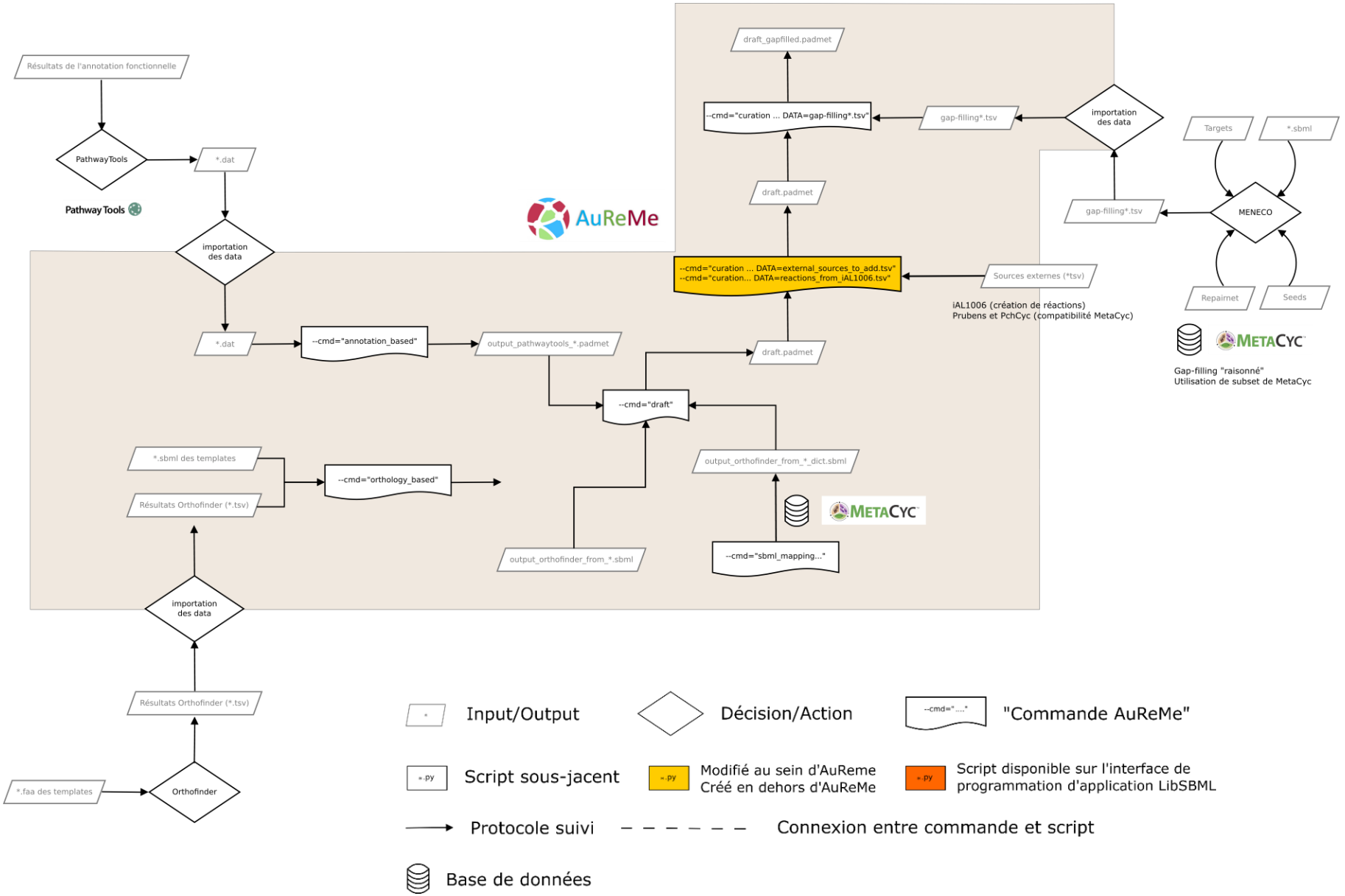
AuReMe (***Au****tomatic* ***Re****construction of* ***Me****tabolic networks*) est une plateforme « à-la-carte » de reconstruction de réseaux métaboliques à l'échelle du génome. Selon les critères de reconstruction établis par la communauté scientifique, ce workflow permet de créer des modèles de haute qualité en se basant sur des données génomiques. Comparé à d'autres pipeline de reconstruction automatique de **GSMNs** : CarveMe, MetaDraft, Merlin., ModelSEED, **Pathway Tools** et RAVEN, **AuReMe** présente des caractéristiques relativement similaires voire légèrement plus compétitives en termes de traçabilité, d'automatisation, et de support pour la curation manuelle *(Mendoza et al. 2019)*.

Conçu pour simplifier les processus de reconstruction de modèles de génomes peu étudiés, **AuReMe** fournit un environnement modulaire, flexible et centralisé qui intègre l'ensemble des outils et ressources nécessaires à l'automatisation des reconstructions, du draft au **GSMN** de haute qualité (*e.g.* fichiers des bases de données **MetaCyc** et MetaNetX, **OrthoFinder**, **Meneco**, **Mene Tools**, le package PADMet, CobraPy, *etc.*).

L'un des principaux atouts d'**AuReMe**, un aspect que nous avons particulièrement apprécié, réside dans son déploiement *via* une image Docker. Ce mode de fonctionnement confère à l'utilisateur une autonomie totale en lui permettant d'administrer et donc de personnaliser à volonté l'ensemble de son environnement de travail sans restriction. Cette structure garantit une transparence complète sur chaque étape du processus de reconstruction métabolique, facilitant la visualisation et le suivi de la chaîne d'actions nécessaires pour obtenir un modèle fonctionnel. Nous avons ainsi pu décortiquer, intervenir, contrôler, voire modifier selon nos besoins, chaque étape du processus de reconstruction (**Figure 2-2**).

Par ailleurs, le processus de reconstruction est documenté de manière exhaustive *via* deux logs distincts : un log synthétique (présenté ci-dessous), résumant les étapes clés de la reconstruction d'***i*Prub22**, et un log détaillé pour chaque traitement effectué sur le réseau en cours de génération. Ces éléments garantissent ainsi une traçabilité complète des opérations réalisées sur le réseau.

```
### LOG ###

aureme --run=P_rubens --cmd="check_input"

# sub-network annotations
aureme --run=P_rubens --cmd="annotation_based"

# sub-network orthologie
aureme --run=P_rubens --cmd="orthology_based"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_A_thaliana.sbml DB=METACYC"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_A_nidulans.sbml DB=METACYC"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_A_niger.sbml DB=METACYC"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_N_crassa.sbml DB=METACYC"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_Pench1.sbml DB=METACYC"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_S_japonica.sbml DB=METACYC"
aureme --run=P_rubens --cmd="sbml_mapping SBML=output_orthofinder_from_S_pombe.sbml DB=METACYC"

# merging data
aureme --run=P_rubens --cmd="draft"

# Adding data from external sources
aureme --run=P_rubens --cmd="curation NETWORK=draft1 NEW_NETWORK=draft2 DATA=external_sources_toadd.tsv"
aureme --run=P_rubens --cmd="curation NETWORK=draft2 NEW_NETWORK=draft3 DATA=reactions_from_iAL1006.tsv"

# Adding modelisation reaction
aureme --run=P_rubens --cmd="curation NETWORK=draft3 NEW_NETWORK=draft4 DATA=transport_reactions.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft4 NEW_NETWORK=draft5 DATA=demand_reactions.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft6 NEW_NETWORK=draft6 DATA=production_reactions.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft6 NEW_NETWORK=draftT7 DATA=uptakes_reactions.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft7 NEW_NETWORK=draft8 DATA=spontaneous_reactions.csv"

# Adding reactions from gap-filling results
aureme --run=P_rubens --cmd="curation NETWORK=draft8 NEW_NETWORK=draft9 DATA=gap-filling1.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft9 NEW_NETWORK=draft10 DATA=gap-filling2.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft10 NEW_NETWORK=draft11 DATA=gap-filling3.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft11 NEW_NETWORK=draft12 DATA=gap-filling4.csv"
aureme --run=P_rubens --cmd="curation NETWORK=draft12 NEW_NETWORK=iPrub22 DATA=biomass_rxn.csv"

# Test & report
aureme --run=P_rubens --cmd="set_fba ID=Biomass_rxn NETWORK=iPrub22"
aureme --run=P_rubens --cmd="summary NETWORK=iPrub22"
aureme --run=P_rubens --cmd="wiki_pages NETWORK=iPrub22"
aureme --run=P_rubens --cmd="report NETWORK=iPrub22"
```


Résultats de l'annotation fonctionnelle
PathwayTools
Pathway Tools
*.dat
importation des data
AuReMe
*.dat
--cmd="annotation_based"
output_pathwaytools_*.padmet
draft_gapfilled.padmet
--cmd="curation ... DATA=gap-filling*.tsv"
gap-filling*.tsv
importation des data
Targets
*.sbml
draft.padmet
gap-filling*.tsv
MENECO
--cmd="curation ... DATA=external_sources_to_add.tsv"
--cmd="curation... DATA=reactions_from_iAL1006.tsv"
Sources externes (*tsv)
iAL1006 (création de réactions)
Prubens et PchCyc (compatibilité MetaCyc)
Repairnet
Seeds
METACYC
Gap-filling "raisonné"
Utilisation de subset de MetaCyc
draft.padmet
--cmd="draft"
*.sbml des templates
Résultats Orthofinder (*.tsv)
--cmd="orthology_based"
output_orthofinder_from_*_dict.sbml
METACYC
output_orthofinder_from_*.sbml
--cmd="sbml_mapping..."
importation des data
Résultats Orthofinder (*.tsv)
*.faa des templates
Orthofinder
Input/Output
Décision/Action
--cmd="...."
"Commande AuReMe"
*.py
Script sous-jacent
*.py
Modifié au sein d'AuReme
Créé en dehors d'AuReMe
*.py
Script disponible sur l'interface de programmation d'application LibSBML
Protocole suivi
Connexion entre commande et script
Base de données

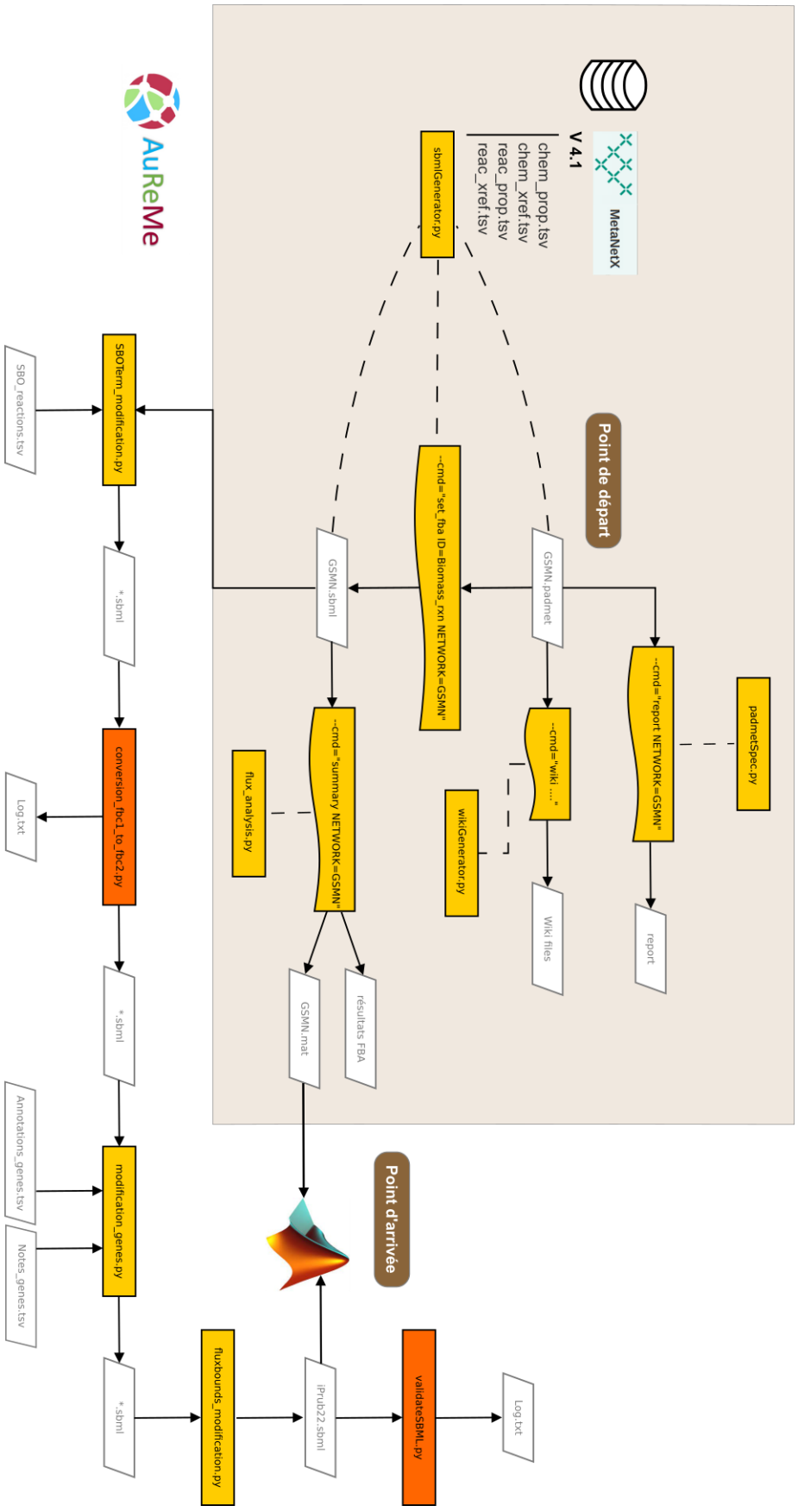


Figure 2-2 : La Toolbox AuReMe au cœur du protocole de reconstruction d'iPrub22. *Résumé graphique de l'enchaînement des étapes pour l'obtention du modèle iPrub22. La figure à gauche illustre la génération des sous-réseaux d'annotation et d'orthologie, leur intégration progressive, et les étapes de gap-filling. La figure de droite détaille le processus allant du draft gap-fillé jusqu'à la finalisation du modèle métabolique iPrub22.*

3.5. Reconstruction – curation & analyses topologiques

Le comblement des lacunes d'un **GSMN** (*i.e. gap-filling*) constitue une étape cruciale et itérative dans le processus de reconstruction de réseaux métaboliques. Cette étape permet de s'assurer que le réseau dispose de toutes les connexions nécessaires pour produire les métabolites indispensables à la croissance et à la survie de l'organisme modélisé. Dans ce contexte, nous avons employé **Meneco** et la suite d'outils **Mene Tools** pour évaluer la productibilité topologique des métabolites. Ces analyses qualitatives, ne tenant pas compte de la stoechiométrie des réactions, peuvent être réalisées dans toutes les étapes de vie d'un modèle, du simple draft au **GSMN** de haute qualité. **Meneco** et **Mene Tools** appliquent une approche basée sur l'algorithme d'expansion de réseau (*Network Expansion Algorithm*) régi par deux règles :

- **Règle d'initiation** : la productibilité d'un composé est initiée par la présence de nutriments, appelés *seeds*.
- **Règle de récursivité** : les produits d'une réaction sont considérés comme productibles si tous les réactants sont eux-mêmes productibles.

Il est essentiel de noter que les solutions proposées durant le *gap-filling* ne sont pas uniques et dépendent fortement des données d'entrée fournies à l'algorithme. Cet aspect, ainsi que ses implications, sont approfondis dans la section dédiée du Chapitre 3, *2.2.1 Un gap-filling « raisonné »*, page 239. Ces outils sont intégrés dans AuReMe, mais pour des questions de praticité et de manipulations (*i.e.* alterner plus facilement entre les diverses combinaisons de *seeds*, de cibles et de réseaux de réparation), ils ont été installés et utilisés localement.

Meneco (version 2.0.0)  *(Prigent et al. 2017)*

Meneco (***Me****tabolic* ***Ne****twork* ***Co****mpletion*) est un algorithme de reconstruction métabolique basé sur une approche par complétion de réseau en utilisant la programmation d'ensembles de réponses, un paradigme de programmation déclarative destiné à résoudre des problèmes de recherche combinatoire. Développé pour guider la reconstruction de réseaux métaboliques, **Meneco** identifie les réactions manquantes dans les modèles incomplets afin d'assurer la production de biomasse ou de métabolites d'intérêt dans des conditions spécifiques. **Meneco** s'appuie sur la théorie des graphes et des calculs d'atteignabilité pour ajouter le minimum de réactions nécessaires dans un réseau métabolique donné, maintenant ainsi une approche parcimonieuse. Ce principe d'économie permet de minimiser l'ajout de réactions non-essentielles, favorisant ainsi la cohérence biologique du réseau complété.

Les résultats obtenus avec **Meneco** reposent sur une heuristique, rendant le *gap-filling* particulièrement sensible à la topologie du réseau et à ses quatre paramètres d'entrée : le réseau à explorer (draft), la liste de métabolites d'initiation (*seeds*), la liste de métabolites cibles (*targets*) et un ensemble de réactions potentielles à interroger (réseau de réparation).

```
optional arguments:
   -h, --help            show this help message and exit
   -d DRAFTNET, --draftnet DRAFTNET
                         metabolic network in SBML format
   -s SEEDS, --seeds SEEDS
                         seeds in SBML format
   -t TARGETS, --targets TARGETS
                         targets in SBML format
   -r REPAIRNET, --repairnet REPAIRNET
                         perform network completion using REPAIRNET a metabolic
                         network in SBML format
   --enumerate           enumerate all minimal completions

meneco -d draft.sbml -s seeds.sbml -t targets.sbml -r repairnetwork.sbml
```

Mene Tools (version 3.3.0)  *(Belcour et al. 2020)*

Mene Tools est une suite d'outils Python dédiée à l'analyse topologique de la productibilité des composés dans un réseau métabolique. Facile à utiliser, ces outils, et plus particulièrement **check**, **scope**, **acti** et **cof** nous ont permis, d'une part, de guider les étapes de *gap-filling* ainsi que de filtrer ses résultats, et d'autre part, d'affiner la liste de *seeds* servant à l'environnement nutritionnel d'***i*Prub22** pour la réalisation de l'approche **OSMAC *in silico*** (cf. Chapitre 3, section *2.2.2 Modélisation de différentes conditions de culture*, page 256).

```
Explore the producibility potential in a metabolic network using the network
expansion algorithm. For specific help on each subcommand use: mene {cmd} --help

subcommands:
 valid subcommands:

 {acti,check,cof,dead,path,scope,seed,scope_inc}
    acti        Get activable reactions in a metabolic network,
                starting from seeds.
    check       Check the producibility of targets from seeds in a
                metabolic network.
    cof         Propose cofactor whose producibility could unblock the
                producibility of targets.
    scope       Get producible metabolites in a metabolic network,
                starting from seeds.
[…]

mene acti -d draft.sbml -s seeds.sbml > out_meneacti_draft.txt
mene scope -d draft.sbml -s seeds.sbml > out_menescope_draft.txt
mene cof -d draft.sbml -s seeds.sbml -t targets.sbml > out_menecof_draft.txt
mene check -d draft.sbml -s seeds.sbml -t targets.sbml > out_menecheck_draft.txt
```

Lorsqu'il s'agit d'étudier le métabolisme dans sa globalité ou d'explorer une voie métabolique spécifique, le recours à la visualisation pour appréhender l'agencement des données constitue une aide précieuse à la compréhension des modèles. En effet, même si l'exploration visuelle de grands modèles métaboliques demeure un véritable défi en raison de la taille et de la complexité de ces réseaux qui peuvent inclure des milliers de métabolites et de réactions, inspecter visuellement les modèles est une solution efficace pour identifier des incohérences ou des discontinuités topologiques.

L'approche la plus basique consiste à retrouver puis dessiner manuellement ou semi-manuellement les voies métaboliques ou sections d'intérêt. Les bases de données métaboliques BioCyc et KEGG proposent d'ailleurs toutes deux leur propre module de cartes de visualisation. Cette simplification vise à rendre les données visuellement inspectables, mais peut également limiter la capacité à obtenir une vue complète d'une voie métabolique, car les voies équilibrant les co-substrats, les co-produits, les pouvoirs réducteurs et l'énergie ne sont pas systématiquement associées aux représentations simplifiées. En revanche, au cours des dernières décennies, plusieurs outils de visualisation particulièrement adaptés à l'analyse de réseaux métaboliques à l'échelle du génome ont été développés. À titre d'exemple, Cytoscape *(Su et al. 2014)*, MetExplore *(Cottret et al. 2018)* ou LMME *(Aichem et al. 2021)* sont des outils qui facilitent grandement l'interprétation des flux métaboliques, en permettant une exploration dynamique et interactive des réactions et des métabolites impliqués.

L'utilisation d'outils interactifs offre des perspectives nouvelles pour la filtration, l'exploration, la navigation et la ségrégation de données *(Pienta et al. 2015)*. Ainsi, les méthodes de visualisation automatique et d'exploration interactive permettent de mieux comprendre les modèles métaboliques tout en aidant à identifier les erreurs dans un modèle, à trouver les lacunes métaboliques, à vérifier la connectivité entre les métabolites, à tester l'impact de la réversibilité ou d'une suppression de réaction et à soutenir les comparaisons de modèles. Ces diverses stratégies ont été explorées lors des phases de curation d'***i*Prub22**, tant pour les tests de *gap-filling* que pour ceux de la fonctionnalité du modèle.

Cependant, la visualisation simultanée de grandes quantités de données peut rapidement devenir gourmande en ressources computationnelles. Combinées à des approches interactives (*e.g.* navigation, partitionnement des voies métaboliques, tests en temps réel de l'ajout ou de la suppression de réactions, *etc.*), la rapidité de réponse de l'outil (*i.e.* sa rapidité à charger et traiter les données), sa fluidité et sa facilité d'utilisation sont des critères majeurs qui nous ont orientés vers l'utilisation de **ModelExplorer** puis de **Fluxer**, détaillés ci-après.

ModelExplorer (version 2.1)  *(Martyushenko et Almaas 2019)*

ModelExplorer est un outil conçu pour identifier les incohérences d'un modèle lors des processus de construction et de raffinement. Cet outil polyvalent aide à comprendre la structure et la dynamique des **GSMNs** et a été développé au laboratoire Almaas, un centre de recherche norvégien axé sur la biologie des systèmes. En combinant méthodes informatiques et expérimentales, ce laboratoire étudie les réseaux biologiques complexes. Les outils et méthodes, mis au point dans leurs recherches, y compris **ModelExplorer**, sont disponibles pour la communauté scientifique *via* leur site internet (https://almaaslab.nt.ntnu.no/). **ModelExplorer** est une application graphique autonome compatible sous Linux et Windows.

Avec **ModelExplorer**, les **GSMNs** sont visualisés sous forme de graphes bipartites où les nœuds correspondent aux métabolites et aux réactions, et les arrêtes représentent les liens existants entre ces deux entités. Son interface graphique permet de naviguer, visualiser et éditer des modèles métaboliques. Des outils interactifs y sont proposés pour détecter les réactions bloquées et retracer les origines des erreurs (*i.e.* impasse métabolique, réversibilité et cycle), notamment à l'aide de l'algorithme ErrorTracer *(Martyushenko et Almaas 2020)*. Cet outil nous a été utile à plusieurs degrés pour visualiser, explorer, comparer et corriger les divers modèles utilisés dans nos travaux (*i.e.* **GSMNs** des *templates* servant à la génération des sous-réseaux d'orthologie, *i*AL1006 et Prubens, ainsi que ***i*Prub22** et ses versions intermédiaires).

Le diagnostic des réactions bloquées (*i.e.* sections non-fonctionnelles du réseau) s'effectue par une approche dichotomique qui évalue la présence ou l'absence de flux à partir d'une analyse FBA. Cette analyse propose trois options : FBA, mode bidirectionnel et mode dynamique. Dans le premier cas, il s'agit d'une analyse classique, dans le second cas, toutes les réactions sont rendues réversibles. La comparaison des résultats de ces deux méthodes (*e.g.* présence de réactions actives avec la méthode bidirectionnelle mais inactives dans le premier cas) permet de concentrer les efforts de résolution d'erreurs sur la directionnalité des réactions. Dans le cadre de la curation d'***i*Prub22**, cette approche s'est avérée précieuse, notamment lors des tests de débogage pour activer la réaction de production des cofacteurs, un des sous-systèmes de la réaction de biomasse. Enfin, le mode dynamique a été spécifiquement conçu pour surveiller la fonctionnalité de la réaction de biomasse. Cette fonctionnalité permet de visualiser et d'identifier les métabolites existants qui pourraient potentiellement participer à la réaction de biomasse sans en compromettre l'équilibre.

En outre, **ModelExplorer** permet une interaction rapide, intuitive et dynamique avec la topologie du réseau métabolique en offrant une visualisation claire des métabolites et des réactions. Une recherche textuelle directe permet de localiser un métabolite d'intérêt ou une réaction spécifique en fonction de leur nom. L'affichage peut être personnalisé en fonction des besoins de l'utilisateur : il est possible de zoomer sur des sous-réseaux spécifiques, d'activer ou désactiver certaines réactions, de modifier les couleurs pour distinguer les catégories de métabolites ou de réactions, et d'ajuster les représentations graphiques selon les données. Ces fonctionnalités rendent l'identification des erreurs plus intuitive, notamment les incohérences dans la connectivité entre les voies. L'outil permet également de simuler des perturbations dans le réseau (*e.g.* suppression ou ajout de réactions ou métabolites) et d'observer leur impact sur l'ensemble du modèle. Un export graphique de l'intégralité ou de portions du réseau est également disponible. Ces diverses options permettent une exploration préliminaire du réseau et nous ont permis de cibler plus rapidement certains points bloquants lors de la curation des drafts. Afin de déterminer la source de l'erreur constatée, en termes d'exploration, **ModelExplorer** propose quatre modes de visualisation qui mettent en surbrillance les données avec des degrés de complexité différents.

- *None* : Seul le nœud sélectionné est affiché en surbrillance et ces métadonnées sont accessibles. Cette visualisation nous permet d'entrer dans le mode d'édition du modèle et de tester directement l'impact et la pertinence des corrections apportées au modèle (*i.e.* utilisé principalement pour débloquer certains sous-systèmes de la réaction de biomasse).
- *Ego-Centric* : Le nœud d'intérêt est affiché en surbrillance ainsi que ces voisins directs (*i.e.* réactant et produits si le nœud sélectionné est une réaction, réactions de consommation et de production si le nœud sélectionné est un métabolite). Ce mode nous a particulièrement été utile pour identifier rapidement les réactions de transport et d'échange manquantes dans notre modèle.
- *Node ancestry* : Le nœud sélectionné est en surbrillance ainsi que le plus petit sous-ensemble d'éléments antécédents nécessaires soit à la synthèse du composé si le nœud est un métabolite, soit à l'activation de la réaction si le nœud initial est une réaction. En présence de cycles dans la reconstruction, ce mode de visualisation permet de les détecter aisément. En revanche, identifier leur cause exacte et les solutions à leur résolution dépend du nombre d'éléments connecté et donc de leur niveau de complexité.
- *Blocked Module* : En survolant un élément bloqué, tout le module (*i.e.* défini comme un ensemble de nœuds bloqués et non connectés à d'autres modules) correspondant est mis en surbrillance et nous pouvons alors choisir d'afficher chaque module bloqué séparément du reste du réseau, rendant plus simple l'identification des sources d'incohérence en réduisant l'encombrement visuel.

Enfin, **ModelExplorer** se distingue par son extrême simplicité d'installation et d'utilisation. Lors de nos expérimentations, les temps de chargement et de traitement se sont révélés très rapides : les modèles métaboliques se chargent en quelques secondes, et même les analyses les plus intensives n'ont pris que quelques minutes. Le lancement de l'interface graphique depuis un terminal permet également de suivre les erreurs et avertissements potentiels, souvent liés à des problèmes d'encodage des modèles analysés. Cette interface fournit un résumé concis des propriétés du modèle et des résultats d'analyses de flux effectuées, qui, bien que succinctes, offre une première évaluation de la qualité du modèle et des pistes d'améliorations envisageables.

```
./ModelExplorer iPrub22.sbml
```

Résumé des propriétés d'*iPrub22*

```
////////////////////////////////////////////////////////////////////////////////
//                                                                            //
//   Opened Model:          Flux Unit:             Species Number:        Reaction Number:   //
//   iPrub22_v1             mmol_per_gDW_per_hr    5464                   5919               //
//                                                                            //
////////////////////////////////////////////////////////////////////////////////

3563 (uni: 3217 bidir: 346) reactions found blocked by the FBA ErrorTracer method in 7(7) iterations!
2805 reactions found blocked by the Bidirectional ErrorTracer method in 261(261) iterations!
2140 reactions and 2877 species found blocked by the Dynamic method!
```

Fluxer (version 2.10) *(Hari et al. 2020)*

Fluxer est un outil interactif conçu spécifiquement pour la visualisation et le calcul par FBA des flux métaboliques d'un **GSMN**. Il se distingue des autres outils par son orientation axée sur l'analyse quantitative des flux plutôt que la simple représentation graphique des réseaux métaboliques. Utilisée en complément de **ModelExplorer**, cette application web intuitive nous a été particulièrement utile pour la curation manuelle et pour l'exploration des sections actives d'***i*Prub22** sous différentes conditions environnementales. Par exemple, l'intensité des flux peut être visualisée par l'épaisseur des arêtes, permettant ainsi une identification rapide des voies métaboliques les plus sollicitées et des points de blocage dans le réseau. Cette caractéristique aide à détecter des flux métaboliques « anormaux », potentiellement indicateurs de connectivités incomplètes, d'erreurs de paramétrage, ou de contraintes mal définies dans le modèle.

Cette application web en expansion offre une interface interactive qui permet aux utilisateurs d'explorer dynamiquement les flux dans les réseaux métaboliques. Les modèles sont représentés sous forme de graphe bipartite où les nœuds représentent les métabolites ou les réactions et les arêtes orientées indiquent l'implication d'un métabolite dans une réaction, soit en tant que produit, soit en tant que réactif. Trois modes d'exploration sont proposés :

- **Mode complet** : affiche tous les nœuds de métabolites et de réactions du modèle.
- **Mode *spanning tree*** : montre les connexions clés entre toutes les réactions et les métabolites menant à un nœud racine, par défaut, la réaction de la fonction objective.
- **Mode *k-shortest paths*** : calcule et affiche les k-plus courts chemins entre deux nœuds sélectionnés par l'utilisateur.

Diverses options de personnalisation viennent s'ajouter à ces modes de visualisation telles que le choix d'afficher ou de masquer les réactions sans flux, la localisation cellulaire, les cofacteurs ou la pondération des arêtes. **Fluxer** facilite également la simulation de *knock-out*s en permettant un nombre illimité de suppressions de réactions. À chaque élimination, une nouvelle FBA est exécutée, et le graphe est mis à jour en fonction des nouveaux flux calculés.

Pour charger un modèle, **Fluxer** propose trois options : accéder directement aux modèles préchargés d'organismes modèles (*i.e.* plus d'une centaine), importer un fichier **SBML** fourni par l'utilisateur, ou charger un modèle *via* son identifiant Biomodel. Ainsi, la reconstruction paramétrée ***i*Prub22** peut être directement visualisée à l'aide de son identifiant Biomodel (MODEL2306150001). Compte tenu de la taille de notre réseau métabolique, son temps de chargement est relativement rapide. En revanche, il est vivement recommandé de cocher l'option « hide » pour les réactions ne présentant aucun flux (*i.e.* réactions bloquées) avant de poursuivre l'exploration du modèle.

Enfin, soulignons que **Fluxer** permet de générer automatiquement un rapport FROG, fournissant des ensembles de données numériques permettant d'assurer la reproductibilité des résultats et la conservation des modèles. Ce rapport inclut notamment les résultats des analyses FVA et la valeur de flux de la fonction objective étudiée, garantissant une documentation détaillée et fiable des simulations effectuées.

3.6. Reconstruction – qualité

Dans nos travaux, les questions relatives à la définition et à la caractérisation de la qualité d'un modèle ont constitué des préoccupations centrales et font l'objet d'une section dédiée dans le Chapitre 3, section *2.1.3 Un GSMN de haute qualité ?*, page 210. **MEMOTE** est, à notre connaissance, le seul outil actuel capable d'évaluer et de quantifier la qualité des reconstructions et des modèles de réseaux métaboliques. Des captures d'écran du rapport généré par cet outil sont fournies en Annexe 4 : *LiveScript MATLAB détaillant les caractéristiques d'iPrub22*, page 451, tandis que sa version interactive, disponible au format HTML, est incluse dans les données supplémentaires déposées sur BioModels en même temps qu'***i*Prub22**.

MEMOTE (version 0.13.0) *(Lieven et al. 2020)*

MEMOTE (***Me****tabolic* ***Mo****del* ***Te****sts*) est un logiciel *open-source* en Python conçu pour assurer la validité structurelle et l'intégrité formelle des **GSMNs**. Cet outil vise à combler l'absence de standards unifiés de contrôle qualité, malgré les directives existantes pour la traçabilité et l'interopérabilité lors de la reconstruction de modèles métaboliques. **MEMOTE** permet d'évaluer les réseaux métaboliques codés en SBML3 FBC, le standard actuel, ainsi que les versions antérieures.

MEMOTE évalue les modèles métaboliques en exécutant plus de 300 tests consensuels dans quatre catégories principales : l'annotation, les tests de base, la réaction de biomasse et la stœchiométrie. Les modèles testés sont notés selon des critères reconnus par la communauté, et un rapport visuel détaillé est généré, affichant un score de qualité en pourcentage. Bien que ce score donne une indication de la conformité structurelle, il nécessite une interprétation critique, car il reflète davantage l'aspect syntaxique (*i.e.* la conformité au format **SBML**) que le contenu scientifique et biologique du modèle.

Dans le cadre de notre travail, l'utilisation de **MEMOTE** s'est exclusivement concentrée sur la compréhension des rapports de qualité pour assurer la curation d'***i*Prub22**. Nous n'avons pas exploité l'ensemble des fonctionnalités avancées, telles que l'intégration continue, qui permettrait de suivre l'évolution des modifications et des améliorations du modèle au fil du temps.

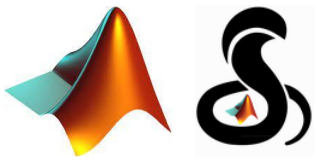
```
Metabolic model testing command line tool.
In its basic invocation memote runs a test suite on a metabolic model.
Through various subcommands it can further generate three types of pretty
HTML reports (snapshot, diff and history), generate a model repository
structure for starting a new project, and recreate the test result history.
```

```
      Take a snapshot of a model's state and generate a report.

Options:
  -h, --help                  Show this message and exit.
  --filename PATH             Path for the HTML report output.
  -a, --pytest-args TEXT      Any additional arguments you want to pass to
                              pytest. Should be given as one continuous string.
  --exclusive TEST            The name of a test or test module to be run
                              exclusively. All other tests are skipped.
                              This option can be used multiple times and
                              takes precedence over '--skip'.
  --skip TEST                 The name of a test or test module to be skipped.
                              This option can be used multiple times.
  --solver [cplex|glpk|gurobi|glpk_exact]
                              Set the solver to be used. [default: glpk]

  memote report snapshot --filename report.html iPrub22.sbml
```

3.7. Analyses

MATLAB (v2018b)
COBRA Toolbox (v3.0)

(Heirendt et al. 2019)



MATLAB (***Ma****trix* ***Lab****oratory*) est un environnement de calcul et de programmation de haut niveau, principalement utilisé pour l'analyse numérique, le traitement de données, et le développement d'algorithmes. Créé initialement par Cleve Moler dans les années 1960-70 pour offrir une « calculatrice matricielle interactive » accessible aux étudiants et chercheurs sans expertise en Fortran, **MATLAB** a évolué pour répondre aux besoins en calcul matriciel en algèbre linéaire *(Moler 1980)*. Sa popularité croissante a conduit à sa commercialisation par MathWorks, un éditeur de logiciels américain fondé en 1984 qui se spécialise aujourd'hui dans les logiciels de calcul mathématique.

Aujourd'hui, Mathworks se positionne comme un leader mondial dans les logiciels de calcul scientifique et ses outils sont utilisés dans plus de 100 000 entreprises et universités. **MATLAB** est particulièrement apprécié dans des domaines comme l'aérospatiale, la défense, l'automobile, les dispositifs médicaux, les services financiers, mais également dans la recherche et l'enseignement. Avec une communauté mondiale de plus de 5 millions d'utilisateurs dans le monde (données extraites sur la page de présentation de MathWorks), **MATLAB** est devenu incontournable, notamment dans les disciplines comme la bio-informatique du fait de sa flexibilité liée aux « Toolboxes » spécialisées qui étendent ses fonctionnalités pour répondre à des besoins spécifiques.

La **COBRA Toolbox** (***Co****nstraint-****B****ased* ***R****econstruction and* ***A****nalysis*) est une extension de **MATLAB** conçue pour la reconstruction et l'analyse de modèles de réseaux métaboliques. Apparue dans les années 2000, en réponse à l'essor des sciences omiques et aux avancées en bio-informatique elle s'est imposée dans les domaines tels que la biologie des systèmes et la modélisation métabolique. À l'aide des capacités de calcul rapides et modulables de **MATLAB**, la **COBRA Toolbox** permet de construire, importer, formater, manipuler et analyser des modèles métaboliques avec une flexibilité de programmation.

Cet outil est fréquemment utilisé pour la modélisation des flux métaboliques, la prédiction de phénotypes cellulaires, et l'analyse de robustesse pour évaluer la stabilité des réseaux face aux perturbations. Il est également appliqué en bio-ingénierie pour l'optimisation de production de biomolécules. En tant que ressource de référence en modélisation métabolique, **COBRA Toolbox** a été choisie pour notre travail, en association avec **MATLAB** 2018b, afin de garantir la compatibilité avec Gurobi v9.03, le solveur de notre installation.

Une limitation notable de **COBRA Toolbox** est sa dépendance à **MATLAB**, un logiciel commercial qui requiert une licence payante, pouvant freiner l'accessibilité d'*i*Prub22 et des tests proposés dans le Livescript mis à disposition avec sa publication. Bien que le contenu du fichier Livescript soit consultable en HTML, une transition vers COBRApy, alternative *open-source* en Python, semble prometteuse pour une diffusion plus large et accessible, alignée avec les principes de la recherche ouverte.

4 - Base de données

Vocabulaire

Bien que les termes « bases » et « banques de données » soient apparentés et souvent interchangés de nos jours, des différences existent entre ces deux notions.

BASE DE DONNÉES : le contenant

Une base de données désigne un **ensemble structuré** de données stockées dans un système informatique. Ces données sont généralement organisées en tableaux, où chaque ligne représente une entrée unique, et chaque colonne, un attribut de cet enregistrement. Ces enregistrements peuvent alors y être facilement **stockés**, **modifiables** (*e.g.* ajout, suppression) et **accessibles** au moyen d'un **système de gestion de bases de données**. À souligner que, lors de la phase de conception d'une base, et en vue de gérer le traitement de tâches complexes, divers aspects liés à la modélisation des données doivent être pris en compte, tels que :

le choix de représentation des données,
le stockage efficace,
la sécurité des accès,
le niveau de confidentialité, notamment pour les données sensibles,
le langage d'interrogation de la base employée,
etc..

Banque de données : le contenu

Une banque de données constitue un **dépôt centralisé** d'informations qui rassemble une collection de données spécifiques sur un domaine particulier. Ces données sont organisées de manière à faciliter leur **stockage**, leur **extraction** et leur **analyse**. Bien que le terme de « banque de données » ait tendance à être remplacé par celui de « base de données », ce dernier demeure fréquemment employé dans le domaine de la recherche scientifique.

EnsemblFungi *(Yates et al. 2022)*

Ensembl est un projet scientifique de génomique dirigé pat l'Institut Européen de Bio-informatique du Laboratoire Européen de Biologie Moléculaire (*European Molecular Biology Laboratory's European Bioinformatics Institute* - EMBL-EBI), initié au début des années 2000.

Cette base de données spécialisée, a pour objectif de constituer une ressource complète pour recueillir, annoter et présenter des données génomiques au public *via* le web. Elle rassemble et maintient des informations provenant d'une variété d'espèces vivantes. Elle offre une richesse de données, notamment sur les gènes, les protéines, les séquences d'ADN et d'ARN et propose de nombreux outils d'analyse et de visualisation de génomes. Ces outils permettent d'explorer en détail les interactions entre gènes, les variations génétiques et les régions régulatrices des génomes, offrant ainsi une vue approfondie des mécanismes moléculaires.

Depuis sa création, **Ensembl** s'est étendue dans des domaines tel que la génomique comparative et dispose de portails web spécifiques aux données génomiques de bactéries, de protistes, de métazoaires invertébrés, de plantes et de champignons. En particulier, **EnsemblFungi**, le portail dédié aux génomes fongiques, répertoriait en septembre 2023 (release 57) 1 505 génomes de champignons.

Les données génomiques de *P. rubens* que nous avons utilisées dans le cadre des travaux présentés dans ce manuscrit sont issues de ce portail.

JGI MycoCosm *(Grigoriev et al. 2014)*

Fondé en 1997, le Joint Genome Institute (JGI) est un institut de recherche américain et un centre de séquençage spécialisé dans la génomique. Créé initialement pour contribuer au projet du Génome Humain, le **JGI** réunit des experts en cartographie du génome, en séquençage de l'ADN, en développement technologique et en sciences de l'information. Aujourd'hui, il met à disposition du public une vaste collection de données génomiques couvrant un large éventail d'organismes tels que des micro-organismes, des plantes et des communautés microbiennes environnementales.

Dans ce contexte, le portail **MycoCosm** a été développé dans le cadre du programme scientifique dédié aux champignons et aux algues, avec l'objectif de mettre à disposition une gamme complète de données génomiques relatives au règne des champignons. **MycoCosm** contribue ainsi de manière significative à l'élucidation de l'arbre de vie fongique et héberge des ressources variées tels que, des séquences d'ADN, des annotations géniques, des informations sur les protéines, des outils dédiés à l'analyse et à la visualisation, ainsi que des données relatives aux voies métaboliques et aux régulations géniques spécifiques aux champignons. En outre, cette plateforme soutient activement l'intégration, l'analyse et la diffusion des séquences génomiques fongiques et d'autres données "omiques" en mettant à disposition des outils interactifs. Ainsi, **MycoCosm** permet aux utilisateurs d'explorer les génomes fongiques séquencés, de réaliser des analyses comparatives et encourage la participation de la communauté en proposant de nouvelles espèces à séquencer, à annoter et à analyser.

MetaCyc *(Caspi et al. 2019)*

Développée dans les années 1990, **MetaCyc** est une base de données non-redondantes de voies métaboliques multi-organismes qui fait partie de la collection **BioCyc**, se distinguant par son contenu exclusivement fondé sur des études expérimentales. Contrairement aux bases de données **BioCyc** spécifiques à un organisme (*i.e.* PGDBs) qui combinent des voies prédictives déterminées *in silico* et confirmées expérimentalement pour des espèces uniques, **MetaCyc** propose une collection exhaustive de voies métaboliques élucidées expérimentalement issues de nombreux organismes. Elle inclut également des informations sur les enzymes, réactions biochimiques, composés chimiques et gènes associés. Utilisée comme ressource de référence pour la prédiction de voies, **MetaCyc** constitue un socle pour diverses applications en biologie des systèmes, biochimie, ingénierie métabolique et modélisation métabolique.

Les données dans **MetaCyc** sont organisées pour permettre un accès intuitif aux informations (*e.g.* SmartTables) sur les enzymes, les métabolites et les voies biologiques. Chaque voie est annotée manuellement avec des détails tirés de la littérature scientifique. **MetaCyc** offre une perspective large et qualifiée sur le métabolisme des organismes, sans chercher à modéliser le métabolisme complet d'un seul organisme, rôle dévolu aux PGDBs de BioCyc. Elle est principalement utilisée pour : la prédiction de voies métaboliques dans de nouveaux génomes, la recherche d'enzymes et de réactions pour des projets d'ingénierie métabolique et l'analyse comparative et fonctionnelle des voies métaboliques. Elle fait office de référence encyclopédique en constante évolution sur les voies et les enzymes (**Figure 2-3**).

Les PGDBs de BioCyc se classent en trois niveaux, Tiers 1, Tiers 2, et Tiers 3, en fonction de leur niveau de curation manuelle, de précision, et d'exhaustivité. Les PGDB de Tiers 1 et 2 sont privilégiés pour des études nécessitant une précision élevée, tandis que les PGDB de Tiers 3 conviennent mieux aux analyses exploratoires et à la comparaison préliminaire des génomes. PchCyc est un PGDB de Tiers 2.

- **Tiers 1** : PGDBs qui bénéficient d'une curation complète et manuelle, intégrant des données métaboliques et génomiques validées avec précision. Les bases de données EcoCyc et **MetaCyc** sont des exemples de PGDB de Tiers 1, contenant des informations fiables et détaillées sur les voies métaboliques, les enzymes et les gènes.
- **Tiers 2** : PGDBs à la curation partielle, intégrant certaines données vérifiées manuellement. Elles incluent également des prédictions automatisées, mais offrent un niveau de fiabilité moindre par rapport au Tiers 1.
- **Tiers 3** : PGDBs générés automatiquement sans intervention de curation manuelle, sont souvent utilisés pour explorer des organismes moins étudiés. Elles contiennent des prédictions brutes, qui nécessitent une validation expérimentale supplémentaire pour une utilisation approfondie en recherche.

Pour faciliter l'accès aux données, **MetaCyc** propose des liens d'unification et de relation avec d'autres bases bio-informatiques (*e.g.* NCBI, UniProt, KEGG). Les liens d'unification relient directement les objets **MetaCyc** à leurs équivalents dans d'autres bases, tandis que les liens de relation relient des objets apparentés, compliquant alors parfois l'interopérabilité inter-bases.

Enfin, au cours de nos travaux, nous avons eu régulièrement recours à l'utilisation de SmartTables. Les SmartTables sont des collections d'objets, tels que des gènes, des réactions, des composés qui peuvent être affichés sous forme de tableaux interactifs avec l'ensemble de leurs données associé. Les SmartTables sont particulièrement utiles dans les études comparatives et dans l'analyse des réseaux métaboliques, en permettant d'organiser, d'annoter et de visualiser de grandes quantités de données de manière simplifiée. Elles permettent d'analyser et de manipuler des données de manière flexible en facilitant l'extraction, l'analyse comparative et la visualisation de données.

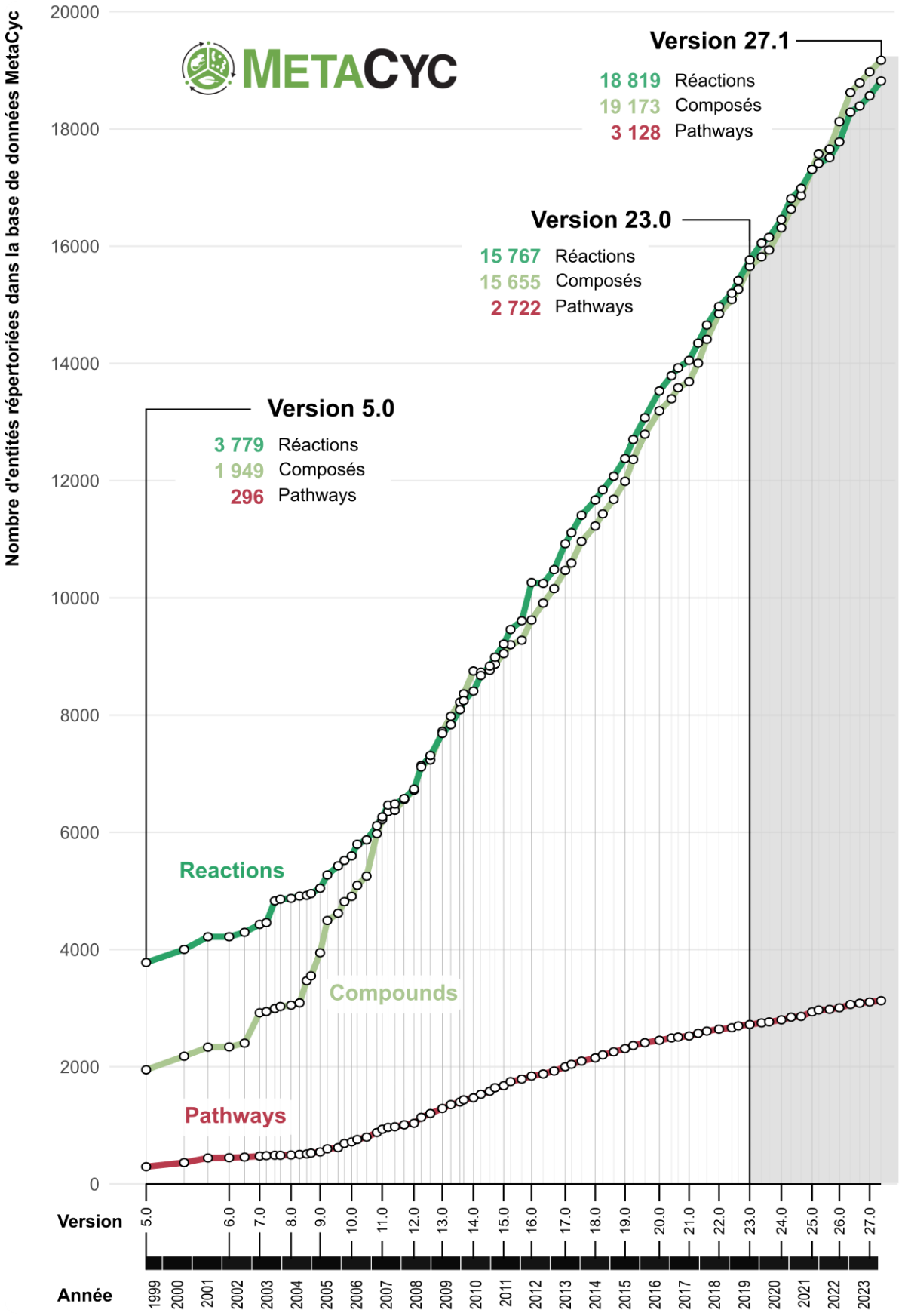
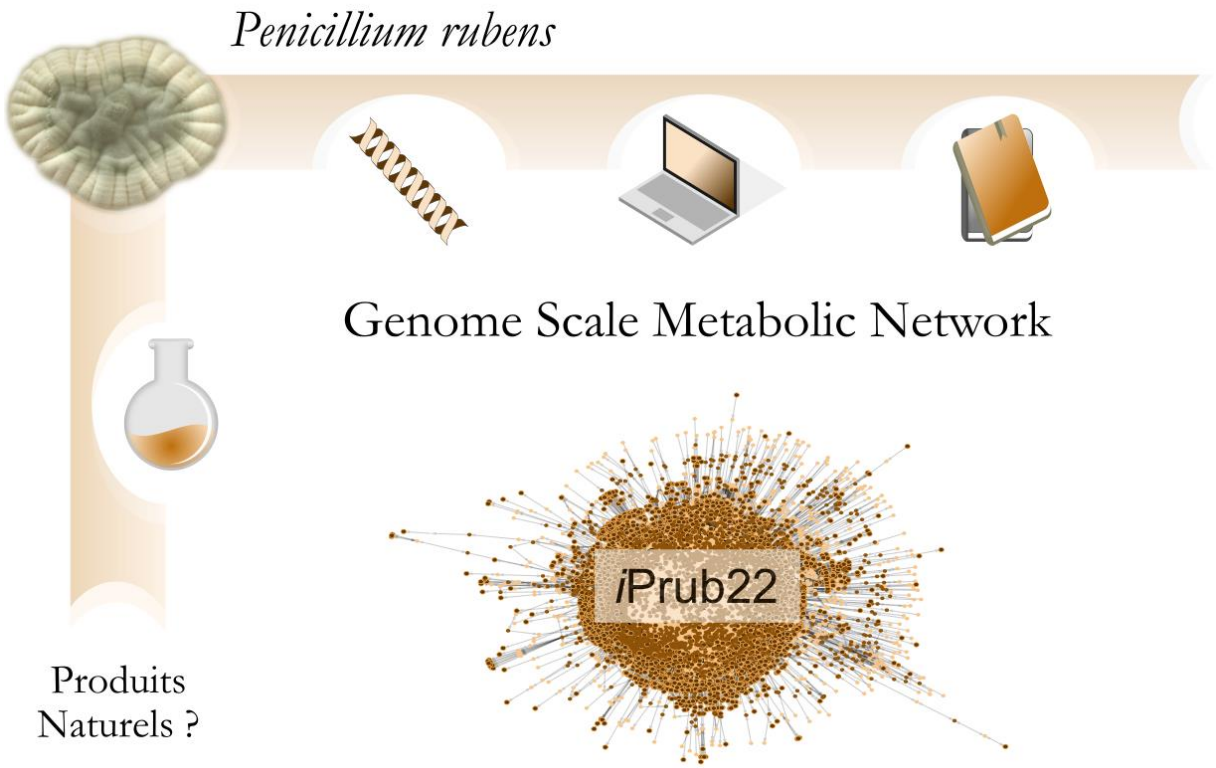


Figure 2-3: Évolution de la base de données MetaCyc à travers ses versions. *Afin de visualiser et quantifier le développement de cette base de données au cours des 24 années écoulées, ce graphique linéaire représente le nombre de réactions, de métabolites et de voies de biosynthèse répertoriées au fil des versions de MetaCyc. Les données ont été recueillies à partir de l'historique des versions disponibles sur le site internet de MetaCyc.*

Entre la version 5.0 publiée le 1er juin 1999 et la dernière version répertoriée, la 27.1 enregistrée le 28 août 2023, il est fait référence de 72 autres versions. Chacune d'entre-elles est représentée sur l'axe des abscisses où seul les numéros de version principaux sont affichés. À l'origine de MetaCyc, le nombre de réactions recensées était presque deux fois supérieur à celui des composés. À partir de la version 11.0, les nombres de réactions et de composés sont devenus sensiblement similaires, témoignant d'un effort intense et rapide dans l'enrichissement des composés. Une tendance nouvelle semble apparaître depuis la version 25.1 car désormais le nombre de composés est légèrement supérieur à celui des réactions. Enfin, nous constatons également que la vitesse dans l'acquisition et l'intégration des données semble relativement constante depuis 2008. Le ***GSMN iPrub22*** *a été reconstruit et amélioré en s'appuyant sur la version 23.0, sortie courant 2019 (marquée par le trait vertical noir). Depuis lors, MetaCyc s'est enrichi de plus de 3000 réactions et 3500 métabolites, comme indiqué dans la section grisée du graphique.*

CHAPITRE 3 :

Reconstruction et Réconciliation d'un Réseau Métabolique à l'Échelle du Génome de Haute Qualité pour *Penicillium rubens* Wisconsin 54-1255



1 - Génération d'une nouvelle reconstruction, *i*Prub22

Pour réaliser des approches **OSMAC *in silico***, en raison de l'évolution constante et croissante des connaissances sur les **GSMNs**, tant sur la forme (*e.g.* évolution des conventions d'écriture, interopérabilité) que sur le fond (*e.g.* intégration et réflexion autour du métabolisme spécialisé ainsi que sur le métabolisme basal), la question d'une nouvelle reconstruction s'est imposée. L'article suivant intitulé *Reconciliation and evolution of Penicillium rubens genome-scale metabolic networks - What about specialised metabolism?*, publié dans le journal PLOS ONE en 2023 (https://doi.org/10.1371/journal.pone.0289757) présente la **reconstruction paramétrée *i*Prub22**. Afin de maintenir la cohérence du manuscrit, une renumérotation des figures et des références a été effectuée. Les données additionnelles discutées ultérieurement dans ce document sont référencées conjointement dans la section *Supporting Information* et lors de leur occurrence dans l'article.

Une question de vocabulaire

Les termes « modèle » et « reconstruction », bien que généralement interchangeables, revêtent toutefois de légères nuances pouvant parfois porter à confusion. Fondamentalement, les enjeux et les questions portés par chacun de ces objets sont sensiblement différents. En conséquence, chacune de ces entités a pu connaître des étapes d'ajustement, peu ou prou détaillées, dont il est impératif d'avoir conscience.

- Un **modèle** (au sens générique)

Le terme « modèle », au sens générique, désigne une représentation simplifiée d'un système complexe. Il sert à décrire des propriétés communes à certains objets et vise à imiter l'objet lui-même. Dans le domaine de la biologie, les modèles sont construits à partir de données expérimentales, de connaissances préalables et de principes fondamentaux, afin de fournir une compréhension plus approfondie du système biologique étudié. Les abstractions qui le composent ont pour objectif de simplifier la complexité biologique afin de permettre des analyses plus accessibles. La validité et l'utilité des modèles dépendent de leur capacité à reproduire le comportement observé dans la réalité, ainsi que de leurs aptitudes à fournir des informations prédictives ou explicatives. Ainsi, et en accord avec cette définition, une reconstruction et un modèle de flux sont tous deux considérés comme des modèles.

- Une **reconstruction métabolique**

Une « reconstruction » métabolique est un agencement de données qui synthétise les connaissances sur le métabolisme d'un organisme. Cette structure de données représente une base d'informations métaboliques dépourvue de contraintes spécifiques. Elle adopte un format souple et neutre, fournissant notamment des perspectives sur l'évolution et la phylogénie des voies métaboliques (*e.g. dead-ends*, associations GPR, patwhays alternatifs).

- Un **modèle** (de flux)

Dans le contexte des **GSMNs**, le terme modèle est généralement associé à un modèle de flux. Le modèle sert ainsi à caractériser le comportement d'un organisme dans des conditions spécifiques (*e.g.* simulation de la croissance sous une source particulière de nutriments ou production de composés d'intérêt). Pour atteindre cet objectif, le modèle est figé en le soumettant à une série de contraintes qui fixent des paramètres variables. Ces contraintes se matérialisent principalement par la détermination de bornes inférieures et/ou supérieures sur des réactions cibles, influant sur leur ouverture ou leur fermeture. Soulignons enfin que le modèle représente un sous-ensemble fonctionnel d'une reconstruction. Lors de son utilisation, seule une partie des informations de la reconstruction est exploitée. En effet, il est estimé qu'environ un tiers des réactions des reconstructions de réseaux métaboliques à l'échelle du génome sont activables. Cette limitation résulte fréquemment de la présence de réactions bloquées ou de cycles infaisables.

- Une **reconstruction paramétrée**

Selon les définitions précédentes, une reconstruction permet la génération d'une infinité de modèles. En conséquence, nous avons choisi de désigner ***i*Prub22** sous le nom de reconstruction paramétrée, permettant ainsi de conserver toutes les informations de la reconstruction tout en fournissant un modèle prêt à exécuter des analyses de bilan des flux.

PLOS ONE

RESEARCH ARTICLE

Reconciliation and evolution of *Penicillium rubens* genome-scale metabolic networks–What about specialised metabolism?

Delphine Nègre[1,2☯], **Abdelhalim Larhlimi**[2☯], **Samuel Bertrand**[1☯*]

1 Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, Nantes, France, **2** Nantes Université, École Centrale Nantes, CNRS, Nantes, France

☯ These authors contributed equally to this work.
* samuel.bertrand@univ-nantes.fr

Citation: Nègre D, Larhlimi A, Bertrand S (2023) Reconciliation and evolution of *Penicillium rubens* genome-scale metabolic networks–What about specialised metabolism? PLoS ONE 18(8): e0289757. https://doi.org/10.1371/journal.pone.0289757

Editor: Bhanwar Lal Puniya, University of Nebraska-Lincoln, UNITED STATES

Received: April 7, 2023

Accepted: July 24, 2023

Published: August 30, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0289757

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

1.1 ABSTRACT

In recent years, genome sequencing of filamentous fungi has revealed a high proportion of specialised metabolites with growing pharmaceutical interest. However, detecting such metabolites through *in silico* genome analysis does not necessarily guarantee their expression under laboratory conditions. However, one plausible strategy for enabling their production lies in modifying the growth conditions. Devising a comprehensive experimental design testing in different culture environments is time-consuming and expensive. Therefore, using *in silico* modelling as a preliminary step, such as Genome-Scale Metabolic Network (**GSMN**), represents a promising approach to predicting and understanding the observed specialised metabolite production in a given organism. To address these questions, we reconstructed a new high-quality **GSMN** for the *Penicillium rubens* Wisconsin 54-1255 strain, a commonly used model organism. Our reconstruction, ***i*Prub22**, adheres to current convention standards and quality criteria, incorporating updated functional annotations, orthology searches with different **GSMN** templates, data from previous reconstructions, and manual curation steps targeting primary and specialised metabolites. With a MEMOTE score of 74% and a metabolic coverage of 45%, ***i*Prub22** includes 5,192 unique metabolites interconnected by 5,919 reactions, of which 5,033 are supported by at least one genomic sequence. Of the metabolites present in ***i*Prub22**, 13% are categorised as belonging to specialised metabolism. While our high-quality **GSMN** provides a valuable resource for investigating known phenotypes expressed in *P. rubens*, our analysis identifies bottlenecks related, in particular, to the definition of what is a specialised metabolite, which requires consensus within the scientific community. It also points out the necessity of accessible, standardised and exhaustive databases of specialised metabolites. These questions must be addressed to fully unlock the potential of natural product production in *P. rubens* and other filamentous fungi. Our work represents a foundational step towards the objective of rationalising the production of natural products through **GSMN** modelling.

1.2 INTRODUCTION

Comparable to Alexander Fleming's pioneering discovery of penicillin in 1929, the elucidation of natural products (NPs) has frequently been serendipitously achieved through the random screening of plant, fungal, and bacterial extracts. Natural molecules, synthesised by living organisms, can be categorised into two types based on their function within them: constitutive metabolites, which are essential for the organism's growth, and specialised metabolites, which provide competitive advantages towards the organism's environment, historically referred to as secondary metabolites. The latter is of great interest in pharmacology, both in the field of health and agriculture, since many drugs are originated from these biologically active natural molecules *(Newman et Cragg 2020)*. However, since the early 2000s, the discovery rate of new biologically active NPs has slowed down or even become scarce. Moreover, the time-consuming, inefficient and costly traditional methods used for bioprospecting new NPs have been slowly abandoned *(David et al. 2015)*. Given the increasing antibiotic resistance *(Gould 2009)*, it is essential to reinvigorate research on NPs, which has faltered in recent years, using more rational and predictive approaches. The confluence of genomics and metabolomics has now enabled the correlation of metabolites with their structural data to their respective biosynthetic gene clusters (BGCs). Combined with the automation and qualitative evolution of tools observed in recent years *(Wolfender et al. 2019)*, these approaches have become increasingly appealing for new bioactive NPs discovery.

Filamentous fungi are known for their ability to produce a diverse array and a substantial amount of NPs *(Wiemann et Keller 2014)*. Over the years, significant knowledge has been accumulated about their metabolism, and fungal genomes exploration has highlighted the high proportion of specialised metabolites produced by enzymes with co-located genes. In addition to biosynthetic capabilities, such BGCs could be associated with resistance, transport, or regulatory genes. Thus, their evolutionary history may provide clues to their biological activities that could guide molecular engineering research *(Eustáquio et Ziemert 2018)*.

However, despite the richness of BGCs exhibited by fungi, one of the extensive risks NP chemists face is the strong tendency to rediscover previously characterised molecules *(David et al. 2015)*. Such a trend is inconsistent with the genomic data that have been acquired, showing a large number of gene clusters associated with unknown specialised metabolism. Indeed, the differences observed between theoretical *in silico* models and *in vitro* cultures indicate that gene clusters not expressed in the laboratory are numerous and form an untapped resource for novel bioactive NPs *(Nielsen et al. 2017; Arora et al. 2020)*.

Penicillium rubens, formerly known as *Penicillium chrysogenum* *(Houbraken et al. 2011)*, is a fungus used in the industry to produce *β*-lactam antibiotics, including penicillin. Consequently, both industrial and wild terrestrial strains of *Penicillium* have been extensively studied over the years. After decades of classical strain improvements using the wild strain NRRL1951 *(Salo et al. 2015; Guzmán-Chávez et al. 2018)*, all industrial strains with high penicillin yields are now derived from the Wisconsin 54-1255 strain, which has therefore become a model species within the filamentous fungi *(Fierro et al. 2022)*. Thus, the availability of its genome *(Van den Berg et al. 2008)* makes it a worthy candidate for studying its metabolism. In 2009, 34 BGCs were identified in *P. rubens*, only 4% of which were associated with a known and isolated compound. Nine years later, 28% of the compounds from the now-identified 50 BGCs were characterised *(Guzmán-Chávez et al. 2018)*. Thus, understanding the conditions under which silent, orphan or cryptic gene clusters could be activated remains of great interest. Following the example of what has been achieved for bacterial NPs *(Zarins-Tutt et al. 2015)*, decoding the genome is the crucial first step toward understanding how to unlock such cryptic genes.

In systems biology, an organism's biochemical and physiological metabolic properties can be described by Genome-Scale Metabolic Networks (**GSMNs**) studies *(Klipp et al. 2016)*. Reconstruction of such models is based on computational methods using both *in silico* and experimental data. Indeed, integrative biology principles applied to genomics and metabolomics allow us nowadays to link the high-quality data obtained by analytical

technologies to the corresponding models. Moreover, increasing model complexity aims to integrate and visualise heterogeneous knowledge-related metabolism more and more efficiently *(Pan et Reed 2018)*, thereby enhancing our understanding of metabolism at the system level *(Hattwell et al. 2020)*. However, sustainability, evolution, and exploitation of models are intrinsically linked to their adoption by the scientific community, as illustrated by the work performed on the *Caenorhabditis elegans* metabolic network reconstruction *(Witting et al. 2018)*. Unfortunately, only a few examples of individual network reconciliation are reported *(Aminian-Dehkordi et al. 2019; Nouri et al. 2020)*, aside from the well-studied model organisms presented in a non-exhaustive list by Gu *et al.* *(Gu et al. 2019)*, which are exceptions. This phenomenon can be attributed to various factors, including the massive generation of data associated with its quality and its updating, the lack of inter- and intra-operability between platforms and software, the multiplication of reconstruction tools linked to the current facility to reconstruct **GSMNs**, and the lack of transparency in the acquisition of previous networks *(Ravikrishnan et Raman 2015; Pham et al. 2019; Community standards to facilitate development and address challenges in metabolic modeling 2020; Bernstein et al. 2021)*. Moreover, a **GSMN** represents a platform of organised and summarised resources intended to reflect the optimal metabolic capabilities of the target organism at the time of reconstruction based on the available knowledge and data at that moment (*e.g.,* knowledge snapshot). In this respect, given the constant and rapid evolution of data, the need to maintain these models up-to-date and standardised is increasingly becoming pressing.

Therefore, since the first *P. rubens* **GSMN** (*i.e., i*AL1006) was published in 2013 *(Agren et al. 2013)*, the work presented here proposes a new **GSMN** for *P. rubens* Wisconsin 54-1255, considering the points mentioned above. The present study is primarily concerned with the reconstruction process, with particular emphasis on providing informative outcomes regarding specific model generation. Additionally, particular attention was paid to specialised metabolism. Thus, the provided model could be compatible with future research on NPs pathways regulation, which could furnish insight into cryptic pathways. Thus, the ***i*Prub22** model, composed of 5,919 reactions and 5,192 unique metabolites, combines data from

previous reconstructions, an updated functional annotation, a homology search with various phylogenetically close and distant organisms, and manual curation based mainly on literature. Furthermore, our reconstruction provides the necessary framework for investigating the growth of P. rubens and exploring the production of its specialised metabolites under varying media conditions through constraint-based modelling.

1.3 RESULTS

1.3.1 Draft generation: reconstruction and reconciliation

The **GSMN** reconstruction ***i*Prub22** started with the automatic generation of two drafts generated using genome functional annotation and homology searches with pre-existing models.

1.3.1.1 Functional annotation subnetwork

Usually, three annotation labels are required to provide a source of 'annotation-based' metabolic reactions. Thus, Enzyme Commission (EC) numbers, Gene ontology terms (GOT) *(Ashburner et al. 2000; The Gene Ontology Consortium 2021)*, and protein domains (PFAM) were sought among the *P. rubens* sequences. Of the 12,556 gene sequences that constitute the *P. rubens* genome *(Van den Berg et al. 2008)* 8,948 (71%) possess at least one annotation (*i.e.,* EC number, GOT and/or PFAM), and 2,270 (25%) are annotated by all three annotations (**Figure 1*** in S1 file). The EC numbers are associated with genomic sequences based on KEGG orthology (KO) identifiers *(Kanehisa et al. 2021)*. Thus, the combined use of KAAS *(Moriya et al. 2007)* and gene information stored in KEGG leads to an 11% enrichment in EC numbers compared to relying solely on Trinotate *(Bryant et al. 2017)* results (**Figure 1*** in S1 file, Table 1 in S1 file). The specificity and complementarity of these three annotation sources are illustrated in **Figure 2.A**† in S1 file. As a result, these annotations allowed the generation of the functional annotation subnetwork composed of 3,390 genes (*i.e.,* 27% of the genome), 1,359 pathways, 3,205 enzymatic reactions and 3,577 metabolites.

* Annexe 1A, page 426

† Annexe 1B, page 429

1.3.1.2 Orthology subnetwork

In parallel, following the search for homology between species, a second subnetwork was reconstructed. Orthologous genes between *P. rubens* and seven other reconstruction templates were searched to achieve the creation of this subnetwork. Orthology templates were chosen according to the quality of their **GSMNs**, their phylogenetic proximity to *P. rubens*, or their reconstruction procedure (see the selected network's topology in **Figure 3*** in the S1 file).

Only the genomic sequences initially present in the **GSMNs** are queried using OrthoFinder *(Emms et Kelly 2019)*. Respectively, the templates *Arabidopsis thaliana* *(de Oliveira Dal'Molin et al. 2010)*, *Aspergillus nidulans* *(Pitkänen et al. 2014; Castillo et al. 2016)*, *Aspergillus niger* *(Pitkänen et al. 2014; Castillo et al. 2016)*, *Neurospora crassa* *(Dreyfuss et al. 2013)*, *Penicillium chrysogenum* species complex *(Pitkänen et al. 2014; Castillo et al. 2016)*, *Saccharina japonica* *(Nègre et al. 2019)* and *Schizosaccharomyces pombe* *(Pitkänen et al. 2014; Castillo et al. 2016)* are composed of 2,330, 1,279, 1,299, 836, 1,352, 5,016 and 810 sequences of which only 0.04%, 1.9%, 0.15%, 1.7%, 11%, 0.04% and, 0.49% were not found in the constitution of the queried proteomes. The 8,519 homologous sequences between *P. rubens* and the different templates *A. thaliana*, *A. nidulans*, *A. niger*, *N. crassa*, *P. rubens* species complex, *S. japonica*, and *S. pombe* are distributed according to 478, 942, 866, 635, 1,003, 1,273 and, 667 orthogroups respectively (Table 2 in S1 file and S2 file). Sequence distribution among the different species templates and orthogroups is displayed in **Figure 4** in the S1 file. Of these homologous sequences, 2,903 are paralogs and will not be considered for reconstruction. Respectively, there are 1,494, 1,144, 1,049, 766, 1,174, 2,302, and 764 orthologs between the templates studied (*i.e., A. thaliana*, *A. nidulans*, *A. niger*, *N. crassa*, *P. rubens* species complex, *S. japonica*, and *S. pombe*) and *P. rubens*. Complementarily, the number of *P. rubens* genes being orthologous to the different templates are 866, 1,818, 1,620, 1,104, 2,368, 2,329, and 1,114, respectively.

* Annexe 2B, page 436

Orthologous genes to the template ones were used to obtain their corresponding reactions related to *P. rubens* Wisconsin 54-1255. As the selected template networks were mostly reconstructed between 2010 and 2018 *via* the KEGG database *(Kanehisa et al. 2021)*, a mapping step on the MetaCyc database *(Nègre et al. 2019)* was necessary for the interoperability between the OrthoFinder results and the subnetworks inference. The draft generation indicates that less than 50% of the detected and potentially informative reactions were lost during this mapping (**Figure 2.B*** in the S1 file). Mapped reactions correspond to 46% of the total reactions for *A. thaliana*, 49% for *A. nidulans*, 49% for *A. niger*, 49% for *P. rubens* species complex, 49% for *S. pombe* and 46% for *N. crassa*. The best mapping rate (87%) is obtained with the *S. japonica* network, whose **GSMN** was reconstructed following a similar protocol to the one presented here and directly *via* the MetaCyc database *(Nègre et al. 2019)*. In terms of gene number, these losses represent 18% of the total number of genes detected by OrthoFinder (*i.e.,* 5,616 genes). However, 70% of the genes (*i.e.,* 716 genes) were finally included in the **GSMN** as they possess correspondence *via* the annotation or with external sources (**Figure 5**† in S1 file). The remaining unmapped 307 sequences were lost and represent a set to be preferentially explored for future network improvements. Thus, the orthology subnetwork is composed of 4,615 genes, 1,550 pathways, 2,946 enzymatic reactions and 3,238 metabolites.

1.3.1.3 Enrichment with external sources

Penicillium rubens is a widely studied model organism, but the available data heterogeneity makes conciliation a delicate task. Therefore, the choice to add external data was made on consistent sources, namely PchCyc, the Pathway/Genome Database (PGDB) from BioCyc *(Karp et al. 2019)*, and the latest version of the organism's network, called Prubens *(Prigent et al. 2018)*. This reconstruction, resulting from a large-scale automatic reconstruction process, takes advantage of a mapping on MetaCyc from the original version of the *i*AL1006 network *(Agren et al. 2013)*. Thus, data were added when reactions, or in default, all their metabolites, had compatible MetaCyc identifiers and were supported by at least one Gene-Protein-Reaction

*Annexe 2C, page 438 † Annexe 2E, page 441

(GPR) association. Respectively, Prubens and PchCyc are composed of 2,533 and 1,291 reactions, of which 650 are shared. Of the 3,824 total reactions queried, 2,418 had a compatible MetaCyc identifier, and 2,112 were associated with at least one gene. Furthermore, of the 1,725 reactions already present in the draft, 313 had identical GPR associations, 123 had no genomic information in the external sources (*e.g.,* in the ***i*Prub22** draft, these reactions are henceforth annotated between 1 and 39 genes), and 1,289 possessed different GPR associations. Finally, the draft was completed by 510 reactions (*i.e.,* 440 reactions belonging exclusively to Prubens, 43 reactions belonging exclusively to PchCyc and 27 reactions coming from both sources). Moreover, 263 supplementary reactions from *i*AL1006 were also included since all the reactants/products of those reactions possess a MetaCyc identifier. To ensure data traceability, the original identifiers of these reactions were kept, and the selection is shown in **Figure 6*** in the S1 file.

1.3.1.4 The initial draft

All the data obtained were then merged to produce a first version of the network (*i.e.,* draft). The complementarity of each approach, displayed in **Figure 3-1** and Supporting Information (**Figures 2.C**, **7**† and **8**‡ in the S1 file), makes it possible to question the meaning and relevance of the added data. These figures trace the origin of each reaction and, therefore, of the genes and metabolites, within the draft. Thus, it is easier to disentangle the information from biological reality or computational bias. For instance, in the orthologous subnetwork, 17 and 1,237 reactions (**Figure 5.C**§ in the S1 file) were exclusively from the most phylogenetically distant templates (*i.e., A. thaliana* and *S. japonica*, respectively). In the case of the 17 *A. thaliana* reactions, 8 (47%) of them are also supported by the annotation source, and for *S. japonica*, the number amounts to 967 (78%) reactions. Similarly, the addition of external sources (*i.e., i*AL1006, Prubens and PchCyc) supported 773 reactions (17% of the draft reactions) which were not retrieved with the annotation and orthology subnetworks. Therefore, reactions supported only by one source should be given special attention in the future. Moreover, according to the nomenclature of EC numbers, reactions are divided into the

* Figure 3-10, page 205 † Annexe 3C, page 449 ‡ Annexe 1D, page 431 - Annexe 2F, page 442 - Annexe 3B, page 447 § Annexe 2D, page 440

following eight categories: oxidoreduction 28%, transfers reactions 24%, hydrolyses 18%, lyses 7.9%, isomerisation 2.8%, reactions involving ligases 3.4%, translocation 0.38% and undetermined reaction type 16%. Nevertheless, this distribution is different within the genes since the most frequent classes are: transferases (29%), hydrolases (26%) and oxidoreductases (24%). However, it should be noted that only 42% of the genes have a clearly identified EC number, predominantly as a functional annotation result (**Figure 9** in the S1 file).Furthermore, from the 4,959 metabolites present at this reconstruction stage, 1,085 (22%) come exclusively from the annotation's subnetwork, 772 (16%) from the orthology subnetwork and 565 (11%) from external sources (**Figure 3-1.C**). Consequently, 2,537 (51%) metabolites are shared between at least two different sources (**Figure 10** in the S1 file). According to the MetaCyc compounds' ontologies, 349 (*i.e.,* 7% of draft metabolites) belong to the specialised metabolism (*i.e.,* 41 (12%) exclusively from the annotation, 92 (26%) exclusively from the orthology, 153 (44%) exclusively from external sources and the remaining 63 (18%) are supported by at least two of these sources). MetaCyc v23.0 contains approximately 12% of compounds annotated "secondary".*In fine*, the reconstructed draft includes 5,585 genes which support 4,916 reactions included in 1,817 pathways and involving 4,959 metabolites.

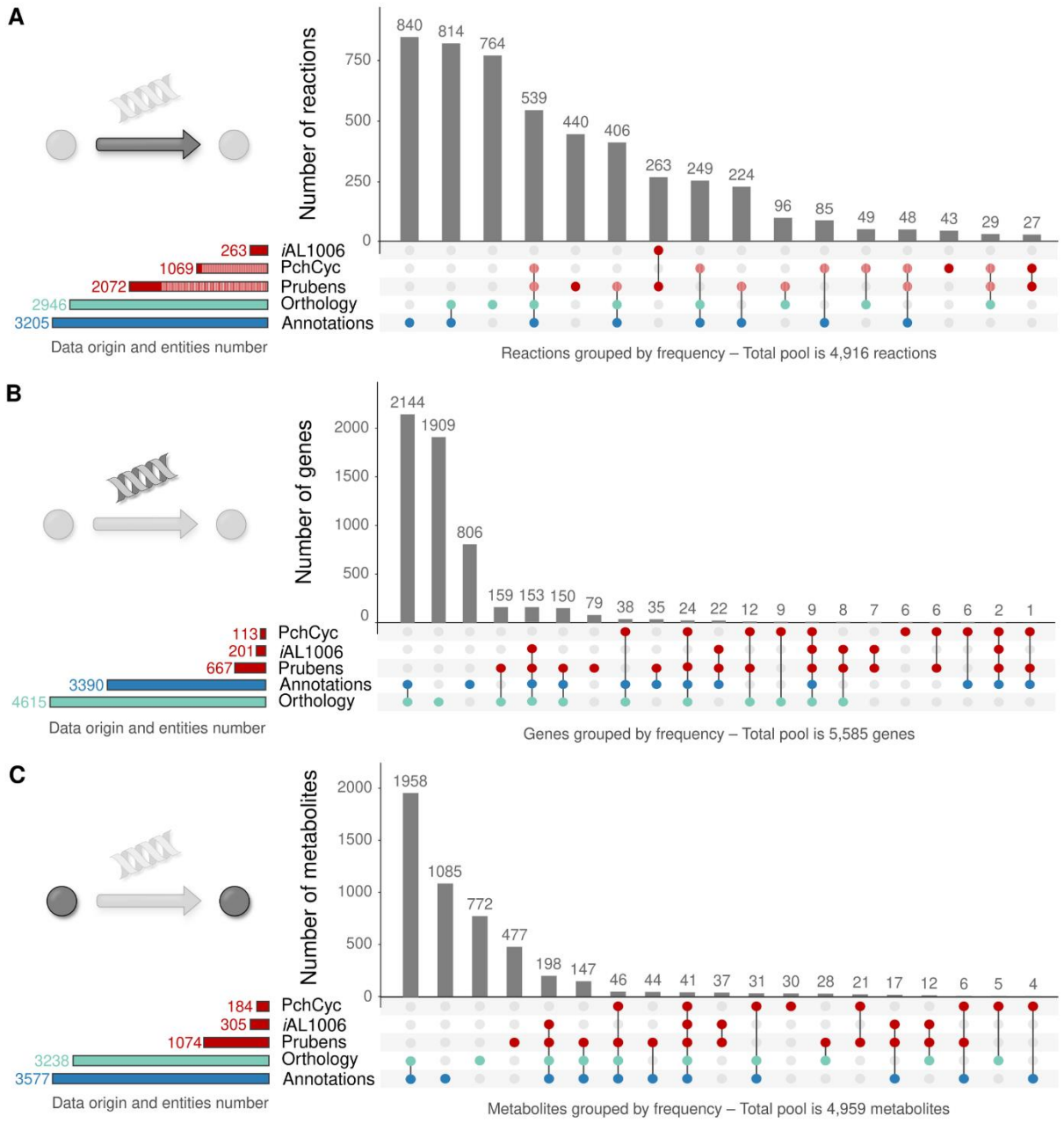


Figure 3-1: UpSet diagram representing the origin of each of the reactions (A), genes (B), and metabolites (C) added to the *Penicillium rubens* Wisconsin 54-1255 draft. This traceability underlines the complementarity of the approaches. Information comes from annotation (■) orthology (■) subnetworks and external sources (■). The red hatched area (■) corresponds to reactions shared between the external sources and the draft from the annotation and/or orthology subnetworks. For a further view and details, see **Figures 7*** and **8**† in S1 File.

* Annexe 3C, page 449

† Annexe 1D, page 431- Annexe 2F, page 442 - Annexe 3B, page 447

1.3.1.5 Reconstruction improvements

Once the initial draft is produced, several modifications must be performed to obtain a functional and searchable model compatible with the experimental observations (*i.e.,* a knowledge platform comprising a set of accessible resources for a given organism at the instance of reconstruction). The first step is to establish a list of metabolites known to be produced by the studied organism. Initially, we investigated the presence of these target compounds in the draft and subsequently explored their topological producibility. When these compounds are either absent or not topologically producible, gap-filling steps should be performed to improve network connectivity. Therefore, the following section deals with the necessary prerequisites for setting up a high-quality **GSMN** ready for subsequent flux analyses.

1.3.1.6 Target selection for model enhancement

As it is not possible to bring the same confidence weight to all available data, the targets selected to study and improve the metabolic network of *P. rubens* were grouped into three distinct sets of targets (*i.e.,* set 'Targets1' with high confidence, set 'Targets2' with medium confidence and set 'Targets3' with low confidence – S3 file for further details).

The first target set corresponds to information from the literature. The target metabolites belong indifferently to specialised or constitutive metabolism in that particular case. Two publications, in particular, caught our attention, one allowing the accurate modelling of the sugar pathway in fungi *(Aguilar-Pontes et al. 2018)* and the other resulting from the penicillin biosynthesis pathway manual curation *(Praußе et al. 2016)*. This list was completed by metabolites related to the biomass function *(Agren et al. 2013)*. In addition, the SMASH tool suite *(Blin et al. 2019)* was used to highlight 19 specialised metabolite pathways through gene cluster search. Although, only the biosynthetic pathways of penicillins, roquefortines and patulin can be exploited due to the lack of MetaCyc identifiers for compounds of the other pathways. However, knowing that in *P. rubens*, the patulin biosynthetic pathway is incomplete *(Nielsen et al. 2017)*, this compound is discarded from 'Targets1', and its pathway curation

was later performed in this sense. Consequently, 11 metabolites were supported by genome mining sources in the first list of targets to be monitored. During this process search, a list of 46 compounds lacking MetaCyc identifiers was also generated (*i.e.,* orphan metabolites). Moreover, even if Melanin and PR-Toxin are not strictly speaking orphan metabolites since they are referenced on MetaCyc, they cannot be produced topologically since no reactions were associated with them. At that stage, those latter compounds cannot be monitored within the network unless focusing on their closest referenced precursor. Finally, set 'Targets1' is thus composed of 243 targets.

The second list is obtained by querying specific NP databases. Using a taxonomy-oriented query from the naturaL prOducTs occUrrence databaSe (LOTUS) *(Rutz et al. 2022)*, 240 compounds for *P. chrysogenum*/*P. rubens* are retrieved along with their InChIKeys, which are used to search within the MetaCyc database. Based on an exact InChIKey match, followed by an analogue of the first two sections of the identifier and finally on the first part (*i.e.,* corresponding to planar chemical structure search), only 8, 11 and 47 metabolites were found, respectively. Depending on these criteria, between 4, 7 or 13 of these metabolites were already present in the draft. Therefore, in the best case, only 20% of the information in LOTUS can be included in the reconstruction. The assignment verification to the studied strain allows the selection of 35 metabolites which constitutes the set 'Targets2'.

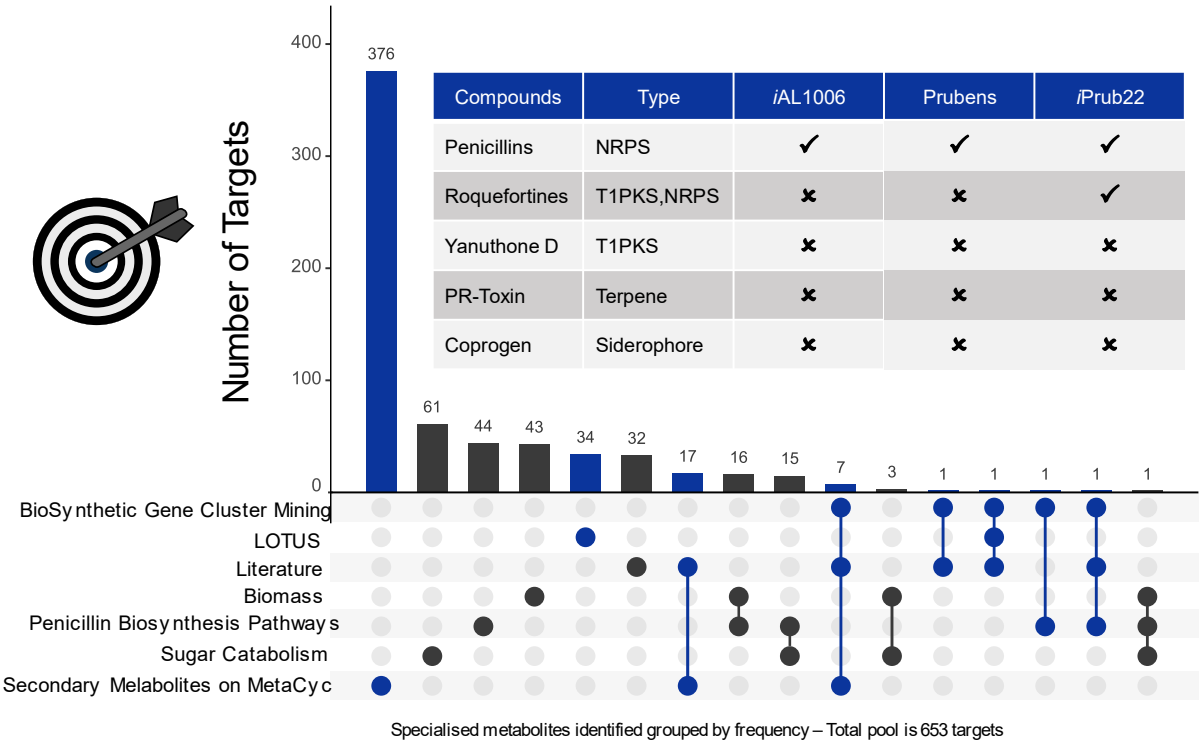
Finally, the third set of targets was established based on the metabolites present in the reconstruction (*i.e.,* at the step intermediate C – "**Tableau 3-1**") and that have annotation with the term 'Secondary-Metabolites' in the MetaCyc compounds ontologies when they were searched with SmartTables. At that stage, the reconstruction contains 400 metabolites classified as "secondary". They are divided into 149 compounds that are at least produced and consumed by a reaction, 90 that are only consumed and 161 that are only produced (*i.e.,* 251 topological dead-ends). Thus, Set 'Targets3' was composed of 400 targets.

Tableau 3-1: Potential topological producibility expressed in the different curation steps of the reconstruction.

Number of entities	Reconstruction progress						
	Draft	A	B	C	D	E	GSMN
Reactions	4,916	5,389	5,475	5,536	5,551	5,823	5,919
Active reactions	-	2,685 (50%)	2,775 (51%)	2,953 (53%)	2,970 (54%)	3,527 (61%)	3,707 (63%)
Metabolites	5,060	5,245	5,292	5,309	5,318	5,437	5,464
Producible metabolites	-	1,700 (32%)	1,760 (33%)	1,887 (36%)	1,899 (36%)	2,406 (44%)	2,545 (47%)
Metabolites only consumed	1,564 (31%)	1,527 (29%)	1,507 (28%)	1,484 (28%)	1,48 (28%)	1,366 (25%)	1,333 (24%)
Metabolites only produced	1,691 (33%)	1,736 (33%)	1,726 (33%)	1,721 (32%)	1,725 (32%)	1,661 (31%)	1,641 (30%)
Metabolites consumed and produced	1,805 (36%)	1,982 (38%)	2,059 (39%)	2,104 (40%)	2,108 (40%)	2,410 (44%)	2,490 (46%)
Producible metabolites (targets 1)	-	193/235 (81%)	198/235 (84%)	226/235 (95%)	226/235 (95%)	226/235 (95%)	234/243 (96%) ▲
Producible metabolites (targets 2)	-	5/35 (14%)	6/35 (17%)	6/35 (17%)	14/35 (40%)	14/35 (40%)	14/35 (40%)
Producible metabolite (targets 3)	-	87/400 (22%)	95/400 (24%)	147/400 (37%)	147/400 (37%)	368/400 (92%)	368/400 (92%)

The main objective of gap-filling was to ensure that the known metabolites in *Penicillium rubens* were present before using the reconstruction as a model for analysis. As gap-filling is an iterative process, it was linearised, and results were grouped for clarity's sake. Thus, between the draft and the **GSMN**, intermediates A, B, C, D and E were generated. First of all, to test the metabolites' producibility, 208 transport, 37 demand, 1 sink and 217 exchange reactions were incorporated (draft to A), followed by 86 spontaneous reactions (A to B) and then 61 (B to C), 15 (C to D), 272 (D to E) and 75 (E to **GSMN**) reactions from the four gap-filling runs. Minor corrections (not described) were made between E and the final reconstruction. The topology analyses were carried out with the seed list defined from the metabolites present in the extracellular medium. The eight metabolites specific to the biomass reaction were only monitored from their inclusion in the reconstruction at the **GSMN** step (▲).

All together, 653 metabolites were used to obtain a high-quality **GSMN** by seeking the network for their presence and connectivity. Based on the MetaCyc annotations, BGCs mining and LOTUS database, this list is divided into 215 and 438 compounds belonging to constitutive and specialised metabolism, respectively (**Figure 3-2**).



Compounds	Type	*i*AL1006	Prubens	*i*Prub22
Penicillins	NRPS	✓	✓	✓
Roquefortines	T1PKS,NRPS	✗	✗	✓
Yanuthone D	T1PKS	✗	✗	✗
PR-Toxin	Terpene	✗	✗	✗
Coprogen	Siderophore	✗	✗	✗

Figure 3-2: Origin of the 653 targets used for reconstruction enhancement. Example of secondary metabolites produced by *Penicillium rubens* according to the main natural product classes: NRPS (Non-Ribosomal Peptides Synthetase and siderophores), T1PKS (Type I Polyketide synthase), hybrid form, and terpenes. The selection was based on data from the literature (*i.e.* Biomass *(Agren et al. 2013)*, Sugar Catabolism *(Aguilar-Pontes et al. 2018)*, Penicillin Biosynthesis Pathways *(Praußе et al. 2016)* and literature details on Supporting Information), genome mining *(Blin et al. 2019)* and LOTUS databases *(Rutz et al. 2022)*. All targets are divided according to constitutive (■ – 215) and specialised (■ – 438) metabolism. Metabolites found (✓) or not (×) in the different versions of **GSMNs** are shown. Metabolites without MetaCyc-compatible information cannot be included in the networks (*i.e.* absence of compound identifier as for yanuthone D or the absence of reactions leading to compound production/degradation as for coprogen or PR-Toxin).

1.3.1.7 Modelling exchanges with the environment

To check the producibility of a compound by topological analysis and then by constraint analysis, it is necessary to define a set of initiation metabolites called seeds. Compounds selection from the extracellular medium led us to review and modify the reactions related to the metabolite transport phenomena between the various model compartments.

Although the proposed model does not include intracellular compartmentation, initially, 76 reactions, supported by a total of 204 genes, involve transport mechanisms. The location of their gene products corresponds to the cell membrane for 107 of them. The 22 reactions that do not possess one of these genes in their GPR association were either corrected or blocked in the model (*i.e.,* reactions that biologically model an intracellular transfer whereas it is extracellular in the reconstruction). The performed curation allowed adding 208 metabolite transfer reactions between the extracellular and intracellular compartments. At least one gene, for a total of 220, is associated with 91 of them. For modelling purposes, 37 demand reactions (*i.e.,* irreversible transfer from the external to the intracellular environment) and 1 sink reaction were also included.

Then, exchange reactions corresponding to 185 uptake reactions and 42 production reactions were added (**Figure 11*** in the S1 file and S4 file). Such exchange reactions are related to only the transfer of the same metabolites between the extracellular compartment and the environment (*i.e.,* system boundaries). These reactions are entirely artificial. However, their addition will allow the simulation of the nutritional conditions of the organism (*i.e.,* for further use of the model, *e.g.,* tests of functionalities and production of metabolites of interest). Thus, the seeds list used for the topological analyses corresponds to the metabolites present in system boundaries (*i.e.,* having an uptake reaction).

Finally, the reconstruction is composed of 284 (4.8%) transport reactions, 37 (0.63%) demand reactions, 1 sink reaction, 185 (3.1%) uptake reactions and 42 (0.71%) production reactions. These reactions, therefore, represent 9.3% of the **GSMN** reactions.

1.3.1.8 Gap-filling and topological producibility

Once the targets and seeds are defined, all that remains to be performed is to select available reactions for gap-filling querying. The database used for gap-filling (MetaCyc v23.0) contains 17,201 reactions. As gap-filling is a heuristic approach, these reactions are filtered into two categories to select the more relevant set of reactions to query. The filtering presented below reduces the space to be interrogated by 46%. Consequently, Gap-filling was run firstly with MetaCyc subset reactions (*i.e.,* 9,281 reactions) with a high confidence weight.

The first category corresponds to reactions related to the fungi kingdom. This list was constructed based on a meta-network obtained by the reconstructions of a species' repertoires belonging to the fungi kingdom *(Belcour et al. 2022)*. This meta-network is composed of 3,810 reactions. Interestingly, 77% of the reactions (*i.e.,* 2,964 reactions) were already present in the draft. This list of reactions was completed by the BioCyc PGDBs of *Candida albicans*, *Exophiala dermatitidis* and *Saccharomyces cerevisiae*, composed of 1,320, 1,714 and 1,562 reactions, respectively. Combining these three latter networks yield a set of 2,905 reactions, among which 70% (*i.e.,* 2,041 reactions) were already present in the draft. The fusion of these two lists of reactions provided a pool of 4,709 reactions, of which 64% (*i.e.,* 3,036 reactions) were already included in the draft reconstruction. Finally, this first category comprises 1,043 reactions of interest.

At this point in the reconstruction, the draft network contains only reactions supported by at least one gene and thus is composed only of reactions catalysed by enzymatic proteins. For instance, biological catalysis carried out by RNA molecules (*e.g.,* intra- or inter-molecular catalysis) or spontaneous reactions are not initially present in the reconstruction. However, their importance in modelling biological systems or under experimental conditions is no longer in question *(Keller et al. 2015; Lerma-Ortiz et al. 2016; Golubev et al. 2017)*. Therefore, the second category includes reactions for which the genomic information of the enzyme is absent or unknown, always according to data on MetaCyc. They were gathered by filtering the reactions on the MetaCyc web server into "Spontaneous reactions for which no enzyme is

required" and "Enzyme-catalysed reactions for which no enzyme has been identified (pathway holes)". It allowed the selection of 705 and 4,971 reactions (downloaded in October 2021 and May 2022, respectively). However, it is noteworthy that 22 (3.1%) and 631 (13%) reactions belonging to these two sets, respectively, are already present in the network, thus questioning the relevance and accuracy of the annotations.

For the classification and traceability of the results, the process leading to the generation of the high-quality **GSMN** involved five intermediate steps labelled A to E. The initial step (*i.e.,* A to B) combined the draft, which incorporated all data supported by at least one gene, with exchange reactions. In this step, 32% of the metabolites (*i.e.,* 1,760 compounds) were topologically producible, and 51% of the reactions (*i.e.,* 2,775 reactions) were active under the evaluated conditions. Through subsequent manual curation, these percentages were improved at the final step to 47% (*i.e.,* 2,545 compounds) and 63% (*i.e.,* 3,707 reactions), respectively. More particularly, out of the 645 monitored metabolites, 285 (43%) were found to be producible in the first intermediate step. In the final **GSMN *i*Prub22**, this percentage reached 94%. Indeed, the refinement process, summarised in "**Tableau 3-1**" involved several gap-filling steps to enhance topological connectivity. For the first set of targets (*i.e.,* 'Targets1'), 61 reactions carefully checked were incorporated into the reconstruction (*i.e.,* B to C). Moreover, GPR associations were included for 20 of them. For the second set of targets (*i.e.,* 'Targets2'), 29 out of 35 metabolites were not producible after the first gap-filling run. Through the addition of 15 reactions from the second gap-filling (*i.e.,* C to D), 8 of the previously not producible metabolites were successfully recovered. The third set of targets (*i.e.,* 'Targets3') was used to improve network connectivity by targeting secondary metabolites initially present in the draft. The addition of 272 reactions (*i.e.,* D to E) in the third gap-filling expanded the model capabilities since 507 supplemental metabolites could henceforth be produced, and 557 more reactions were activated. A final gap-filling step was performed with all the metabolites present in the initial draft against reactions with unknown enzyme sequences. It resulted in the addition of 75 more reactions (*i.e.,* E to **GSMN**). In summary, 510 reactions were included in ***i*Prub22** through the gap-filling processes (S5 file). About half

of these reactions were sourced directly from the MetaCyc pre-filtered set (**Figure 12*** in the S1 file). Furthermore, there was an enrichment of 86 specialised metabolites, as the draft contained 405 metabolites labelled as “secondary metabolites” compared to 491 in the final **GSMN**.

1.3.1.9 Reconstruction transformation into a constrained model

From a reconstruction perspective, achieving a consistent flux model involves addressing common and known issues to prevent inconsistencies and inaccuracies in the model's flux behaviour. Therefore, the ***i*Prub22** model underwent refinement through adjustments, such as determining appropriate reaction reversibility, resolving redundant reactions or identifying and eliminating infeasible cycles.

Firstly, during the reconstruction loading, it was discovered that 61 reactions contained errors due to the ambiguous compounds representation serving as both reactants and products (*i.e.,* stoichiometry represented by the metabolite name). The lack of clarity in their definition, attributed to the utilisation of metabolites belonging to classes or superclasses, led to 56 blocked reactions (*i.e.,* reactions involving overly generic and unbalanced compounds) and the correction of 5 others. Furthermore, the intracellular compartmentation is not explicitly modelled within ***i*Prub22**. Consequently, 11 reactions related to the transport between the cytoplasm and organelles (*i.e.,* mitochondria or peroxisomes) are inadequately annotated. As these reactions are not relevant to the current objectives of the model, they have been blocked.

Secondly, we focus on the significance of the reversibility of reactions. When the reaction directionality is unknown, it is often assumed as reversible, which can introduce biases in model development and comprehension. To address this, we utilised evolving database annotations and improved knowledge in the field. Thus, we subjected the 967 reversible reactions of the initial reconstruction to further scrutiny using the SmartTable tool on the MetaCyc server. Among the 787 reactions retrieved from MetaCyc, 98 had their reversibility updated by closing one of their bounds. In addition, 165 reactions were without annotation labels, indicating unknown directionality, and thus were blocked. These initial modifications

* Figure 3-18, page 253

yielded a model that exhibits sensitivity to nitrogen variations. Specifically, by manipulating the openings of various nitrogen sources, we discern changes in the flux distribution within the model. This sensitivity highlights the significance of nitrogen availability and its impact on metabolic processes.

Thirdly, we dealt with duplicate reactions contained within the reconstruction. As a result, we identified 49 pairs of similar reactions, with one possessing a MetaCyc identifier and the other using a “homemade” identifier due to the reconciliation process. In such a case, we retained the more balanced reaction from each pair. Furthermore, we meticulously examined the GPR association, ensuring either their complete identity or the inclusion of all genes from one reaction into the other. Then, to maintain consistency within the stoichiometric matrix and ensure the reaction uniqueness, we identified and blocked 79 instances of duplicated reactions.

Fourthly, we encountered issues related to energy production when the model was fully closed, as well as complications with metabolite production. These observations suggested the presence of infeasible cycles within the network, which can disrupt proper metabolic functioning. Furthermore, at this stage, the model did not exhibit sensitivity to variations in carbon availability and sources, highlighting the need for further adjustments. To address these challenges, we blocked some of the non-equilibrated reactions. This step ensures the overall mass and charge balance of the model, guaranteeing that it adheres to fundamental principles of stoichiometry and improves its accuracy and reliability for subsequent analyses. Out of the 5,919 reactions present in the reconstruction, 1,398 (23%) were found to be mass imbalanced, indicating inconsistencies in the stoichiometry of the participating metabolites. Additionally, 969 reactions (16%) exhibited charge imbalance, implying a discrepancy in the overall charge of the reactants and products. Regarding the 5,464 metabolites encompassed in the reconstruction, 3,517 (65%) were exclusively involved in balanced reactions, demonstrating that their participation maintained the system equilibrium. Conversely, a small fraction of 192 metabolites (3.3%) lacked molecular formulae, emphasising the need for further

annotation and characterisation. Among the 1,402 imbalanced reactions identified, 228 were exchange reactions, 10 were associated with biomass production and its assimilated reactions, 170 had been closed during previous curations mentioned earlier, 25 were indispensable for biomass production, 23 were involved in the synthesis of specialised metabolites, and 946 were effectively blocked. This intervention helped resolve the problems related to energy production and metabolite synthesis, leading to a more functional and realistic network representation. Indeed, this approach prevented the occurrence of energy production from non-physiological sources when the model is closed, ensuring the model's compliance with fundamental principles of energy conservation and maintaining its biological realism.

Thus, once all these modifications have been applied (*i.e.,* 1,389 reactions, 23% of the reactions in the reconstruction had at least one of their bounds modified), and when all uptakes are unrestricted, the resultant model consists of 1,667 active reactions, accounting for approximately 28% of the overall reconstruction. Notably, 392 (24%) of these reactions are reversible, indicating the system's capability to accommodate bidirectional fluxes. This proportion underscores the importance of considering the flexibility and adaptability of metabolic pathways in response to changing conditions. All of the above model modifications and information are reported in Supporting Information (S7 file*) to be compliant with the MIASE standards *(Waltemath et al. 2011)*.

Finally, the model quality was assessed based on its ability to simulate organism growth. Biologically, the biomass reaction models the organism synthesising essential amino acids, membrane lipids, and sugars. When the predicted growth rate is null, this implies the incompleteness of the **GSMNs**. It results either from missing reactions in the biosynthetic pathway or an accumulation of one or more reaction products due to the absence of a degradation reaction of these metabolites. Thus, to overcome this issue, manual analysis was carried out either by adding outward transport reactions or determining the missing reactions through similar organisms' **GSMNs** analysis. According to previous studies, *P. rubens* has a maximum growth rate of around 0.17 h^{-1} when it grows on sucrose *(Robin et al. 2001)*.

* Annexe 4, page 451

Wild-type *Penicillium* species were reported to grow on glucose/sucrose at a rate between 0.14 and 0.22 h^{-1} *(Grijseels et al. 2017)*. In the ***i*Prub22** model, we conducted simulations by setting import rates of sulphur, riboflavin, thiamine, phosphate, and iron to 10 $mmol.gDW^{-1}.h^{-1}$, while oxygen import is unlimited. The carbon source is supplied at a rate of 15 $mmol.gDW^{-1}.h^{-1}$, and the nitrogen source at 5 $mmol.gDW^{-1}.h^{-1}$. When amino acids serve as both carbon and nitrogen sources, their import rate is adjusted to 15 $mmol.gDW^{-1}.h^{-1}$. All other uptake reactions are closed. In natural conditions, *P. rubens* requires a nitrogen source for growth, and our model accurately reflects this dependency. Moreover, the inability of the fungus to grow when cysteine is used as a nitrogen source aligns with characteristics expressed by *P. rubens* in real-life conditions. Under the reference conditions of using glucose and ammonia as carbon and nitrogen sources, the model predicts a theoretical growth rate of 0.2944 $mmol.gDW^{-1}.h^{-1}$. Reasonably, a potential correlation is observed between fungal growth and the increasing carbon availability in the environment, as depicted by the variations in the carbon source represented in **Figure 3-3**. However, when we substituted glucose with sucrose as the carbon source, a threefold increase in production was observed compared to the theoretical value *(Robin et al. 2001)*. Furthermore, **Figure 3-3** illustrates additional simulations conducted to evaluate the growth capacity of the **GSMN** model under various media conditions.

Effect of Different Media Compositions on *Penicillium rubens* Growth: Optimal Solution Values for Biomass Production

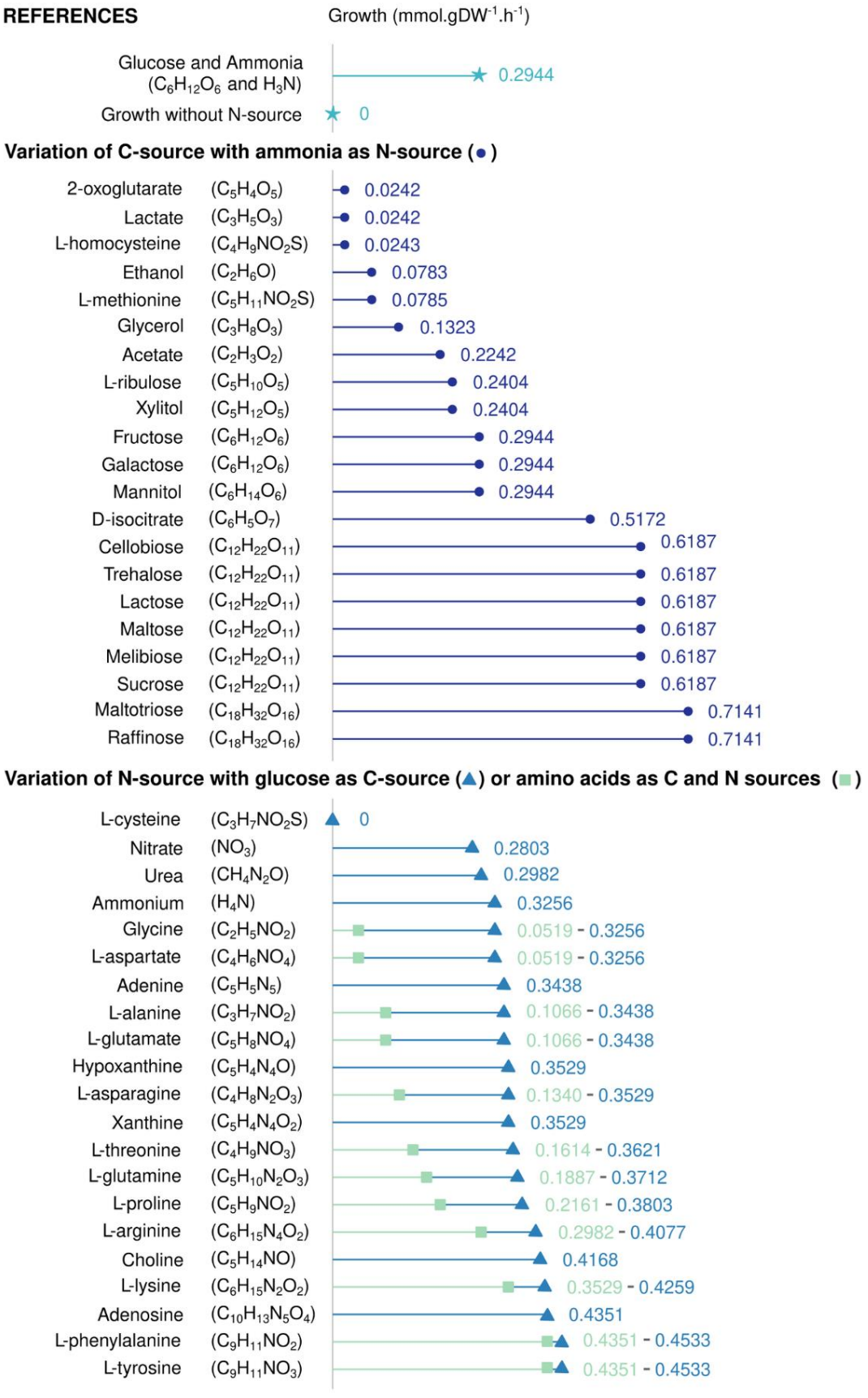


Figure 3-3: Theoretical growth of *Penicillium rubens* on different simulated media. Reference conditions are denoted by the light blue star in the figure (★). The variations in the carbon source are represented by dark blue circles (●). The blue triangles (▲) represent the optimum biomass flux observed when the nitrogen source is altered. The green squares (■) indicate the maximum biomass flux achieved when an amino acid functions as carbon and nitrogen sources. Import rates of sulphur, riboflavin, thiamine, phosphate, and iron are set to 10 mmol.gDW^{-1}.h^{-1}, while oxygen import is unlimited. The carbon source is supplied at a rate of 15 mmol.gDW^{-1}.h^{-1}, and the nitrogen source at 5 mmol.gDW^{-1}.h^{-1}. When amino acids serve as both carbon and nitrogen sources, their import rate is adjusted to 15 mmol.gDW^{-1}.h^{-1}. All other uptake reactions are closed. For more detailed information, please refer to the additional data presented in the S6 file.

1.3.2 Comparisons and evolution of *Penicillium rubens* GSMN

1.3.2.1 Connectivity and metabolic coverage

The **GSMNs**, *i*AL1006, Prubens and our reconstruction encompass 1,660, 2,574 and 5,919 reactions that connect 2,429, 3,058 and 5,464 metabolites, respectively. These networks are supported by 1,006, 1,786, and 5,703 different genes, bringing the number of *P. rubens* genes into ***i*Prub22** (*i.e.,* metabolic coverage) to 45%, surpassing the respective percentages of 8% and 14% observed in the first and second reconstructions from 2013 and 2018. Therefore, besides an increasing number of reactions associated with better metabolic coverage links to recent improvements in databases and annotation methods, a significant evolution observed over the years is logically the enhanced connectivity within the models themselves (**Figure 3-4**).

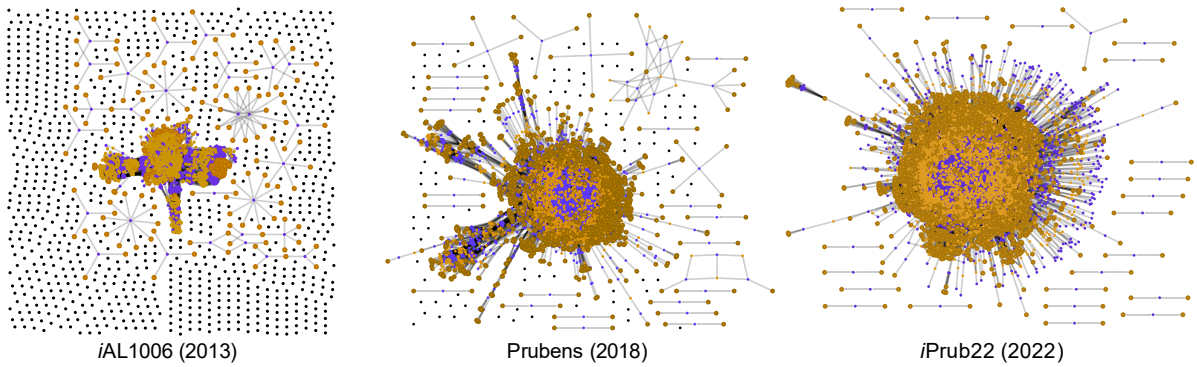


Figure 3-4: *Penicillium rubens* network visualisation through bipartite graphs. Illustration of the progressive evolution and improved topology of the *Penicillium rubens* network over time. The purple nodes (■) correspond to reactions, while the orange nodes (■) represent metabolites. Disconnected metabolites are depicted as black nodes (■). These images were generated using ModelExplorer (version 2.1) *(Martyushenko et Almaas 2019)*.

Among the 5,703 genes in ***i*Prub22**, 30% (3,771 genes) are newly added to the reconstruction, while 6% (800 genes) are shared with both previous network versions. Additionally, 7% (939 genes) are shared with the most recent version of the network, Prubens, while 2% (193 genes) are mutual with the oldest, *i*AL1006. The 53 genes contained in previous versions and not found in ***i*Prub22** are mainly involved in metabolite transport (data obtained with FungiFun, not shown). According to the GOT annotation provided by FungiFun (Fig 13 in S1 file), the genes contributing to this new network have annotations covering various biological and metabolic processes such as antibiotic responses, signal transduction, protein phosphorylation and molecular binding. Also, the KEGG annotation indicates that processes involving DNA and RNA (*e.g.,* nucleotide metabolism, transcription, translation, replication and repair) are present. These categories are reflected in the annotation provided by FunCat. The 3,771 genes included in the reconstruction are associated with a range of 1 to 61 reactions. On average, each gene is linked to approximately 4.5 reactions, with a median value of 2 reactions. Furthermore, among the 6,884 sequences that are not included in the ***i*Prub22** reconstruction, approximately 61% of them have a substantial length capable of encoding functional enzymes, with an average size of approximately 370 amino acids (**Figure 3-5**).

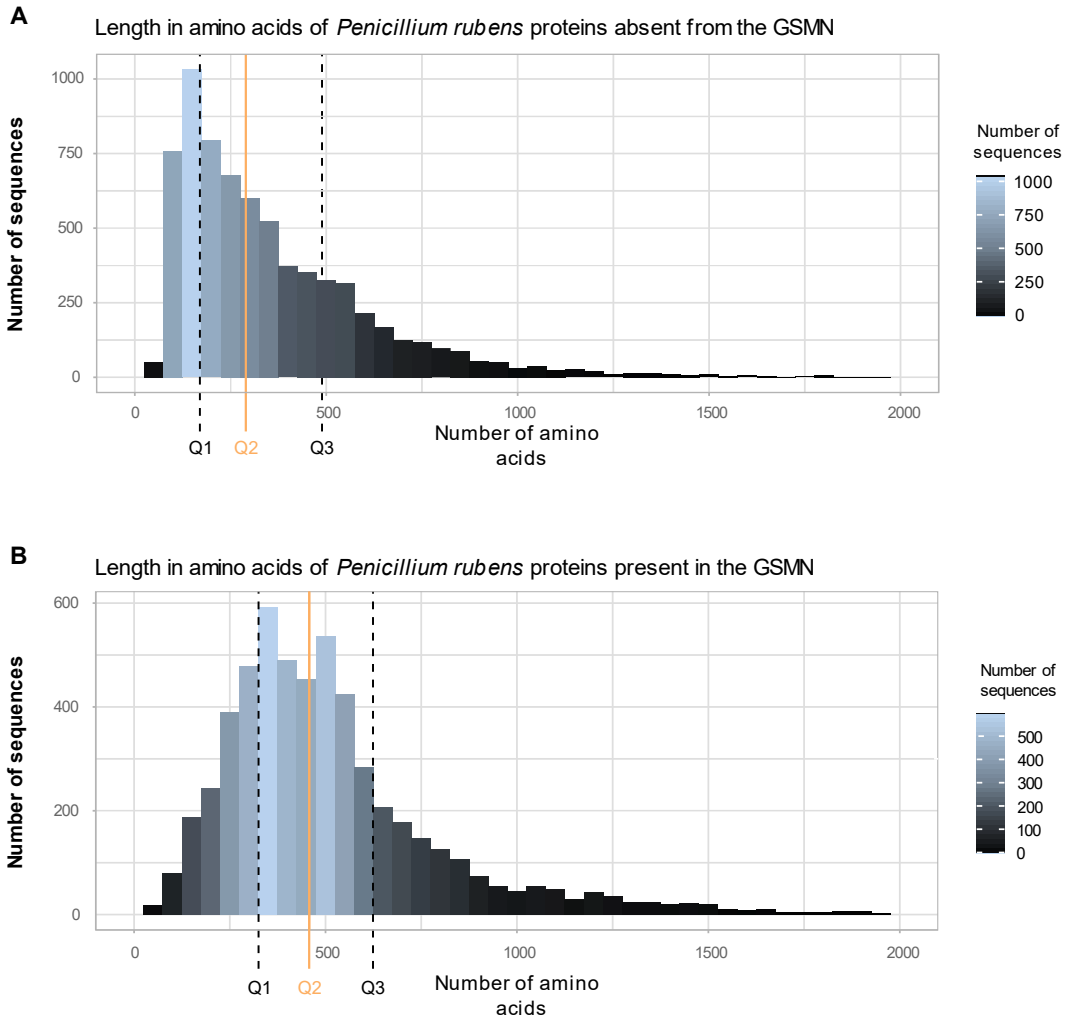


Figure 3-5: Amino acid length of *Penicillium rubens* sequences not included (A) and present (B) in the reconstruction. Median (continuous orange line ■) and quartiles (black dot lines) are shown. A range of 50 amino acids is displayed. (A) The average size of the 6,884 *P. rubens* sequences is 371 amino acids, and the median is 286. Measuring intervals: [15(min)-169];]169-286];]286-480];]489- 3,852 (max)]. (B) The average size of the 5,703 *P. rubens* sequences is 539 amino acids, and the median is 462. Measuring intervals: [28(min)-328];]328-462];]462-624];]624-7,287 (max)].

1.3.2.2 Specialised metabolism through versions

The initial **GSMNs**, *i*AL1006 and Prubens, involved 1,395 and 3,058 metabolites distributed across 5 and 14 compartments, respectively. The evolution and comparison of ***i*Prub22** metabolites with those of previous networks are based on MetaCyc (v23.0) identifiers for Prubens and the first section of InChIKeys generated from InChI for *i*AL1006. Indeed, due to the identifiers incompatibility, the comparison with the 2013 network was not direct and required caution in the analysis of the results.

This latter **GSMN**, *i*AL1006, included 849 unique metabolites (*i.e.,* after removing compartment information), of which 554 (62%) are annotated with an InChI. Conversion of these identifiers to InChIKeys *(O'Boyle et al. 2011)* yielded 512 different first sections InChIKeys (*i.e.,* the first section of InChIKey corresponds to the structure without stereochemistry). The ***i*Prub22** reconstruction is associated with InChIKeys identifiers for 3,651 (67%) of its metabolites, with 3,382 being unique in their first section. A comparison of the InChIKeys revealed that 24 metabolites were exclusively present in *i*AL1006, while 488 metabolites were shared between ***i*Prub22** and *i*AL1006. Among these 488 metabolites and continuing to disregard stereoisomerism, their first InChIKey section allows retrieving 617 compounds in ***i*Prub22**. Interestingly, 6 of these compounds were classified as "a secondary metabolite" according to the MetaCyc ontology.

Additionally, Prubens contained 2,594 unique metabolites, of which 2,415 (93%) possess a MetaCyc identifier (v23.0). Among these compounds, 2,377 were found in ***i*Prub22**, and only 38 were absent. Thus, ***i*Prub22** encompassed 93% of Prubens' metabolites and was enriched with 2,825 additional entities. In the MetaCyc compound ontology, Prubens had 218 (8.4%) metabolites classified as "a secondary metabolite", 211 of which are shared with ***i*Prub22**, which itself contains 434 (8.4%) entities of this class.

Subsequently, our focus shifted towards understanding the predictive production capabilities of ***i*Prub22** for a well-characterised list of specialised metabolites displayed in **Figure 3-6**. Thus, we employed FVA to assess the potential production of compounds, whereby the maximisation of biomass was set at an arbitrary threshold of 80%. It is necessary to relax the constraints imposed on the system, allowing for the evaluation of flux variations in the exchange reactions of the targeted metabolites. As expected, in the absence of constraint relaxation, the flux variations were found to be null, indicating a lack of superficial production. This last point ensures the validation of the model's constraint definitions while simultaneously enabling an assessment of the potential production of specialised metabolites within the examined system. We evaluated eight biosynthetic pathways: seven were identified using the

SMASH tool, and the eighth used mining reconstruction (*i.e.,* pathways involving averufin and versicolorin B). In the context of databases such as MetaCyc, a pathway refers to a curated collection of biochemical reactions and associated genes or enzymes that are known to be involved in a specific biological process or metabolic. To provide a comprehensive analysis, we adopted a unified categorisation approach by grouping all intermediates, specialised metabolites, and by-products under the collective term "specialised metabolites" without differentiation between them. Consequently, our classification encompasses two distinct categories: specialised metabolites and precursors. Indeed, the production of metabolites, such as andrastin A, yanuthone D, and PR-toxin, poses challenges due to the absence of a MetaCyc identifier or associated reactions for their synthesis within the database. Therefore, it becomes necessary to focus on studying their precursors to gain insights into their theoretical distribution fluxes. Furthermore, the potential production of isoepoxydon, roquefortins and assimilates, averufin and assimilates requires the addition of Red-NADPH-hemoprotein reductase, an unbalanced compound with a radical in its empirical formula. This particular compound serves as a flavoprotein catalyst that facilitates the reduction of cytochrome. Additionally, the availability of phenoxyacetate is essential to unblock the flux of penicillin V, while the import of decanoate or octanoate was essential for the production of penicillin K.

Determining the Maximum Variation in Fluxes for Biosynthesis of Specialised Metabolites with 80% Biomass Maximisation, Glucose as C-Source, and Ammonia as N-Source

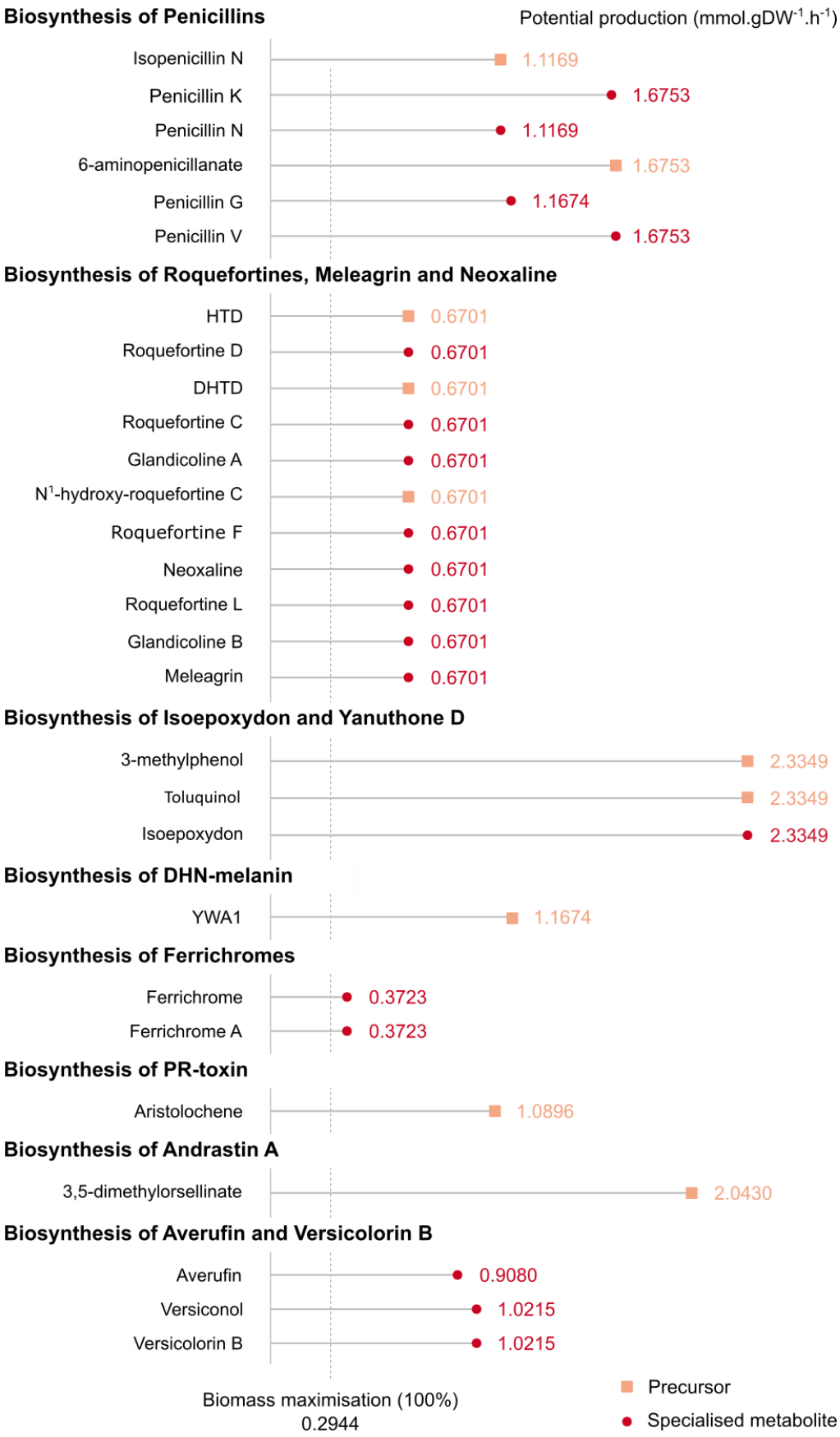


Figure 3-6: Potential production of specialised metabolites in the *i*Prub22 model under reference conditions. The figure displays the maximum flux of reactions associated with specialised metabolites biosynthesis (represented by red circles ●) or their closest known precursors (indicated as salmon squares ■) present in the ***i*Prub22** model when the biomass is maximised to 80%. In addition to the imports required for biomass production (*i.e.* sulphur, thiamine, riboflavin, phosphate, oxygen, iron, glucose as C-source and ammonia as N-source), the uptakes of decanoate, phenoxyacetate and a general flavoprotein (NADPH-hemoprotein reductase) are open to allow the production of the depicted compounds shown.

1.3.2.3 Interoperability

To ensure the interoperability of ***i*Prub22**, the identifiers of metabolites and reactions were kept consistent with MetaCyc and cross-referred through MetaNetX *(Moretti et al. 2021)*. We focus on various annotation enhancements (*e.g.,* identifiers interoperability of metabolites, reactions, and genes between known databases) and on the addition of Systems Biology Ontology terms *(Lieven et al. 2020)* (SBO) to match the standards established *via* MEMOTE. SBO annotations provide semantic information that characterises the model components *(Courtot et al. 2011; Bernasconi et Masseroli 2019)*. A relatively low match rate on the BiGG (13%) and KEGG (43%) historical reconstruction databases was noted concerning metabolites annotations. On the other hand, more than two-thirds of the metabolites had a chemical identification (*e.g.,* ChEBI, PubChem). Details are given in "**Tableau 3-2**", and ***i*Prub22**'s different characteristics are available in Supporting Information in the form of a MATLAB live script (S7 file*). The reconstruction presented here is rated with a MEMOTE score of 74% *(Lieven et al. 2020)*. Finally, based on these various annotations, it remains possible to complete them with mass spectral information from the MassBank *(Horai et al. 2010)* and GNPS *(Wang et al. 2016)* mass spectral databases for further identification in real biological samples. Based on a mapping from the InChIKey, 508 metabolites were associated with at least one spectral data.

Finally, interoperability between tools was ensured by upgrading the network distribution format to comply with **SBML** Level 3 *(Hucka et al. 2018)*.

* Annexe 4, page 451

Tableau 3-2: Details of the GSMN enrichment (nature and number of metadata).

Metabolites		
Entities number		5,464
▪ **Structure identifiers**		
○ InChI	3,886	71%
○ InChIKey	3,887	71%
○ SMILES	5,276	97%
○ Molecular weight	5,270	96%
▪ **Database identifiers**		
○ BiGG	717	13%
○ BioCyc	5,441	99.6%
○ CAS	959	18%
○ ChEBI	3,568	65%
○ Chemspider	1,487	27%
○ DrugBank	237	4.3%
○ HMDB	1,549	28%
○ KEGG	2,371	43%
○ KNApSAcK	96	1.8%
○ LIPIDSMAPS	138	2.5%
○ MetaboLights	1,052	19%
○ MetaNetX	5,436	99.5%
○ ModelSEED	4	0.07%
○ PubChem	3,693	68%
○ SABIORK	12	0.22%
○ SwissLipids	26	0.48%
○ UMBBD	50	0.92%
▪ **SBOTerm**		
○ Simple chemical SBO:0000247	5,464	100%
Reactions		
Entities number		5,919
▪ **Database identifiers**		
○ BiGG	4	0.07%
○ BioCyc	5,162	87%
○ Brenda	4,343	73%
○ KEGG	2,470	42%
○ MetaNetX	5,436	92%
○ Rhea	2,873	49%
○ Seed	1,329	22%
▪ **SBOTerm**		
○ Biochemical or transport reaction SBO:0000167	3	0.05%
○ Biochemical reaction SBO:0000176	5,280	89.20%
○ Translocation reaction SBO:0000185	60	1.01%
○ Exchange reaction SBO:0000627	228	3.85%
○ Demand reaction SBO:0000628	37	0.63%
○ Biomass production SBO:0000629	1	0.02%
○ ATP maintenance SBO:0000630	1	0.02%
○ Sink reaction SBO:0000632	1	0.02%
○ Transport reaction SBO:0000655	53	0.90%
○ Active transport SBO:0000657	38	0.64%
○ Passive transport SBO:0000658	108	1.82%
○ Symporter transport SBO:0000659	21	0.35%
○ Antiporter transport SBO:0000660	2	0.03%
○ Spontaneous reaction SBO:0000672	86	1.45%

1.4 DISCUSSION

1.4.1 The strengths and limitations of reconstruction automation

Since the publication of a reference protocol for the reconstruction of **GSMNs** *(Thiele et Palsson 2010a)* and the emergence of omics data, numerous microbial models have emerged, essentially of bacteria. Methods established primarily from micro-organisms have been automated *(Faria et al. 2018; Gu et al. 2019)*, allowing simultaneous and multiple eukaryotic **GSMN** creations *(Pitkänen et al. 2014; Prigent et al. 2018)*. In the meantime, several tools *(Faria et al. 2018; Mendoza et al. 2019)* have also been developed, raising interoperability issues and the well-known need for standardisation. Thus, AuReMe *(Aite et al. 2018)* was chosen for its data traceability and network visualisation capabilities. This toolbox merges data from functional annotation, orthology search from previously published models and manual expertise. Traceability of the reaction origin has a double objective. It provides information both on evolutionary and phylogenetic concepts, thus highlighting the core genome, which mainly characterises the genes involved in essential intermediary metabolism, DNA replication and repair, and transcription and protein synthesis. Additionally, traceability provides a quality estimation of the reconstruction by giving a 'confidence weight' to the reactions. This manual curation can be performed in two ways, focusing either on reactions or on genes. Thus, all of the information presented in **Figure 3-1** and **Figure 7*** in the S1 file are clues to guide the refinement of the draft toward obtaining a **GSMN**. In addition, the complementarity between functional annotation, orthology search and external sources addition during the reconstruction process is also visible in **Figure 8**† in the S1 file. Hence, the used approaches and data complementarity are the heart of the **GSMN** reconstruction process for *P. rubens*. Therefore, is there a legitimate interest in cross-referencing all of this information which is fundamentally and originally the identical mechanism result, which is sequence homology?

The first question that can be asked concerns the choice between carrying out a reconciliation of pre-existing data or a *de novo* reconstruction. As described by Oberhardt *et al.*, *(2011)* the reconciliation process implies that for the two previously

* Annexe 3C, page 449

† Annexe 1D, page 431- Annexe 2F, page 442 - Annexe 3B, page 447

reconstructed networks, if there is “no discriminating biological evidence for a given reaction to being included in one reconstruction but not the other, the reactions should be identically included in both”. The struggle is real since working with reconstructions using incompatible or “homemade” identifiers is a significant source of error and a complicated task (**Figure 6*** in S1 file). However, in terms of network evolution perspectives, the consensus and systematic comparison of these data could lead to improvements (*i.e.,* refining genomic information by updating the GPR associations) for future manual curation.

These comparisons provide information on the complementarity of the approaches and the sensitivity/specificity of the GPR associations. Thus, one of the limits of the automated reconstruction lies in the assignment of genes for non-specific reactions (*i.e.,* a large number of genes being associated with such type of reaction). Moreover, intra-operability and especially identifiers durability as the databases are updated is sometimes not guaranteed. For example, of the 1,291 reactions from PchCyc, 79 (6%) have an incompatible identifier with MetaCyc version 23.0. For these reasons, few tools currently exist that allow automatic comparison of these networks, although efforts are being made in this direction *(Hari et Lobo 2022)*. As a result, relatively few model organism reconstructions benefit from pre-existing data *(Aminian-Dehkordi et al. 2019; Nouri et al. 2020)*, and when they do exist, they require a significant community effort *(Witting et al. 2018)*.

Another example of information loss is related to database mapping, which arises in particular when integrating subnetworks obtained by orthology search. During the creation of the orthology subnetwork, more than half of the reactions were discarded due to a lack of MetaCyc-compatible identifiers. These reactions, supported by genomic data, could, therefore, constitute a preferred pool for future completion of the **GSMN** (*i.e.,* addition of reactions or refinement of existing GPRs). For example, from the 546 lost reactions during the mapping on *N. crassa*, 505 were allocated to metabolite transport and had model-specific identifiers. Only manual curation will allow for the recovery of such information. Thus, this known interoperability problem (*i.e.,* lack of MetaNetX *(Moretti et al. 2021)* identifiers or model-specific identifiers)

* Figure 3-10, page 205

highlights the need for uniform writing standards both at the database level and in the intrinsic writing of models. Also, this points out the issue of the relevance of reconciling the subsequently reconstructed networks. While these networks carry relevant genomic information, their integration is tricky, as shown in the section on modifications of the reconstruction to obtain a viable flux model.

The second question is whether orthology search should be worth using in addition to functional annotation. Are these two approaches redundant or complementary? As orthologous genes arise by speciation events, these groups of genes are assumed to share identical biological functions *(Fang et al. 2010)*. By definition, the search for orthologous sequences allows the metabolism section detection common to organisms (*i.e.,* consensus), thus highlighting their evolutionary history. Constitutive metabolism will therefore be mainly targeted by this type of approach since it is shared between various species. On the other hand, due to its specificity, specialised metabolism can only be incorporated into the reconstructions by functional annotation or, if it fails, by manual curation. The functional annotation approach should, therefore, cover the one performed by orthology search. However, 16% of the reactions in the draft still come exclusively from one or more subnetworks resulting from the orthology search. Therefore, this method also helps to benefit from the manual expertise within the reconstructions used for orthology search, as it would appear that the information they contain is not systematically referenced in the functional annotation database (*i.e.,* 764 reactions are exclusively specific to the orthology subnetwork – **Figure** 3-1). In addition, one would also expect the most informative templates to be the most phylogenetically related. In the case of the ***i*Prub22** reconstruction, most information obtained by orthology search came from *S. japonica*. Therefore, the most informative template for the reconstruction is not fundamentally the most biologically relevant. Since the selected **GSMNs** were published between 2010 and 2018, and *S. japonica* **GSMN** was the most recently reconstructed network using a protocol very similar to the one used in this study, this reflects a lack of details in the past and present data rather than proper evolutionary concepts.

As mentioned above, the specialised metabolism of an organism can probably only be raised by functional annotation. What about its presence within the proposed reconstruction? This point can be illustrated by focusing on two compound families: roquefortines and penicillins. First, roquefortines and meleagrin are indole alkaloids produced by several *Penicillium* and *Aspergillus* species from tryptophan and histidine. Knowing that all the information related to the synthesis of these compounds is encoded in MetaCyc (since version 20.5) and is grouped under the PWY-7609 super-pathway, why is there no trace of them, even partial, in the draft? As described in Guzmán-Chávez *et al.* *(2018)*, roquefortines and meleagrin are catalysed by 13 reactions mediated by 6 enzymes/genes. Since their pathways are not described in previous reconstructions or orthology templates, the only possibility to incorporate this specific pathway in our reconstruction is from the functional annotation subnetwork. However, although the 6 genes involved in the biosynthesis of these compounds were present in the draft and encoded between 2 and 20 reactions, no reactions related to the pathways concerned were found in the draft. Backtracking its absence, through traceability provided by AuReMe, indicated that this absence is related to an error in the PathoLogic module use. It unintentionally and automatically pruned 9 genomic sequences whose 6 belong to the roquefortines pathways. Concerning the second example, within MetaCyc, the reactions leading to different penicillins biosynthesis are listed in 3 pathways: PWY-5629 (isopenicillin-N biosynthesis – 2/2 reactions found in the draft); PWY-7716 (penicillin G and penicillin V biosynthesis – 3/4 reactions found in the draft); PWY-5630 (penicillin K biosynthesis – 1/1 reactions found in the draft). Although their names vary from one database to another and within the literature, the genes responsible for the production of these compounds were identified as pcbAB or ACV or ACVS (*Pc21g21390*); pcbC or IPN or IPNS (*Pc21g21380*); pcbDE or iAT or AAT or PenDE (*Pc21g21370*); phl or pclA (*Pc21g14900*). PenDE, ACVS and phl were discarded by PathoLogic during reconstruction like the roquefortines biosynthetic pathways genes. It would appear that discarded genes and associated reactions are deleted if, and only if, they are not supported by any other gene. For example, the following two reactions are catalysed by

PenDE: RXN-1702 and RXN-17100 (*i.e.,* production of penicillin G and interconversion of penicillin G to penicillin V, respectively). For the first reaction, PathoLogic infers two sequences (*i.e.,* one corresponding to PenDE based on the EC-number and another based on the GO annotation). Since this reaction is supported by two genes, the tool does not perform any pruning step. Conversely, the second reaction is only supported by PenDE and is consequently excluded from the reconstruction, as are all reactions related to the synthesis of roquefortines. Nevertheless, the reconstruction methods' complementarity allows us to recover almost all of the penicillin biosynthetic reactions, even if the GPRs are not precise enough. Moreover, the genes associated with penicillins synthesis, *Pc21g21380* (IPNS), *Pc21g21370* (PenDE), and *Pc21g14900* (phl), were present in the draft and supported between 2 and 12 reactions. However, *Pc21g21390* (ACVS) was absent from the draft even though the reaction it catalyses was found in it and supported by 10 other genes found by orthology. These results raise questions about the relevance and accuracy of the automatically deduced GPR associations and, thus, more broadly, about the quality of the networks used for reconstruction through orthology. Therefore, in addition to the entire PWY-7609 pathway for roquefortine biosynthesis (Table 3 in S1 file), manual curation was limited to the inclusion of the RXN-17100 reaction (PenDE-mediated interconversion of penicillin G to penicillin V) and the ACVS gene for the 6.3.2.26-RXN reaction (the first reaction inducing isopenicillin-N synthesis). Even if GPRs' specificity and accuracy are essential for model simulations (*e.g., in silico* knock-out), verbose GPRs within a reconstruction remain sources of information to explain evolving concepts. Thus, GPR associations cleaning (Table 4 in the S1 file) was not performed.

The third and last point concerns the limits of reconstruction automation, notably the gap-filling steps. Gap-filling makes it possible to complete the pathways from a structural point of view. As this process is based on heuristics and the absence of genomic data, additional results from such tools must be done sparingly and carefully. Here, sets of targets were used to perform gap-filling against proper subsets of MetaCyc depending on the purpose (*e.g.,* preferential selection of spontaneous or fungi reactions). Iterative processes were performed with the repair networks and targets from most to least relevant to achieve

'reasoned gap-filling'. Nonetheless, beware that performing a complete gap-filling, *i.e.,* using the set of putative metabolites present in our reconstruction against the entire MetaCyc database, is absurd from a biological and evolutionary point of view. This set was used only against controlled versions of repair networks (*i.e.,* for the detection of spontaneous reactions and for the last round of gap-filling whose supporting genomic sequences are unknown in MetaCyc). By reasoning in this way, it is possible to give an additional confidence weight to almost half of the gap-filling reactions (**Figure 12*** in S1 file). As shown in the results presented in "**Tableau 3-1**", gap-filling improves the connectivity of the **GSMN**. The semi-automated process for the draft generation allows topologically reach 86% and 22% of the selected targets belonging to the constitutive and specialised metabolisms, respectively. These numbers demonstrate the methods' efficiency for the automatic reconstruction of primary metabolism. It also points out how far we still have to go with regard to specialised metabolism. Furthermore, it should be remembered that, as described in Dusad *et al.* *(2021)*, topological approaches are limited by two well-known issues, namely hub metabolites and reactions reversibility. This last point will be discussed later in the section on the reconstruction evolution to the flux model.

Finally, what parameters should be considered as a plausible possible model, as close as possible to biological reality (*e.g.,* addition, modification, deletion of reactions and constraints)? This point should be addressed at various scales, from the most generic (*e.g.,* addition of layers of information related to compartmentation, spontaneous reactions and exchange reactions) to the most specific and precise (*i.e.,* refinement of metabolic pathways with a focus on the nature of GPR associations). Due to their nature, these steps often require manual expertise and thus highlight the limits of such automation. As ***i*Prub22** is intended to be improved in this sense with feedback from the scientific community, a user-friendly wiki is available to facilitate communication and improvement for future versions. The wiki associated with the **GSMN** offers a less cumbersome virtual alternative to the efforts developed during the nonetheless indispensable processes of community meetings over the past years (*e.g.,* jamboree or workshop) *(Thiele et Palsson 2010b; Wang et al. 2012; Naithani et al. 2019)*.

* Figure 3-18, page 253

1.4.2 GSMN quality

The **GSMN** origins date back to 1999, with the first model publication for the bacterium *Haemophilus influenzae* *(Edwards et Palsson 1999)*. Thiele and Palsson published a protocol for reconstructing metabolic networks in four key steps ten years later, which is still used as a reference today *(Thiele et Palsson 2010a)*. However, in the history of **GSMN** reconstruction, the development of uniform criteria for judging the quality of a model is relatively recent, encouraging the tendency to prefer quantity over quality *(Monk et al. 2014)*. Except for the MEMOTE test suite *(Lieven et al. 2020)*, there are still few tools or measures available to assess the completeness of a **GSMN**.

Thus, the **GSMN** presented here has been reconstructed following, among others, these three minima and general criteria defined by the work of Monk *et al.* *(2014)*:

- the extent of manual curation: what aspects and means are used to refine the model? This point was addressed in the previous section.
- metabolic coverage: what is the metabolic space, *i.e.,* the genome proportion, covered by the data present in the **GSMN**?
- the use of established standard operating procedures: what are the means and standards of writing to be used to encode the information? This step is independent of biological knowledge and allows the **GSMNs** application scope to be extended.

With the improvement of methods and the data proliferation, the *P. rubens* **GSMNs** evolution underlines a trend towards improved metabolic coverage over the years that seems to be enhanced substantially. The function of an enzyme is intrinsically dependent on its three-dimensional shape, *i.e.,* structure and folding, which in turn is induced by the amino acid sequence of which it is composed. It should also be noted that eukaryotic proteins are generally longer (*i.e.,* median size of 360 amino acids *(Brocchieri et Karlin 2005)*) than those of other living organisms. Therefore, the fewer amino acids a *P. rubens* sequence contains, the less likely it is to carry a functional catalytic domain. The sequence size distributions presented in **Figure 3-5** shows that of the sequences discarded from the **GSMN** reconstruction, almost half

are less than 300 amino acids in length. The remaining half is of sufficient size to encode functional enzymes and constitute a reservoir/pool of candidates for future improvement of the network. This set of genes with unknown functions and probably specific to the studied organism offers research perspectives for an improvement of future models (*e.g.,* unlocking the topological accessibility of certain dead-ends).

Computer modelling is one of the means used by Systems Biology to understand, organise and integrate large quantities of data from diverse sources (*e.g.,* molecular, cellular, *etc.*). Relying on ontologies (*i.e.,* set of controlled vocabulary) allows the exchange, interoperability and reuse of the models produced. This step, independent of model functionality, can occur at all stages of its life cycle. One aspect related to computational interoperability is associated with the annotation of the model entities in terms derived from the 'Systems Biology Ontology' (SBO) *(Courtot et al. 2011; Bernasconi et Masseroli 2019)*. This type of annotation is not specific to metabolism modelling. However, it allows, like any ontology, to generate a classification and a hierarchy of the data making them easier to interrogate for the user since controlled vocabulary use is part of good practice recommendations for computer modelling. In addition to the distribution format, network sustainability depends on its usability. It is well known that interoperability between databases is a delicate task. Therefore, the most exhaustive annotation of the entities in the model is essential and is now an undeniable criterion for increasing the quality of a network *(Ravikrishnan et Raman 2015; Community standards to facilitate development and address challenges in metabolic modeling 2020)*. Firstly, when these annotations belong to specific databases such as KNApSAcK (*i.e.,* a database dedicated to specialised metabolites), they constitute an entry point for model exploration. For information purposes, ***i*Prub22** contains 80 metabolites annotated with an identifier from this database. Secondly, the molecule's structural identifiers allow the entities to be identified and used as an anchor for querying other databases. For instance, using an InChIKey mapping considering only the first identifier section, thus disregarding the stereochemistry and compound's charge, 9.8% of **GSMN** metabolites are linked to mass spectral information retrieved from the GNPS and MassBank databases. This coverage is

lower than the average 40% detected by Frainay *et al.* work *(2018)*. This confirms the ever-limited aspect of obtaining and integrating metabolomics data acquired in given organisms within **GSMNs**. Such difficulty could be one more problem in validating **GSMN** behaviour, at the chemical level, with respect to biological observation.

However, as mentioned above, MEMOTE *(Lieven et al. 2020)* is a suite of tests aimed at qualifying a reconstruction or model by providing a summary of its characteristics both in terms of topology and flux. Its quality score considers the annotation content of the model. For example, a complete annotation in SBO terms guarantees a quality of 25%. This tool provides an easily accessible resource aiming toward model standardisation for detecting blocking points and/or abnormalities, thus helping manual curation. The **GSMN** we propose possesses a MEMOTE score of 74% against 31% and 16% for the 2013 and 2018 networks, respectively. Nevertheless, due to the biases taken during the reconstruction, such as the creation of artificial genes to help their identification, the systematic conservation of all data supported by genomic information and the annotation of some SBO terms not currently monitored, the 74% score is slightly underestimated (the MEMOTE report is available in S8 file*). Furthermore, this tool is designed and optimised to evaluate models that preferentially respond to the SBML3 FBC2 format *(Hucka et al. 2018; Olivier et Bergmann 2018)*. Since the previous networks were not encoded with the flux balance constraints (FBC) package, some characteristics cannot be evaluated properly, and thus, their MEMOTE results should be put into perspective. This is particularly true for analyses related to genes and their products, as this notation was only introduced in Version 2 of the FBC package.

The recurrent non-reproducibility of results, known as the reproducibility crisis *(Peng 2015)*, affects all areas of science and *in silico* modelling is no exception *(Tiwari et al. 2021)*. In fact, limited interoperability between software may yield incorrect models parsing, which can lead to errors or misunderstandings when analysing **GSMNs** *(Ebrahim et al. 2015)*. Thus, the proposed model was standardised in the most recent Versions of **SBML** and its extension package FBC. However, even if the emphasis on reconstruction has been placed on the

* Annexe 5, page 479

maximum transparency and traceability of data acquisition, with the constant evolution of databases and tools, providing all the information for the strict whole model reproducibility remains challenging. For this reason, **GSMNs** are regarded as knowledge state snapshots about an organism at the precise moment of its reconstruction, so a particular effort to keep on watch would be necessary to ensure their sustainability. Ultimately, the main objective of this reconstruction is to propose a resource grouping of known data for *P. rubens* based on these three key complementary notions: traceability, accessibility and visibility, focussing additionally on the exploration of specialised metabolisms' presence.

1.4.3 From reconstruction to a model that simulates environmental exchanges

Transforming a genome-scale reconstruction into a computational model for predicting properties and behaviour is a considerably time-consuming process that requires expertise. Even if steady-state constraint analyses are neither informative on the concentration of the metabolites nor on the temporal dynamic system aspects, it is still the more appropriate tool for exploring complex biological systems. These analyses are based on linear optimisation theories and constraints implementation based on three aspects: physicochemical (*i.e.,* steady-state, stoichiometry, thermodynamics), topological (*i.e.,* spatial modelling of intracellular and extracellular compartments) and environmental (*i.e.,* various nutrient media simulations defined by exchange and transport reactions) *(Price et al. 2004)*. The previously obtained reconstruction is used as a backbone for mathematical approach modelling, which will perform *in silico* predictive hypotheses-driven. However, the transition from the reconstruction model development to the fluxes models requires various adaptations, both computational and biological.

The uncertainty inherent in the **GSMN** predictions is influenced by the numerous tools, methods and algorithms that exist for both the reconstruction phase (*i.e.,* network structure) and the analysis phase (*i.e.,* simulation results). However, the specific sources and magnitudes of these uncertainties remain complex to quantify, which presents challenges for

evaluating the relevance and reliability of model predictions. In this respect, the recent great review by Bernstein *et al.* *(2021)* identifies five uncertainty areas surrounding the reconstruction and analysis of **GSMNs**. These fields include **(1)** genome annotation, **(2)** specification of environmental conditions, **(3)** formulation of biomass equations, **(4)** network gap-filling, and **(5)** flux simulation. Their work highlights the need to assess, communicate and understand the uncertainty sources associated with a model to discuss their impact on the relevance of predictions resulting from the model analysis. As functional annotation and gap-filling have been dealt with previously, we will focus here on the specification of the environment and the formulation of the objective biomass function. We would also like to stress that the focus of the work presented in this paper is on reconstruction considering specialised metabolism, and thus, we propose preliminary work on the network analysis of the flux distribution. Finally, an insight into the potential avenues for improving the model will be discussed in this section.

The biomass reaction is a commonly used objective function in constraint-based modelling, including FBA and FVA, that simulates the growth of a specific organism. It identifies essential growth compounds and assigns weights based on their occurrence, normalised to represent the dry-weight biomass of 1 gram. Understanding and optimising the growth of organisms requires defining the biomass reaction and its subsystems, which correspond to the synthesis of macromolecules from building blocks and precursors. Experimental values for growth-associated ATP maintenance (*i.e.,* GAM) and non-growth-associated ATP maintenance (*i.e.,* NGAM) are added to the reaction, representing the energy required for macromolecule polymerisation during growth and for maintaining the organism's viability, respectively *(Thiele et Palsson 2010a)*. However, in the absence of specific data, the biomass composition of a model organism is often used as a template, generating biases in the interpretation and aggregation of results when comparing **GSMNs** *(Bernstein et al. 2021)*. *Penicillium rubens* is a well-studied model organism, and experimental data indicate that its relative macromolecule content is distributed as follows: 45% proteins, 25% carbohydrates (including 22% cell wall compounds and 3% glycogen), 5% lipids (including 0.5% fatty acids,

1% sterol esters, and 3.5% phospholipids), 9% nucleic acids, and 8% each of ASH and the soluble pool *(Henriksen et al. 1996; Nielsen 1997)*. As the biomass reaction in *i*AL1006 was constructed based on these experimental data *(Agren et al. 2013)*, the same reaction is used in ***i*Prub22**. However, ***i*Prub22** lacks several subsystems found in *i*AL1006 that regulate the biosynthesis of fatty acids (*i.e.,* r1434), phospholipids (*i.e.,* r1460), lipids (*i.e.,* r1464), sterol esters (*i.e.,* r1462), and glycerides (*i.e.,* r1461). These subsystems could not be included in ***i*Prub22** due to the absence of compatible or relevant MetaCyc identifiers. Finally, refining or improving the biomass reaction for more accurate growth prediction can be achieved through sensitivity analysis, identifying critical reactions and key parameters for maximum biomass yield, or experimental measurements to determine the exact quantities of components needed for growth, allowing adjustment of the biomass reaction coefficients. For instance, the BOFdat workflow *(Lachance et al. 2019a)* may offer a prospective solution to incorporate lipid monitoring, which is currently not present in ***i*Prub22**.

Defining the chemical composition of the environment for metabolic modelling is a delicate task fraught with uncertainties, as described by Bernstein *et al. (2021)* (*e.g.,* inconsistent media definition in specific databases or undefined chemical input in experimental settings). However, environmental specification is a critical component of flux analyses and serves as the cornerstone for such investigations. As filamentous fungi are known to possess cellular machinery that enables extracellular nutrient absorption, uptake reaction selection was carried out by focusing on extracellular metabolites. Specifically, we integrated reactions from the well-curated *i*AL1006 model and complemented them with information on the extracellular localisation of gene products (*e.g.,* inclusion of reversible or non-reversible transport reactions between the intracellular and extracellular compartments, regardless of whether they are artefactual). Thus, we adopted the "seed" approach (*i.e.,* metabolites occurring naturally in the environment of the organism studied and essential input for the **GSMN**) to generate biologically relevant models that have the potential to enhance our understanding of metabolic processes and facilitate future investigations. Indeed with 185 uptake reactions included in the reconstruction, ***i*Prub22** offers a broad range of possibilities for modelling simulations.

Moreover, proposing diverse combinations will give clues on organism behaviour by identifying aspects of ***i*Prub22** predictions that exhibit either high sensitivity or resilience to changes in the environmental composition. Finally, once the nutrient nature has been defined, the subsequent step entails relying on experimental data to establish proper uptake bounds for flux simulation analyses.

Then, the modifications performed to deal with problems encountered when converting the reconstruction to a flux model focus mainly on updating and correcting **(1)** the directionality of reversible reactions (*i.e.,* adding constraints by closing reactions whose directionality remains unknown on MetaCyc), **(2)** the redundancy contained in the reconstruction and **(3)** the identification of blocked reaction.

The redundancy point is addressed at different levels. On the one hand, automatic and manual curation of duplicated reactions is carried out (*i.e.,* reactions with the same reactants and products). The presence of such reactions is explained by the duplicate's existence within the databases and by the data reconciliation contribution (*i.e.,* addition of reactions whose unbalanced equations do not allow their automatic detection). On the other hand, there is redundancy within the networks linked to varying metabolite names. Indeed, the same chemical entity may have several forms, and thus identifiers, that coexist within the model without having a bridge reaction between the different forms. Examples include the case of coexisting linear or cyclised states of molecules (*e.g.,* carbohydrates), the chirality encoding as a function of the different stereoisomers, or the use of more generic concepts with compounds class against single chemical species (*e.g.,* molecules ontology). For instance, if D-galactopyranose is a compound class composed of ALPHA-D-GALACTOSE and GALACTOSE (*i.e.,* beta form) and if an interconversion reaction of alpha and beta form exists, it seems reasonable to replace D-galactopyranose by ALPHA-D-GALACTOSE. From a computational perspective, this identifiers multiplication, or impoverishment when the annotation is too specific, leads to ultra-connected or under-connected network sections. Uniformity of identifiers on parent or child terms can thus reduce such artefacts. Concerning

Blocked reactions in a metabolic model, their presence indicates infeasibility and highlights potential errors, connectivity issues, lack of necessary metabolite inputs, or gaps in our understanding of the metabolism. These reactions are unable to sustain any metabolic flux under steady-state conditions and require investigation and correction to ensure model accuracy. As these corrections directly influence the outputs of these constraints-based analyses, all the modifications applied to the model generation are reported as additional data. By employing these constraint-based modelling techniques, we gained valuable insights into the metabolic behaviour and potential of ***i*Prub22**, further validating its utility as a predictive tool for *P. rubens* metabolic engineering and biotechnological applications.

The above-mentioned modifications (S7 file*) have yielded a model that exhibits diverse growth behaviours in response to variations in 35 different carbon sources and 35 different nitrogen sources (**Figure 3-3** displays a sample of the results, and the complete data are provided in the S6 file). This approach provided valuable insights into the models' performance and their responses to varying nutrient availability scenarios. Interestingly, the ***i*Prub22** model demonstrates its ability to reflect some observed behaviours with *P. rubens*. In accordance with Nielsen *et al.* observations *(1997)*, it is established that *P. rubens* requires a source of carbon, such as sugars, polysaccharides, organic acids, lipids, or certain amino acids, and a source of nitrogen, which can be provided by organic or inorganic compounds like ammonia, nitrate, or nitrite, to support its growth. However, it is worth noting that the carbon availability for respiration is relatively more limited compared to nitrogen. Thus, we have arbitrarily set the nitrogen limit at one-third of the carbon limit. Furthermore, it is observed that this fungus cannot grow in the nitrogen source absence. Besides, the amino acid L-cysteine alone is not a nitrogen source sufficient for the growth of *P. rubens*. These various behaviours are expressed by ***i*Prub22**. When ammonia and glucose are present, the model predicts a growth rate of 0.2944 $mmol.gDW^{-1}.h^{-1}$, which falls slightly outside the range of 0.14 and 0.22 h^{-1} established by the work of Grijseels *et al.* *(2017)*.

* Annexe 4, page 451

Nevertheless, by recalibrating the uptake boundaries, it will be possible to better align with the experimental data. However, it is crucial to acknowledge the presence of more divergences between the predictions of the model and the experimental findings. These disparities serve as valuable insights for further refining and improving the model, ensuring its accuracy and reliability in predicting biological phenomena. The study conducted by Allam *et al.* *(1969)* demonstrated that among the tested nitrogen conditions, the highest growth rates were observed in the following decreasing order: hypoxanthine, adenine, sodium nitrate, xanthine, urea, and ammonia. However, in our model, the optimal growth conditions were found to be with xanthine, hypoxanthine, adenine, ammonia, urea, and nitrate, as illustrated in **Figure 3-3**. Moreover, the *in silico* prediction of biomass flux on sucrose *(Robin et al. 2001)* surpasses the experimentally measured growth rate by a factor of 3.5. On the other hand, contrary to what the experimental data suggest *(Nielsen 1997)*, the model appeared not to be able to simulate growth on media whose carbon source is D-isocitrate, cellulose or acetate.

Finally, the model expresses mixed responses when amino acids can serve as both carbon and nitrogen sources. *In vivo*, the fungus shows a greater preference for amino acids that can rapidly degrade into a carbon source in one or two steps. These amino acids can be classified into three groups: L-glutamine, L-asparagine, L-arginine, and L-proline (group 1); L-glutamate, L-aspartate, L-alanine, and L-ornithine (group 2); L-histidine, L-glycine, L-isoleucine, L-lysine, L-leucine, L-methionine, L-phenylalanine, L-tyrosine, L-threonine, and L-valine (group 3). The model cannot offer a solution when L-histidine, L-tryptophan, L-methionine, L-isoleucine, L-leucine and valine are used as nitrogen and carbon source; amino acids from the third group, which are less suitable for the fungus. However, L-phenylalanine, L-tyrosine, and L-lysine, which also belong to this group, are the amino acids for which the simulated growth potential of *P. rubens* is the highest. These discrepancies suggest that the model may have insufficient constraints, allowing an exaggerated increase in the growth rate. Consequently, growth differences in different media may be attributed either to the high metabolic robustness of the

fungus or to the extreme pathways' presence, such as futile cycles inside the model *(Katz et Rognstad 1976)*. Thus, one of the most plausible hypotheses for correcting this inaccuracy is to investigate the presence of thermodynamically infeasible cycles (*e.g.,* type II or III extreme pathways, also known respectively as futile cycle or internal cycle) within the model *(Price et al. 2002)*. In metabolic network modelling, thermodynamically infeasible cycles are a group of reactions that violate the laws of thermodynamics when active (*i.e.,* unbalanced production or consumption). As thermodynamically infeasible cycles do not generate advantageous metabolites while consuming energy, they are not considered feasible pathways within the metabolic network. The presence of these cycles is generally attributed to inaccuracies or errors in the reconstruction process, stemming from incomplete knowledge of metabolic pathways or oversimplified assumptions regarding, for instance, reaction reversibility or the absence of regulatory processes. The most commonly used method for internal cycle detection is the II-Cobra method *(Schellenberger et al. 2011)*, and other algorithms have emerged over the years, such as the CycleFreeFlux algorithm *(Desouki et al. 2015)*. GlobalFit *(Fritzemeier et al. 2017)* identifies erroneous energy-generating cycles (*i.e.,* cycles that charge energy metabolites without a source of energy since they consume cofactors to generate motive power), which are futile cycles that are more difficult to identify and highly prejudicial to the simulations. Therefore, the presence of infeasible cycles can significantly impact the predictive capability of models, making it essential to identify and eliminate them to enhance the accuracy and reliability of the models.

Finally, a significant improvement of the proposed model will lie in GPRs verification and correction. For reconstruction, the false positive presence within a GPR carries information (*i.e.,* genomic sequence encoding an enzyme with a similar mode of action). Thus, the cross-referencing of false positives with GPR and gap-filled reactions would enable the pool creation of preferred candidates to be interrogated to detect, for example, new functionalities or to highlight evolutionary concepts. On the other hand, from the computational model side, erroneous GPRs greatly complicate the analysis of knock-out results *in silico*. Incorrect

predictions, linked to imperfect knowledge of the models studied, are an opportunity for biological discovery. Indeed, the abnormal behaviours of the model allow us to pinpoint the most relevant aspects of the metabolism and are, therefore, sources of hypothesis for addressing these failures. It is a valuable help in guiding experiments that often lead to new organism functionality characterisation.

1.4.4 What about specialised metabolites?

The reference strain *P. rubens* Wisconsin 54-1255 is the result of a long classical strain improvement programme *(Salo et al. 2015)*. Various rounds of random mutagenesis have led to the selection of a strain capable of producing higher quantities of penicillins, notably at the expense of other specialised metabolites production expressed by *P. rubens* NRRL 1951 wild-type strain *(García-Estrada et al. 2020)*. Proteome analysis *(Jami et al. 2010)* and the study of these mutations *(Salo et al. 2015)* revealed lower expression levels for some genes related to specialised metabolism and the formation of non-functional proteins. The proposed **GSMN** expresses the potential for metabolite production carried by the genome and thus provides a platform for knowledge that can be enhanced from such data. However, it should be noted while the presence of a compound in the model offers a valuable starting point for further investigation, it does not necessarily guarantee its production within the model.

According to the classification of metabolites established by the MetaCyc ontologies, 13% of ***i*Prub22** compounds would be associated with specialised metabolism. However, it is not uncommon to find metabolisms building blocks relative to constitutive metabolisms such as geranyl diphosphate or myoinositol in the specialised metabolites lists. This fact illustrates that the specialised metabolites concept, as understood by biologists or chemists, diverges from what is implemented in the databases. Thus, based on the information provided in the literature and the results of genome mining, the percentage of specialised metabolites contained in ***i*Prub22** should be further minimised. It raises the question of what constitutes a specialised metabolite for these communities and highlights the constant need for annotation standardisation.

Fungal specialised metabolites' particularity lies in their biosynthetic genes being co-located and forming sets called "Biosynthetic Genes clusters" (BGCs). Today such clusters are commonly identifiable *via* genome mining, but many of their products remain unknown. The review by Iacovelli *et al.* *(2021)* reports 33 essential biosynthetic genes encoding 10 non-ribosomal peptides synthetases (NRPSs), 20 polyketide synthetases (PKSs), 2 hybrid NRPS-PKSs, and 1 dimethyl-allyl-tryptophan synthetase (DMATS). Non-ribosomal peptides identified include fungisporin, roquefortine C, D, F, M, and N, as well as associated products such as meleagrin, glandicolins A and B, histidyltryptophanyldi-ketopiperazine (HTD) and dehydrohistidyltryptophanyldiketopiperazine (DHTD). This list of NRPSs is also enriched by the following three siderophores, coprogen, ferrichrome and fusarinin family compounds (*i.e.,* being classified by their function instead of their biosynthetic pathways, here NRPSs). Among the polyketides identified are compounds belonging to the sorbicillinoid family and chrysogin. All these metabolites are well-documented, studied and commonly accepted in the literature *(Van den Berg et al. 2008; Houbraken et al. 2012; Salo et al. 2015; Nielsen et al. 2017; Guzmán-Chávez et al. 2018)*.

However, the identification of BGCs and their end products slightly differed depending on whether fungiSMASH or antiSMASH was used (Table 5* in the S1 file). Thus, three BGCs coding for the following five compounds were found exclusively with fungiSMASH: naphthopyrone, chrysoxanthones A, B and C and depudecin. Concerning the antiSMASH results, the BGC related to penicillin synthesis emerged more exhaustively since isopenicillin N, phenoxymethylpenicillin (penicillin V) and the precursor *δ*-(L-*α*-aminoadipyl)-L-cysteine-D-valine (ACV) were identified. In addition, BGCs leading to the synthesis of PR-toxin, neurosporin A, ACT-Toxin II, melanin, NG-391, aspercryptins and chaetoglobosins are also raised only with antiSMASH results. It is also worth mentioning that the similarity percentages (*i.e.,* the confidence that can be placed on the actual presence of the BGC in the strain under consideration) can also be different from one version to another, as illustrated by the sorbicillin-encoding cluster.

* Tableau 3-13, page 313 et Tableau 3-14 page 314

Of the 42 BGCs identified using the fungiSMASH tool suite, 29% have yielded at least one distinct compound. However, only three of these BGC products (*i.e.,* involved in the biosynthesis of penicillins, roquefortins, and initial steps of patulin) have been integrated into **GSMN**. Collecting and processing the information on specialised metabolites biosynthesis raises several questions. Firstly, are the metabolic pathways leading to the biosynthesis of those NPs comprehensively elucidated and documented in the literature? Furthermore, if so, are these pathways indexed in the relevant databases such as MetaCyc? For instance, CobraMod *(Camborda et al. 2022)* is a pathway-centric curation tool designed to enable the modification and extension of **GSMN**. This tool extracts metabolic information from BiGG, KEGG, and BioCyc databases. However, depending on the existence of biosynthetic pathways for the selected metabolites and their degree of completeness, what measures are envisaged to enhance the modelling accuracy? This point leads to the two lists creation, one containing the metabolites for which an identifier exists in MetaCyc, and the other composed of orphan metabolites for which, among other things, the closest known antecedents (*i.e.,* precursors) will have to be determined. Subsequently, if the metabolic pathways are fully present, curation consists of verifying the actual presence or absence of the reactions in the draft and, if the information in the literature provides it, checking the corresponding GPR associations. The accuracy of the GPR associations is pivotal for the biological and computational interpretation of the *in silico* knock-outs. Thus, to conform with the literature data, manual curation was applied to ***i*Prub22** for penicillin, meleagrin, roquefortine and the last known precursor of patulin produced *(Nielsen et al. 2017)*. On the other hand, if the metabolic pathways are semi-complete, like for yanuthone D *(Holm et al. 2014)* or absent, like for chrysogine *(Viggiano et al. 2018)*, the analyses will focus on the closest known precursors, and further work will aim to find the most accurate method for integrating and generating the missing information. These gaps could potentially be resolved by adding modules or a single reaction, balanced and supported by all the genes involved in the biosynthesis of the compounds of interest. Recently an automatic pipeline called BiGMeC (Biosynthetic Gene cluster Metabolic pathway Construction) has been developed to automatically reconstruct metabolic pathways associated

with BGCs, specifically targeting PKS and NRPS *(Sulheim et al. 2021)*. Based on the analysis of antiSMASH outputs, the resulting enzymatic reactions take into account redox cofactors and energy demand. The reactions determined in this way are compatible with the BiGG database, promising a facilitated integration of the biosynthesis pathways encoded in the well-characterised or not BGCs into the **GSMNs**. Finally, in the case of orphan metabolites, it will be necessary to create both the metabolites and the reactions underpinned by genomic sequences. To maintain optimal traceability, the addition of these entities will be supported by as much metadata as possible such as bibliographic references for the constitution of biosynthetic pathways, InChIKey, InChI, PubChem identifiers, molecular weights, charges and SMILES until their incorporation into public databases. In addition to the biological relevance and the creation of a model as close to reality as possible, the interest in precisely recreating these pathways lies in the possibility of subsequently analysing in greater detail the distribution and flux reallocation from the precursors of all the targeted metabolites.

The last point concerns the interoperability between specialised metabolic databases and reconstruction databases. The LOTUS database *(Rutz et al. 2022)* (*i.e.,* database of specialised metabolites classified by taxonomy) is queried to target specialised metabolism more specifically. Of the 240 hits found for *P. rubens*, only 47 compounds are potentially usable (*i.e.,* metabolites mapped based on their InChIKey versus MetaCyc). This low proportion of integrable data is even more marked if we focus on the Natural Product Atlas database *(Van Santen et al. 2019)*, a microbial NP resource containing, among others, referenced data for structure, compound names and source organisms. Still, based on the InChIKey string, 102 compounds were reported for *P. rubens*, but only 6-aminopenicillanate was present in MetaCyc (v23.0) as in the network structure. Extending the search to the genus *Penicillium*, 2,159 NPs are reported. However, only 20 of these compounds have an exact match on MetaCyc, and 44 molecules are raised, looking only at the first section of the InChIKey. Among them, 17 are present in ***i*Prub22**, including the following seven exact matches: citreoisocoumarin, griseophenone C, brevinamide F, verruculogen, paxilline, 6-aminopenicillanate and *β*-10-hydroxy-12-demethyl-11,12-dehydropaspaline. These

17 metabolites are involved in 6 consumption reactions, 8 production reactions and 1 reversible transport reaction. These 15 reactions were introduced in the **GSMN** from either previous networks (*i.e.,* 3 from *i*AL1006 and 6 from Prubens), gap-filling (*i.e.,* 4 reactions) or functional annotation (*i.e.,* 2 reactions). Once again, this reflects the limited compatibility and communication between databases for specific data integration since most of the specialised metabolism in the strict sense found in ***i*Prub22** comes from manual processes.

1.5 CONCLUSION

This study presented a revised version of the *P. rubens* Wisconsin 54-1255 **GSMN**, encompassing 5,919 reactions, 5,703 genes and 5,192 unique metabolites. The reconstruction process involved updating functional annotations, establishing orthology links with various organisms, and reconciling data from previous *P. rubens* **GSMNs**.

Developing a functional **GSMN** is a complex and time-consuming endeavour that necessitates iterative and cyclical steps. These steps are crucial for enhancing our comprehension of the organism being studied and ultimately improving the accuracy of the models. Indeed, the iterative and cyclical nature of the reconstruction process allows for continuous refinement and enhancement of the **GSMN**. As more data is incorporated and new insights are gained, the models become increasingly precise and representative of the organism's metabolic behaviour. This iterative approach ensures that the models are continuously improved and aligned with the available knowledge. Thus, we have placed significant emphasis on ensuring interoperability, accessibility, and transparency during the ***i*Prub22** generation. Consequently, ***i*Prub22** respects the actual convention's standards, and we believe that providing a model easily usable by the community will facilitate advancements in the field and encourage knowledge sharing.

The ***i*Prub22** model fulfils the fundamental criteria for metabolic modelling, effectively simulating the growth of *P. rubens* by considering variations in carbon and nitrogen, reflecting environmental changes and diverse nutrient sources encountered by the fungus. Moreover, the ***i*Prub22** reconstruction serves as a robust backbone for generating different models and

investigating the growth and production of specialised metabolites under diverse environmental conditions. It contributes to our understanding of *P. rubens*' metabolic capabilities and provides a valuable resource for future studies in the field of systems biology. Furthermore, we demonstrate that ***i*Prub22** exhibits a predictive capacity for specialised metabolite production on the reference medium (*i.e.,* glucose as C-source, ammonia as N-source).

Finally, a particular effort was made to explore the specialised metabolism integrated into this model. Although the generation of more data in recent years has led to an increasing size of networks, the percentage of specialised metabolites in the models remains unchanged. Therefore, the effort undertaken to update the reconstruction databases seems linear, but the proportion devoted to specialised metabolites warrants further investigation. Exploring the presence of specialised metabolism further highlights the lack of information within the databases and shared identifiers. Such problems need to be addressed before any use of **GSMN** at a large scale in NP studies.

1.6 MATERIALS AND METHODS

Reconstruction is performed following the steps recommended in the 2010 Reference Protocol *(Thiele et Palsson 2010a)*. The ***i*Prub22** reconstruction is provided in the **SBML** community standard format *(Keating et al. 2020)* with a quality score of 74% evaluated by MEMOTE (v0.13.0) *(Lieven et al. 2020)*. To ensure transparency and reproducibility, the reconstruction process and related documents are available in Supporting Information. As a final step, the model was registered in BioModels *(Malik-Sheriff et al. 2020)* and given the identifier MODEL2306150001.

1.6.1 The primary stages: generating the draft

1.6.1.1 Data origin

The genomic data for *P. rubens* Wisconsin 54-1255 (*i.e.,* 12,557 protein and gene sequences), as well as the General Feature Format file associated, were downloaded from the Ensembl Fungi browser under accession number GCA_000226395 *(Van den Berg et al. 2008)*.

1.6.1.2 Functional annotation

Penicillium rubens functional annotation is performed using the Trinotate pipeline (v3.2.1) *(Bryant et al. 2017)* for its boilerplate SQLite database. Thus, homology analyses are launched with blastx and blastp (v2.5.0) with an *e*-value cut-off of 0.001 against the UniProt database (download on 27 June 2020). The best hit of a sequence is considered its annotation. Search for protein domain is carried out by HMMER (v3.3) *(Potter et al. 2018)* against the PFAM database (download on 27 June 2020) with an *e*-value cut-off of 0.001 and a threefold in the per-domain output of 0.01. SignalP (v4.1g) *(Petersen et al. 2011)* and TMHMM (v2.0c) *(Krogh et al. 2001)* are used to search for peptide signals and transmembrane domains, respectively, with their default settings. Protein subcellular localization is determined using DeepLoc (v1.0) *(Almagro Armenteros et al. 2017)*.

Data that are mandatory for the **GSMN** reconstruction, Gene Ontology Terms (GOTs) *(Ashburner et al. 2000; The Gene Ontology Consortium 2021)* and EC numbers are captured respectively using an internal Trinotate script and by querying the KEGG database *(Kanehisa et al. 2021)* using its own orthology identifiers (KO). The resulting KO list is enriched by a *P. rubens* proteome analysis performed on the KAAS webserver (Automatic Annotation Server v2.1) *(Moriya et al. 2007)* against a dataset composed of 367,616 genomic sequences exclusively selected from 37 fungi kingdoms (Best Bidirectional Hit-method and other settings by default – July 2020) and by the exploitation of pre-existing *P. rubens* data *(Van den Berg et al. 2008)* on KEGG (Genome information accession – T01091). EC-numbers are then retrieved from these three selected KO sets using KEGG-API REST (REpresentational State Transfer) application programming interface in July 2020.

Annotations combined with the information contained in the *P. rubens* General Feature Format are aggregated to generate a GenBank file using the script: https://github.com/ArnaudBelcour/gbk_from_gff. This file was then used as input to the PathoLogic software from the Pathway Tools suite (v23.0 default settings) *(Karp et al. 2021)*. This qualitative metabolic model reconstruction module allows the inference of the reactome from the annotated genome. The database containing the information from the annotation was then exported in attribute-value flat files, which were necessary for further analysis in the AuReMe workspace *(Aite et al. 2018)* (see *1.6.1.5. Data consensus and integration* for further details).

1.6.1.3 Orthology completion

Using OrthoFinder (v2.3.12) *(Emms et Kelly 2019)* and Blastp (v.2.5.0), Reciprocal Best Normalized Hit (RBNH) strategy was applied to identify functionally identical genes. Blastp was parametrised with an *e*-value of 0.001 and performed against seven selected templates: *Arabidopsis thaliana* *(de Oliveira Dal'Molin et al. 2010)*, *Aspergillus nidulans* *(Pitkänen et al. 2014; Castillo et al. 2016)*, *Aspergillus niger* *(Pitkänen et al. 2014; Castillo et al. 2016)*, *Neurospora crassa* *(Dreyfuss et al. 2013)*, *Penicillium* species complex *(Pitkänen et al. 2014; Castillo et al. 2016)*, *Saccharina japonica* *(Nègre et al. 2019)* and *Schizosaccharomyces pombe* *(Pitkänen et al. 2014; Castillo et al. 2016)*. Upstream, data from the different selected templates were filtered to keep only the sequences present in the associated **GSMN** models.

1.6.1.4 External sources

The reconstruction is enriched and completed by external data sources. First, data from the Pathway/Genome Database Concepts PchCyc (PGDB belonging to Tier 2) available on BioCyc are searched from the special SmartTables *(Karp et al. 2019)* provided. Second, compliant information present in the first *(Agren et al. 2013)* and latest *(Prigent et al. 2018)* versions of the network are extracted and parsed to be added to the draft. Only those reactions supported by genomic information that either has a MetaCyc compatible reaction ID or reactions for which all reactants and products have a MetaCyc ID (in which case the original reaction ID is retained) are incorporated into the draft.

1.6.1.5 Data consensus and integration

AuReMe (AUtomatic REconstruction of MEtabolic models – v2.4) dedicated to **GSMN** reconstruction *(Aite et al. 2018)* was used to reconstruct the *P. rubens* **GSMN**. This ToolBox encapsulates the various programs needed to create a high-quality network and maintain a consistent steps record. Results obtained from PathwaysTools and OrthoFinder are injected into AuReMe to generate the intermediate subnetworks resulting from the functional annotation and the orthology search, respectively. Reconstruction was based on the data present in MetaCyc (v23.0) *(Caspi et al. 2020)* As data templates for orthology refer to other databases like KEGG *(Kanehisa et al. 2021)* or BiGG *(Norsigian et al. 2020)*, a mapping operation using MetaNetX dictionary *(Moretti et al. 2021)* (*i.e.,* intrinsic to AuReMe MNXref version 2018/09/14) was performed to obtain identifiers compliant with MetaCyc. These data are then merged and completed with the information from external sources. Gene annotations are evaluated and compared using the FungiFun *(Priebe et al. 2015)* web server (v2.2.8 Beta). The resulting draft is then analysed both qualitatively using topological analysis and quantitatively using constraint-based analysis, refined by manual curation and enriched with data extracted from the literature (See the following section).

The majority of the Figures are drawn using R (v3.5.1) and the following packages: ggplot2 (v3.3.5) *(Wickham 2016)*, hrbrthemes (v0.8.0), ggpubr (v0.4.0) for visualisation and plyr (v1.8.6), dplyr (v1.0.7) and forecast (v0.5.1) for data formatting. Comparisons are made using the Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn) for the simplest cases and upsetplot with UpSetR (v1.4.0) *(Conway et al. 2017)* for the most complex cases. For the visualisation of the network with KEGG maps, KEGG mapper was used *(Kanehisa et Sato 2020)*.

1.6.2 The final stages: from draft to high GSMN quality

1.6.2.1 Potential producibility – topological analyses and manual curation

The reconstruction is refined by examining the presence of known compounds (*i.e.,* called targets) in *P. rubens*. The metabolites list to be verified is first established according to data described in the literature, molecules isolated in the laboratory and results identified by antiSMASH/fungiSMASH *(Blin et al. 2019)*. This list is enriched by metabolites present in the LOTUS database (https://lotus.naturalproducts.net) *(Rutz et al. 2022)* to focus more precisely on the specialised metabolism. Lastly, the metabolites present in the draft are filtered according to the annotations containing the term “secondary” in the MetaCyc ontologies.

Then, the potential topological producibility tests were carried out using the Mene Tools suite (v3.2.0) *(Belcour et al. 2020)*. It was necessary to define a list of initiating compounds to perform topological analyses (*i.e.,* starting point called a seed). All metabolites found in the extracellular compartment were considered as seeds to model the nutritional environment impact (*i.e.,* system boundaries). Although detailed intracellular compartmentation was not modelled within ***i*Prub22**, particular attention was given to the exchange between intracellular and extracellular environments due to the nutritional regime of the fungus (*i.e.,* extracellular absorption). The added transport reactions were mainly the result of the information concatenation from *i*AL1006, PchCyc, and some literature. GPR associations were checked and cleaned up according to the annotation provided by DeepLoc *(Almagro Armenteros et al. 2017)*. The subcellular assignment of each gene product is available within the **SBML** in the notes tag of the gene entities. For modelling purposes, artificial exchange reactions (*i.e.,* uptake for the input of a compound and production for the output) were added to the reconstruction. They were defined according to the metabolites present in the extracellular medium, which possessed a transport reaction towards the intracellular compartment. The seeds, therefore, corresponded to all the metabolites that have an uptake reaction (S4 file). Artificial genes of form t001 to t208 (*i.e.,* transport), d001 to d037 (*i.e.,* demand), u001 to u185 (*i.e.,* uptake), sk001 (*i.e.,* sink) and p001 to p042 (*i.e.,* production) were assigned for each of them to discretise specific reactions more efficiently than those coming from a simple gap-filling.

Network gap-filling was performed using the MENECO tool (v2.0.0) *(Prigent et al. 2017)* This topological gap-filling tool seeks to make accessible from a given list of seeds a set of targets. As the gap-filling process is based on heuristics, it was decided to use filtered repair networks corresponding to MetaCyc subsets. They were built and queried successively, independently, and then jointly until the entire MetaCyc database (v23.0) was used. The MetaCyc subsets were divided into two main categories. On the one hand, the subset was composed of reactions reported in fungal reconstructions (*i.e.,* using the following 3 PGDBs available on MetaCyc – *Candida albicans* SC5314, *Exophiala dermatitidis* NIH/UT8656 and *Saccharomyces cerevisiae* S288c – and a fungi meta-network *(Belcour et al. 2022)*). On the other hand, the subset was composed of reactions for which no genomic data is associated (*i.e.,* spontaneous reactions or unknown enzymes). Gap-filling processes are performed iteratively from the targets with the highest confidence on the most relevant repair networks. The resulting reactions are manually verified and incorporated into the network (reactions and their sources are available in the S5 file).

For instance, this process was used to enrich the reconstruction with spontaneous reactions specifically. As Palsson *et al. (2010a)* recommended, only reactions for which at least one of the entities (*i.e.,* reactant or product) was already present in the draft were added to the reconstruction. As it was done for previous specific reactions, an artificial gene of the form s001 to s086 is assigned for each of them.

Objects such as metabolites or reactions are stored, listed, filtered and sorted using SmartTables available on MetaCyc *(Karp et al. 2019)*. Reconstruction curation and debugging are manually assisted by the network visualisation *via* ModelExplorer (v2.1) *(Martyushenko et Almaas 2019)*.

1.6.2.2 From reconstruction to models

The transformation of the reconstruction into a usable model was carried out using MATLAB (v2018b) and diverse functions of the COBRA Toolbox (v3.0) *(Heirendt et al. 2019)*.

As with reconstruction, model refinement requires several iterative correction and adjustment processes to optimise the model's performance. Moreover, these crucial steps improve model predictions' alignment with experimental observations, increasing its fidelity to *P. rubens* biological system.

The first point we addressed concerned the reversibility of reactions. To ensure accuracy and reliability, we meticulously updated the reversibility of reactions following the evolving information and reactions with indeterminate directionality were appropriately blocked. The second improvement concerns the naming redundancy issues related to compound class encompassing multiple specific metabolites. For example, if a compound class like D-galactopyranose includes ALPHA-D-GALACTOSE and GALACTOSE, and there is an interconversion reaction between their alpha and beta forms, we prioritised the use of α-D-GALACTOSE to replace D-galactopyranose. Then, we focused on reaction redundancy related to identifier inconsistency (*e.g.,* a reaction with MetaCyc identifier and a redundant reaction with “homemade” identifier). In such a case, we retained the more balanced reaction. Additionally, we ensured that the GPRs were identical or that all the genes from one reaction were included in the other. Discarded reactions were thus closed. The third aspect covered imbalanced reactions. The `checkMassChargeBalance()` function was employed to check mass and charge balance for every reaction. As a result, a large majority of reactions detected to be unbalanced were blocked.

Consequently, to guarantee that the simulations run as expected, we have established three models: the default, the open and the closed model. The default model represents the minimum uptakes necessary for biomass production (details provided in the next section). The open model allows for all uptakes with an upper bound of 10 $mmol.gDW^{-1}.h^{-1}$, while the closed model restricts all uptakes, disallowing any external metabolites from entering the system.

Complementarily, `cycleFreeFlux()` and `findMassLeaksAndSiphons()` functions were used on the closed model. Theoretically, without nutrient input, the results of these functions must be null or at least minimal. The first function allows the identification of thermodynamically infeasible cycles within the metabolic network to ensure a more realistic representation of cellular metabolism. The second function identifies any potential mass leaks (*i.e.,* molecular species can be generated from nothing) or siphons (*i.e.,* molecular species consumed without yielding anything) in the model. Furthermore, based on the work of Fritzemeier *et al.* *(2017)* and their recommendation regarding energy generating cycles, we verified by adding an energy-dissipating reaction for 13 compounds (*i.e.,* ATP, CTP, GTP, UTP, ITP, NADH, NADPH, $FADH_2$, $FMNH_2$, Ubiquinone-8, acetyl-CoA, glutamate, proton) that no flux was generated for these reactions when the model is closed.

Lastly, after addressing all these aspects, we proceeded to evaluate the predictive performance of ***i*Prub22** using analyses related to constraint-based modelling, namely FBA and FVA. Additionally, Fluxer *(Hari et Lobo 2020)*, a web application combining FBA and graph theory to offer an interactive graph visualisation, was also used to help us in the flux model refinement. These analyses allowed us to assess the metabolic capabilities of the model, predict optimal flux distributions under different conditions, and explore the range of possible flux values for each reaction.

1.6.2.3 Model functional capabilities: biomass and specialised metabolites production

The 2013 network, *i*AL1006, proposes a biomass reaction adapted from the organism's behaviour *(Agren et al. 2013)*. Thus, ten subsystems that model the production of amino acids, cell wall components, cofactors, DNA, RNA, fatty acids, phospholipids, glycerides, sterol esters, and some generic lipids define the biomass reaction (file S5). The complete adaptation of this physiologically relevant objective function to the data on MetaCyc was made impossible due to the lack of clear mapping between the identifiers and the use of highly generic terms. To test the validation of our proposed model, the biomass reaction was identical to that present in the 2018 network, Prubens *(Prigent et al. 2018)*.

Then, for biomass production analyses, ***i*Prub22** was subjected to growth simulations on different media compositions (S6 file), including modifications of the carbon source (35 simulations), nitrogen source (35 simulations), and the use of amino acids as combined carbon and nitrogen sources (21 simulations). Considering that the C-source required for respiration is significantly higher than the N-source, C-source was fixed at 15 $mmol.gDW^{-1}.h^{-1}$, while the N-source was set at 5 $mmol.gDW^{-1}.h^{-1}$. When simulating the utilisation of amino acids as both carbon and nitrogen sources, their uptake bounds were also set to 15 $mmol.gDW^{-1}.h^{-1}$. The uptakes of thiamine (Uptake_171), ferrous ion (Uptake_062), sulphur (Uptake_169), riboflavin (Uptake_157), and phosphate (Uptake_146) were set to a value of 10, while the uptake of oxygen (Uptake_136) remained unrestricted. The defined constraints allowed us to systematically evaluate and analyse the models' behaviour and metabolic capabilities across diverse growth conditions.

For specialised metabolite production analyses, FVA was performed with the biomass production maximised arbitrarily at 80% and reference constraints were applied (*i.e.,* glucose as C-source, ammonia as N-source, and uptake openings at the same rates mentioned earlier). Given that the biosynthesis pathways of many specialised metabolites involve a flavoprotein containing both FMN and FAD, the corresponding uptake was open (Red-NADPH-Hemoprotein-Reductases - Uptake_155). It was necessary to unblock the decanoate (Uptake_044) and phenoxyacetate (Uptake_144) uptakes for penicillin K and V production, respectively. These 3 uptakes were also opened arbitrarily at 10 $mmol.gDW^{-1}.h^{-1}$.

Finally, in adherence to the MIASE (Minimum Information About a Simulation Experiment) recommendations *(Waltemath et al. 2011)* to maintain a record of the modifications made to the reconstruction and to ensure comprehensive documentation of the model modifications, the details of the changes operated on the model were organised in a LiveScripts file from MATLAB (S7 file*) and the simulation results (S6 file) are available in the Supporting Information.

* Annexe 4, page 451

1.6.2.4 SBML format

The network resulting from AuReMe was modified following the current writing conventions *(Keating et al. 2020; Community standards to facilitate development and address challenges in metabolic modeling 2020)* to ensure the sustainability and viability of the model. To comply with these requirements (**SBML** Level 3, FBC2 *(Hucka et al. 2018; Olivier et Bergmann 2018)*), a format update was performed by relying on the official recommendations available on the COMBINE resource (https://co.mbine.org – community for the coordination of modelling standards in biology). The interoperability and enrichment of the identifiers of the various entities of the model were achieved using MetaNetX (v4.1) *(Moretti et al. 2021)*. These annotations were encoded within **SBML** according to the standards of the MIRIAM resource *(Novère et al. 2005)* (Minimum Information Requirements in Biochemical Model Annotation). Each entity of the network was also associated with an SBO term *(Courtot et al. 2011; Bernasconi et Masseroli 2019)* (Systems Biology Ontology – a nested classification scheme for grouping model components). The conversion to FBC2 *(Olivier et Bergmann 2018)* was performed *via* the script available on the LibSBML application programming interface *(Bornstein et al. 2008)*. These format modifications were then checked and validated *via* the scripts made available to the community on this same application (SBML Validator – testing the syntax and internal consistency of an **SBML** file). Finally, the model is tested by MEMOTE (MEtabolic MOdel Tests – v0.13.0) *(Lieven et al. 2020)* and its report is available in Supporting Information (S8 file*).

To conclude, as we have decided not to remove any information from the reconstruction, the model that we propose on BioModels (MODEL2306150001) with its FROG report corresponds to a parametrisation of the reconstruction where potentially incorrect reactions are closed and where the minimal uptakes that ensure both the production of biomass and specialised metabolites are open.

*Annexe 5, page 479

1.7 ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of the French National Research Agency (ANR), project number ANR-18-CE43-0013. We gratefully thank Erwan Corre (Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Roscoff, France) and David Touboul (Université Paris-Saclay, CNRS, Institut de Chimie Des Substances Naturelles, UPR 2301, Gif-Sur-Yvette, France) for their valuable advice. We also gratefully acknowledge Jeanne Got and Anne Siegel from the Institute for Research in IT and Random Systems (IRISA) for their helpful advice and for generously sharing their data. Furthermore, we express our gratitude to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), which is part of the Institut Français de Bio-informatique (ANR-11-INBS-0013) and BioGenouest network, for providing computing and storage resources.

1.8 SUPPORTING INFORMATION

S1 file. List of additional tables and figures. Table 1: Functional annotation details; **Table 2:** Features and results of the seven templates used for the reconstruction of the intermediate subnetwork from the orthology search; **Table 3:** List of reactions involved in the synthesis of roquefortines and meleagrin added to the reconstruction (MetaCyc identifiers); **Table 4:** Lists of reactions involved in penicillins biosynthesis; **Table 5:** Identification comparison of BGCs between antiSMASH and fungiSMASH; **Figure 1:** Overview of *Penicillium rubens* functional annotation; **Figure 2:** Overview of *Penicillium rubens* reactions performed with KEGG Mapper; **Figure 3:** Bipartite graph representing the seven templates networks' topology for the orthology subnetwork generation; **Figure 4:** Sequences number distribution per orthogroups according to species detected with OrthoFinder; **Figure 5:** Visualisation of orthologous genes detected by OrthoFinder and inference of their reactions to the orthology subnetwork; **Figure 6:** Sankey plot of the data selection from external sources; **Figure 7:** Scatter plot of reconstruction sources complementarity; **Figure 8:** Classification of genes and reactions according to their source of integration in the

draft; **Figure 9:** Classification of genes and reactions according to their associated Enzyme Commission number; **Figure 10:** Origin of metabolites in ***i*Prub22** according to reaction reconstruction sources; **Figure 11:** Sources of transport and exchange reactions added to the reconstruction; **Figure 12:** Distribution of the 510 reactions added to the reconstruction during the gap-filling steps; **Figure 13:** Annotations enrichment of the 3,771 genes added to the *Penicillium rubens* **GSMN** reconstruction (Results from FungiFun). (PDF)

Les éléments non mentionnés ci-dessous sont accessibles via https://doi.org/10.1371/journal.pone.0289757.

Table S5 :	Tableau 3-13, page 313 et Tableau 3-14, page 314
Figure S1 :	Annexe 1A, page 426
Figure S2 :	Annexe 1B, page 429 et Annexe 2C, page 438
Figure S3 :	Annexe 2B, page 436
Figure S5 :	Annexe 2D, page 440 et Annexe 2F, page 442
Figure S6 :	Figure 3-10, page 205
Figure S7 :	Annexe 3C, page 449
Figure S8 :	Annexe 1D, page 431, Annexe 2F, page 442 et Annexe 3B, page 447
Figure S11 :	Figure 3-23, page 266
Figure S12 :	Figure 3-18, page 253

S2 file. Orthology results. This workbook presents an overview of the OrthoFinder results, including selected protein sequences, the number of orthologous sequences, orthogroups between templates, and species-specific statistics. (XLSX)

S3 file. Metabolites belonging to *Penicillium rubens*. This workbook contains a traceability record of the three target lists used for reconstruction refinement, supplemented by a list of orphan metabolites. (XLSX)

S4 file. List of transport and exchange reactions added to the reconstruction. This workbook comprises a comprehensive list of transport and exchange reactions incorporated into the model. It includes transport reactions initially present in the draft, those added with or without gene support, and the necessary exchange reactions for model simulation (Uptake, Demand, Sink, and Production). (XLSX)

S5 file. List of reactions from external sources and gap-filling used for *i*Prub22 reconstruction. This workbook contains reactions from external sources (with or without MetaCyc identifier), spontaneous reactions, the four sets of reactions from gap-filling, and the three reactions required for biomass synthesis (*i.e.,* biomass formulation, transport, and exchange). (XLSX)

S6 file. List of modified reaction bounds and model simulations. This workbook includes information on reaction bounds adjusted during the model generation and presents simulations of *P. rubens* growth on different media using FBA with diversified constraints. It also explores the potential production of specialised metabolites using FVA on the reference medium. (XLSX)

S7 file. Features of reconstruction and model. This Livescript MATLAB showcases the characteristics of the reconstruction process and outlines the modifications required for generating the model. It provides a detailed analysis of the reconstruction and model development. (ZIP)

Un extrait de ce document est disponible en **ANNEXE 4 : LiveScript MATLAB détaillant les caractéristiques d'*i*Prub22**, page 451.

S8 file. MEMOTE report of *i*Prub22 reconstruction. This file is an informative report that details the MEMOTE test suite results conducted on the ***i*Prub22** reconstruction. (HTML)

Rapport présenté en **ANNEXE 5 : Rapport MEMOTE**, page 479.

2 - Des améliorations apportées mais une reconstruction et un modèle perfectibles

Dans le contexte de la reconstruction d'***i*Prub22**, cette section expose les choix méthodologiques et leurs considérations sous-jacentes, visant à fournir une ressource globale garantissant une utilisation et une prise en main rapides de notre réseau. Ainsi, nous commencerons par un focus lié aux principes d'accessibilité et d'interopérabilité. À cet effet, **(1)** nous présenterons une description détaillée des données utilisées, **(2)** nous nous interrogerons sur la pertinence liée à la réconciliation de données externes, **(3)** nous fournirons une évaluation critique de la qualité de la reconstruction selon MEMOTE, puis nous aborderons la nécessité et la mise en œuvre d'un format interopérable et exportable. **(4)** Nous examinerons ensuite le processus de *gap-filling*, une étape déterminante dans la qualité d'un **GSMN**. Nous exposerons ainsi la mise en œuvre d'un *gap-filling* raisonné, entrepris pour pondérer et classer les réactions ajoutées à la reconstruction. Nous constaterons toutefois, que, malgré des améliorations significatives dans la connectivité de la reconstruction, des défis subsistent, notamment concernant les réactions bloquées et les associations GPR. **(5)** Nous explorerons les diverses options de modélisation de médias que nous proposons pour une représentation plus fine des conditions environnementales. **(6)** Enfin, nous reviendrons alors en détail sur la production de biomasse, un indicateur clé de la fonctionnalité d'un modèle. À ce titre, nous nous attarderons sur sa conception et nous apporterons une réflexion sur ses limites.

2.1. Transparence, réconcillation, accessibilité et interopérabilité des données

Les principes FAIR, acronyme de « *Findable*, *Accessible*, *Interoperable, and Reusable* », constituent un ensemble de règles qui visent à améliorer la gestion et l'utilisation des données scientifiques. Leur première mention formelle remonte à 2016 et émane des travaux de Wilkinson *et al.* *(2016)*, intitulés « *The FAIR Guiding Principles for scientific data management and stewardship* ». Cet article, cité plus de 10 275 fois (*i.e.* moyenne du nombre de citations sur Google scholar, Scopus, Semantic scholar et ResearchGate), décrit les principes FAIR comme étant axés sur l'amélioration de la capacité des machines à localiser et à utiliser automatiquement les données, tout en soutenant leur réutilisation par les individus. Les recommandations FAIR, bien qu'intrinsèquement liées, sont séparables, et chaque point peut être traité indépendamment (cf. encart : *« The FAIR Guiding Principles » tels que définis dans les travaux de Wilkinson et al*, page 196). Depuis leur introduction, les principes FAIR ont progressivement modifié les pratiques de gestion et de partage des données, gagnant une adoption croissante dans toutes les disciplines scientifiques. Afin de soutenir ce dernier point, nous présentons en **Figure 3-7** une brève analyse bibliométrique, basée sur Scopus, des principes mentionnés.

Scopus

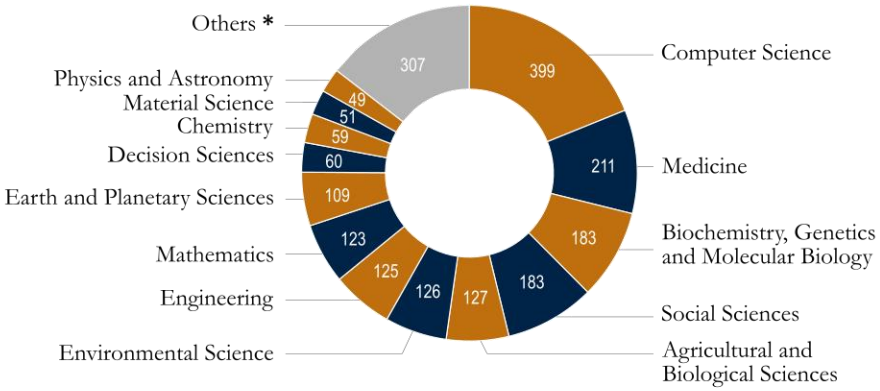
Query :
"reusable" AND "findable"
AND "interoperable"
AND "accessible"

Résultats: 1 059 documents publiés entre 2015 et 2024

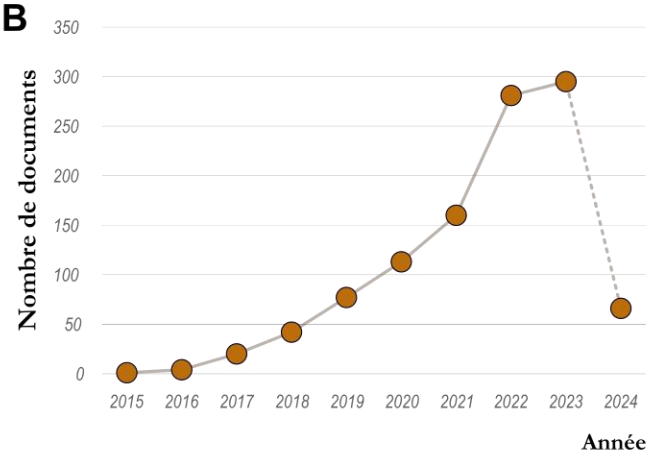
- 645 Articles
- 136 Reviews
- 21 Chapitres de livres
- 257 Autres

A

Nombre de documents recensés par domaine



B



*** Catégorie "Others"**

Chemical Engineering (42)
Health Professions (42)
Neuroscience (40)
Immunology and Microbiology (37)
Pharmacology, Toxicology and Pharmaceutics (34)
Arts and Humanities (31)
Multidisciplinary (28)
Energy (17)
Nursing (13)
Business, Management and Accounting (10)
Economics, Econometrics and Finance (4)
Dentistry (3)
Psychology (3)
Veterinary (3)

C

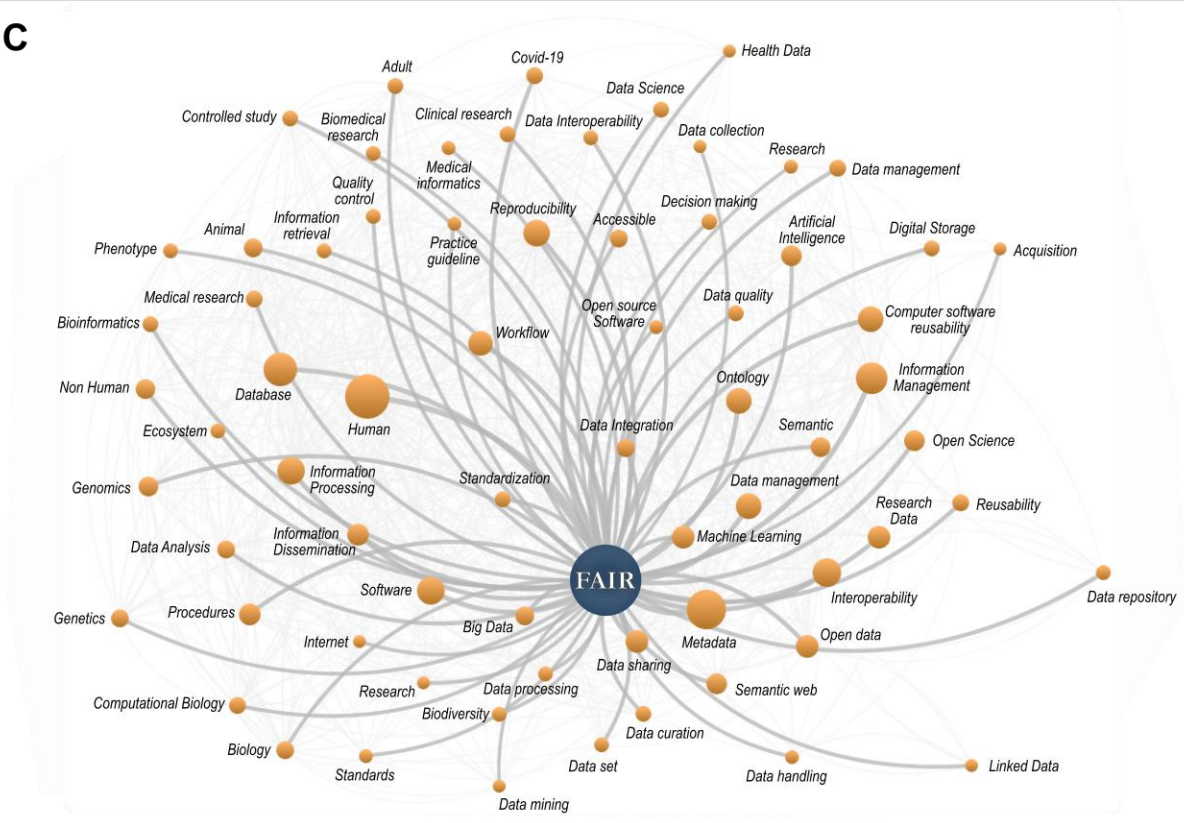


Figure 3-7 : Analyse bibliométrique des publications sur le principe FAIR : vue d'ensemble depuis Scopus. *Le corpus de documents étudiés, accompagné de ses métriques associées (i.e. panels A et B), provient de la base de données bibliographiques Scopus. Développé par Elsevier depuis 2004, Scopus est un outil de recherches multidisciplinaires qui indexe des articles de revues scientifiques, des conférences, des livres et des brevets dans divers domaines académiques. En recherchant tous les termes de l'acronyme FAIR, nous avons recensé 1 059 documents couvrant la période de 2015 à avril 2024. Comme les documents traités sont en langue anglaise, aucune traduction n'a été effectuée.* ***(A)*** *Graphique en forme de donut présentant le nombre de publications par domaine. La catégorie "others" regroupe toutes les catégories dont le nombre de documents représente moins de 2 % de l'ensemble du corpus. Nous observons une diversité étendue de domaines, incluant les sciences sociales, la médecine, les disciplines liées à la santé, la physique, la chimie, l'économie et l'informatique.* ***(B)*** *Graphique linéaire décrivant le nombre de documents publiés par an. Même si les notions sous-jacentes au concept FAIR ne sont pas nouvelles, la préoccupation de leur visibilité et de leur application sont relativement récentes et connaissent dès lors un intérêt croissant.* ***(C)*** *Carte de co-occurrences des mots clés, basée sur les métadonnées extraites des documents du corpus et générée avec VOSviewer (v 1.6.18). Les relations entre les termes sont établies en fonction du nombre de documents dans lesquels ils apparaissent (i.e. matrice de co-occurrence). Le corpus comprend un total de 7 234 mots clés indexés, attribués soit directement par les auteurs des publications, soit par Scopus. Après l'application d'un seuil minimal de 20 occurrences par mot-clé (i.e. doit apparaître dans au moins 20 documents différents) et leur nettoyage (e.g. fusion des termes similaires ou standardisation des écritures), 70 termes différents, matérialisés sur la carte par des nœuds pondérés en fonction de leur occurrence, ont été retenus. Sur ce réseau, nous avons mis en exergue les liens existant avec l'expression FAIR.*

Les quatre éléments constitutifs de l'acronyme FAIR se définissent de la manière suivante :

- ***Findable*** : implique que les données doivent être facilement localisables et que des métadonnées appropriées ainsi que des identifiants uniques doivent être utilisés.
- ***Accessible*** : garantit que les données doivent être accessibles tant techniquement que légalement, permettant à quiconque d'y accéder sans restriction excessive.
- ***Interoperable*** : exige que les données soient structurées de manière à être compatibles avec différents systèmes et plates-formes, facilitant ainsi leur intégration et leur analyse conjointe.
- ***Reusable*** : stipule que les données doivent être bien documentées et accompagnées d'une licence appropriée, permettant leur utilisation future par d'autres chercheurs, ce qui favorise la transparence, la reproductibilité et l'avancement de la recherche.

En prenant en considération et en abordant les questions liées à la gestion, à l'accessibilité et à la réutilisation des données en amont de tout projet, ces principes tendent à devenir un pilier de la recherche, concrétisant ainsi l'idéal d'une Science plus ouverte et participative. Cependant, bien qu'ils ouvrent la voie sur ces problématiques et fournissent une ressource puissante pour encourager le partage et la gestion des données, ils ne résolvent pas à eux seuls toutes les interrogations et de nouvelles questions émergent, telles que l'organisation du partage des données, le choix de conception des données, le respect des données sensibles ou la valorisation du partage des données *(Boeckhout et al. 2018)*. En conséquence, les principes FAIR bien que fondateurs, devront être enrichis par de nouvelles considérations pour promouvoir une recherche pleinement responsable *(Jacobsen et al. 2020)*.

Au cours de nos travaux, conscients des défis liés à la reproductibilité et à la transparence dans la modélisation des **GSMNs** *(Ebrahim et al. 2015; Baker 2016; Tiwari et al. 2021)*, nous avons particulièrement mis l'accent sur ces aspects. Notre objectif était de fournir une reconstruction et des modèles qui soient facilement utilisables par quiconque. Dans cette section, nous exposerons la manière dont nos recherches s'alignent avec les principes FAIR présentés précédemment, identifiant les points de convergence et de divergence. Le cas échéant, nous proposerons des axes d'amélioration visant à renforcer leur application et leur intégration.

« *The FAIR Guiding Principles* » tels que définis dans les travaux de Wilkinson *et al*

- *To be Findable*

F1. Les (méta)données sont dotées d'un identifiant unique et persistant à l'échelle mondiale.

F2. Les données sont décrites avec des métadonnées riches (cf. R1).

F3. Les métadonnées incluent clairement et explicitement l'identifiant des données qu'elles décrivent.

F4. Les (méta)données sont enregistrées ou indexées dans une ressource consultable.

- *To be Accessible*

A1. Les (méta)données sont récupérables par leur identifiant en utilisant un protocole de communication standardisé.

A1.1. Le protocole est ouvert, gratuit et universellement implémentable.

A1.2. Le protocole prévoit une procédure d'authentification et d'autorisation, si nécessaire.

A2. Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.

- *To be Interoperable*

I1. Les (méta)données utilisent un langage formel, partagé et largement applicable pour la représentation des connaissances.

I2. Les (méta)données utilisent des lexiques qui respectent les principes FAIR.

I3. Les (méta)données incluent des références qualifiées vers d'autres (méta)données.

- *To be Reusable*

R1. Les (méta)données sont richement décrites avec une pluralité d'attributs précis et pertinents.

R1.1. Les (méta)données sont publiées avec une licence d'utilisation claire et accessible.

R1.2. Les (méta)données sont associées à une provenance détaillée.

R1.3. Les (méta)données respectent les normes communautaires pertinentes pour le domaine.

2.1.1. DESCRIPTION DES DONNÉES UTILISÉES

En préambule de cette partie, il convient d'apporter un éclaircissement concernant la sélection des données utilisées pour la reconstruction d'***i*Prub22**. Il peut sembler trivial de rappeler que la sélection des données est la clé de voute de tout projet scientifique. Les résultats obtenus et la crédibilité qui leur est accordée dépendent indubitablement de la qualité des données initialement interrogées. En effet, le processus de nettoyage et de sélection des données est une étape commune à toute étude. Nous pouvons citer, à titre d'exemple, la nécessité lors du travail sur des organismes non modèles et nouvellement séquencés, de s'assurer de l'absence de séquences contaminantes. Cependant, comme il l'a été mentionné précédemment *P. rubens* Wisconsin 54-1255 est un organisme modèle largement étudié au cours des dernières décennies. Nous pourrions donc aisément supposer que, contrairement aux organismes non modèles, nous disposons de données propres, corrigées et facilement accessibles. Cependant, qu'en est-il factuellement ?

Nous avons choisi de collecter nos données à partir d'EnsemblFungi, une branche d'Ensembl Genome, la base de données de référence européenne. Toutefois, nous allons constater qu'il existe de légères différences entre ces données et celles hébergées dans d'autres bases de données. En conséquence, nous débutons par une exposition des données d'origine mentionnées dans la littérature, suivie de la présentation des données archivées sur NCBI, KEGG, EnsemblFungi, JGI MycoCosm, PchCyc, et UniProt. Les divergences constatées entre ces diverses sources sont représentées graphiquement dans la **Figure 3-8**.

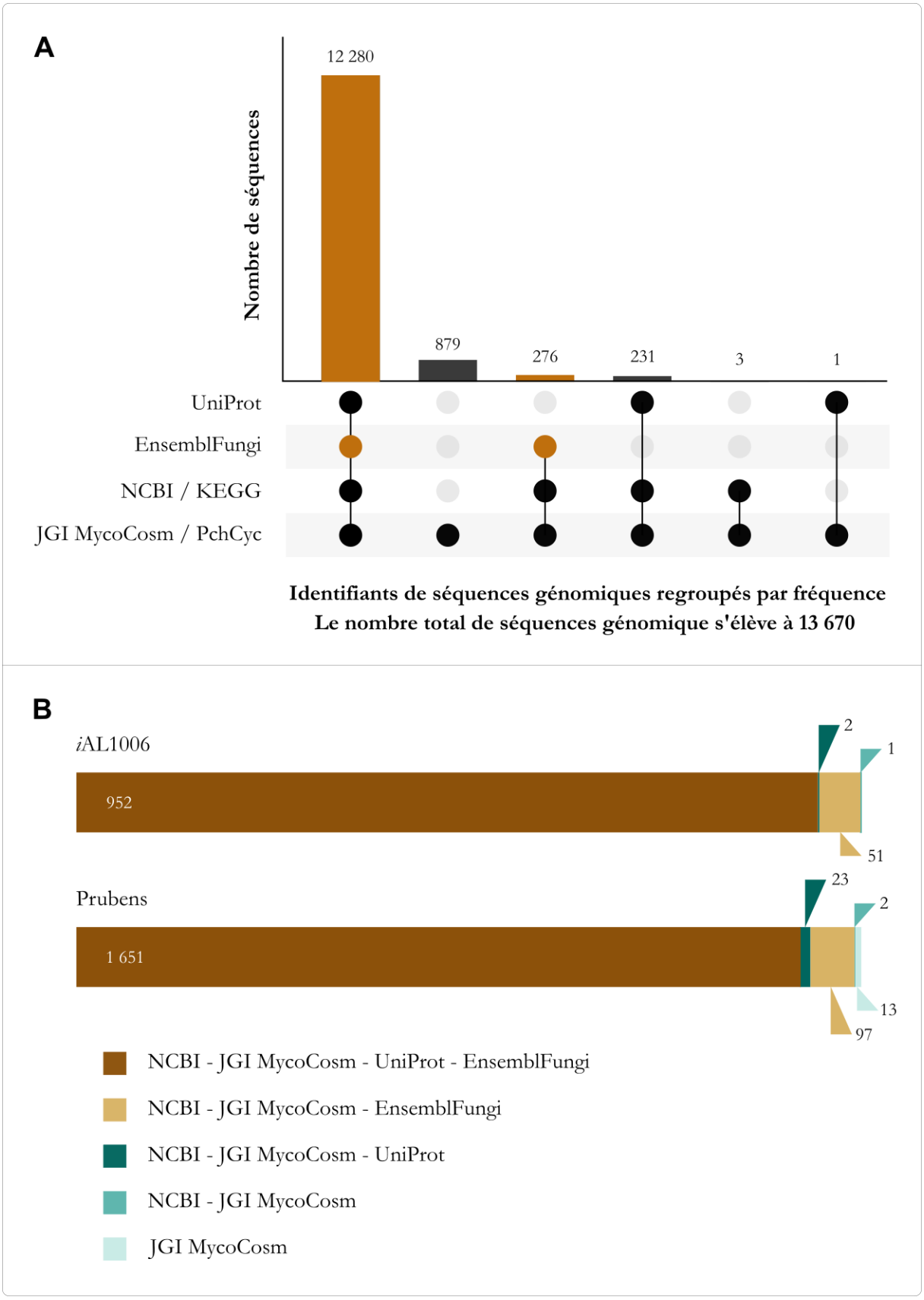


Figure 3-8 : Origine des données de Penicillium rubens Wisconsin 54-1255 à travers les différentes bases et banques de données. (A) *Diagramme d'ensembles retraçant l'origine des séquences de P. rubens Wisconsin 54-1255 disponibles sur Internet. Seules les séquences comportant le nom des gènes (i.e. Pc\d{2}g\d{5}) ont été comptabilisées. Les données sélectionnées, provenant de la base de données EnsemblFungi sont mises en évidence en orange (■).* ***(B)*** *Diagramme à barres empilées en pourcentage représentant l'origine des séquences génomiques présentes dans les deux* ***GSMNs*** *connus de P. rubens, iAL1006 (Agren et al. 2013) et Prubens (Prigent et al. 2018).*

Le séquençage du génome de *P. rubens* Wisconsin 54-1255 a été réalisé en 2008. Selon les travaux menés par Van den Berg *et al.* *(2007)*, qui servent de référence pour l'ensemble des données qui seront citées ultérieurement, le génome nucléaire de cette souche s'étend sur une longueur de 32,19 mégabases et est divisé en 49 *scaffolds*. L'annotation génomique, basée sur la détection de cadres de lecture ouverts (ORF) d'une taille minimale de 100 acides aminés, a révélé 13 653 ORFs parmi lesquels 592 sont identifiés comme pseudogènes, et 116 ORFs présentent des signes de troncature potentielle en raison de leur emplacement en bordure des contigs. Enfin 56,6 % du génome de *P. rubens* Wisconsin 54-1255, qui comprend 12 943 gènes, est constitué de séquences codant pour des protéines putatives.

Penicillium rubens Wisconsin 54-1255 possède sur NCBI l'identifiant taxonomique 500485. Les données associées à cet organisme qui sont hébergées sur GenBank (numéro d'accession : GCA_000226395.1) et RefSeq (numéro d'accession GCF_000226395.1) font état de 13 911 gènes dont 12 791 séquences codant pour des protéines avec 12 790 identifiants uniques (*i.e.* `Pc[0-9]{2}g[0-9]{5}`). À noter également que ces données sont accessibles *via* la base de données KEGG sous l'entrée T01091. Enfin, 25 758 séquences protéiques sont également stockées sur NCBI.

L'annotation disponible sur le portail EnsemblFungi, soumise à l'*International Nucleotide Sequence Database Collaboration* (INSDC), provient de l'annotation de l'assemblage de référence GCA_000226395.1. Cette annotation est également supplémentée par des gènes non codants provenant de la base de données Rfam, spécialisée dans les familles d'ARNs. Au total, 12 557 gènes codants et 496 gènes non codants y sont dénombrés. Toutefois, une seule séquence correspondant à un gène codant mais ne possédant pas l'identifiant de gène standard (cf. encart ci-dessous : *Dénomination des séquences génomiques de P. rubens Wisconsin 54-1255*) a été exclue des jeux de données que nous avons interrogés.

Dénomination des séquences génomiques de *P. rubens* Wisconsin 54-1255

Pour la réalisation du draft d'***i*Prub22** nous avons utilisé les deux fichiers fasta suivants ainsi que le fichier d'annotation de ces séquences au format gff3. Tous trois ont été téléchargés sur le portail EnsemblFungi :

- *Penicillium_rubens_wisconsin_54_1255_gca_000226395.PenChr_Nov2007.cds.all.fa*
  Fichier contenant les séquences codantes correspondant aux gènes Ensembl.
  Les CDS ne contiennent ni séquences introniques, ni séquences UTR.
- *Penicillium_rubens_wisconsin_54_1255_gca_000226395.PenChr_Nov2007.pep.all.fa*
  Fichier contenant les traductions des protéines des gènes Ensembl.
- *Penicillium_rubens_wisconsin_54_1255_gca_000226395.PenChr_Nov2007.57.gff3*
  Fichier d'annotations permettant de faire le lien entre les différents élements génomiques au moyen d'identifiant unique et d'identifiants parents.

Les identifiants de séquences que nous avons utilisés correpondent aux codes des gènes composés de 14 caractères alphanumériques. Les quatre premiers caractères, relatifs au *locus* tag des séquences, sont les lettres « PCH » en majuscules suivies d'un soulignement (*i.e. underscore* « _ »). Ces caractères sont suivis des lettres « Pc » pour *Penicillium chrysogenum*, puis, par deux chiffres correspondant au numéro du contig, et enfin, de la lettre « g » et de cinq chiffres représentant le numéro du gène. Pour simplifier la lecture, nous avons omis l'utilisation du préfixe « PCH_ » pour désigner les gènes.

Ces identifiants de séquences sont associés aux gènes mais ils peuvent également faire référence aux ORFs, aux CDS et par extension aux produits géniques, à savoir les protéines et plus particulièrement les enzymes.

Les données disponibles sur le portail web du JGI MycoCosm regroupent 13 671 gènes sous l'entrée ***P. chrysogenum*** Wisconsin 54-1255. Toutefois, alors que les données de NCBI et d'EnsemblFungi indiquent la présence d'un seul chromosome pour ***P. rubens*** Wisconsin 54-1255, il est fait état dans la description de l'organisme d'une taille de génome de 34.1 mégabases distribués sur quatre chromosomes. En outre, sur NCBI, la fiche de *P. chrysogenum*, dont l'identifiant taxonomique est 5 076, présente effectivement quatre chromosomes, mais une taille de génome de 32.41 mégabases pour 11 981 gènes codants répertoriés. En 2014, la fiche bioproject PRJNA246109 rapporte une amélioration du génome de *Penicillium chrysogenum* Wisconsin 54-1255 en réalisant un assemblage de 32.4 mégabases en quatre chromosomes. Rappelons ici que la proposition de reclassification de la souche Wisconsin 54-1255 date de 2012. Il est troublant de constater la similitude entre ces chiffres, mais en l'absence de données liées à la fiche bioproject, nous ne pouvons que supposer une confusion et un mélange des données survenus lors de la reclassification de l'espèce. Nous ne sommes pas en mesure d'expliquer en revanche, l'écart de près de deux mégabases annoncé dans la taille du génome. Enfin, il est spécifiquement fait mention au sein de la page MycoCosm que les données proviennent bien du papier de référence du séquençage du génome de *P. rubens* Wisconsin 54-1255, mais que la copie du génome qui y est présentée n'est plus soutenue. En raison de tous ces éléments, et malgré la somme d'informations contenues sur MycoCosm, nous avons préféré exclure les données provenant exclusivement de cette base de données, que ce soit lors des étapes du reconstruction ou de curation.

PchCyc est un PGDB classé au niveau Tier 2 de BioCyc. Il a été construit à partir du génome annoté de *P. rubens* en utilisant l'outil Pathologic ainsi que la base de données MetaCyc. Comme précisé dans sa fiche descriptive, le réseau de réactions biochimiques résultant a ensuite été amélioré par une curation manuelle limitée. Le diagramme de la carte métabolique de l'organisme, présenté en **Figure 3-9**, offre une vue d'ensemble cellulaire de sa machinerie biochimique, structurée en voies regroupées en clusters apparentés. Ces cartes, accessibles *via* BioCyc, suivent une architecture générale où le cycle de l'acide citrique (TCA) sépare les voies principalement cataboliques à droite des voies d'anabolisme et de métabolisme intermédiaire à gauche. L'ensemble des réactions individuelles du côté droit du diagramme correspond aux réactions du métabolisme des petites molécules qui n'ont pas encore été spécifiquement attribuées à une voie particulière. Nous observons dès lors un défaut dans la connexion des composés. Au sein de PchCyc, un total de 13 931 gènes est répertorié, dont 13 670 peuvent être directement associés à l'aide du code de référence mentionné précédemment. Ces identifiants sont identiques à ceux retrouvés dans les données issues du portail JGI MycoCosm. Nous y trouvons également 879 pseudogènes, ainsi qu'une grande quantité de données organisées et accessibles *via* les SmartTables. Comme PchCyc bénéficie d'une curation manuelle, les données contenues dans ce PGDB ont été utilisées pour enrichir le draft d'***i*Prub22**. En revanche, il est à noter que les séquences utilisées pour la construction de cette base de données sont regroupées sous le numéro d'accession GenBank NS_000201, un enregistrement aujourd'hui obsolète et supprimé de RefSeq.

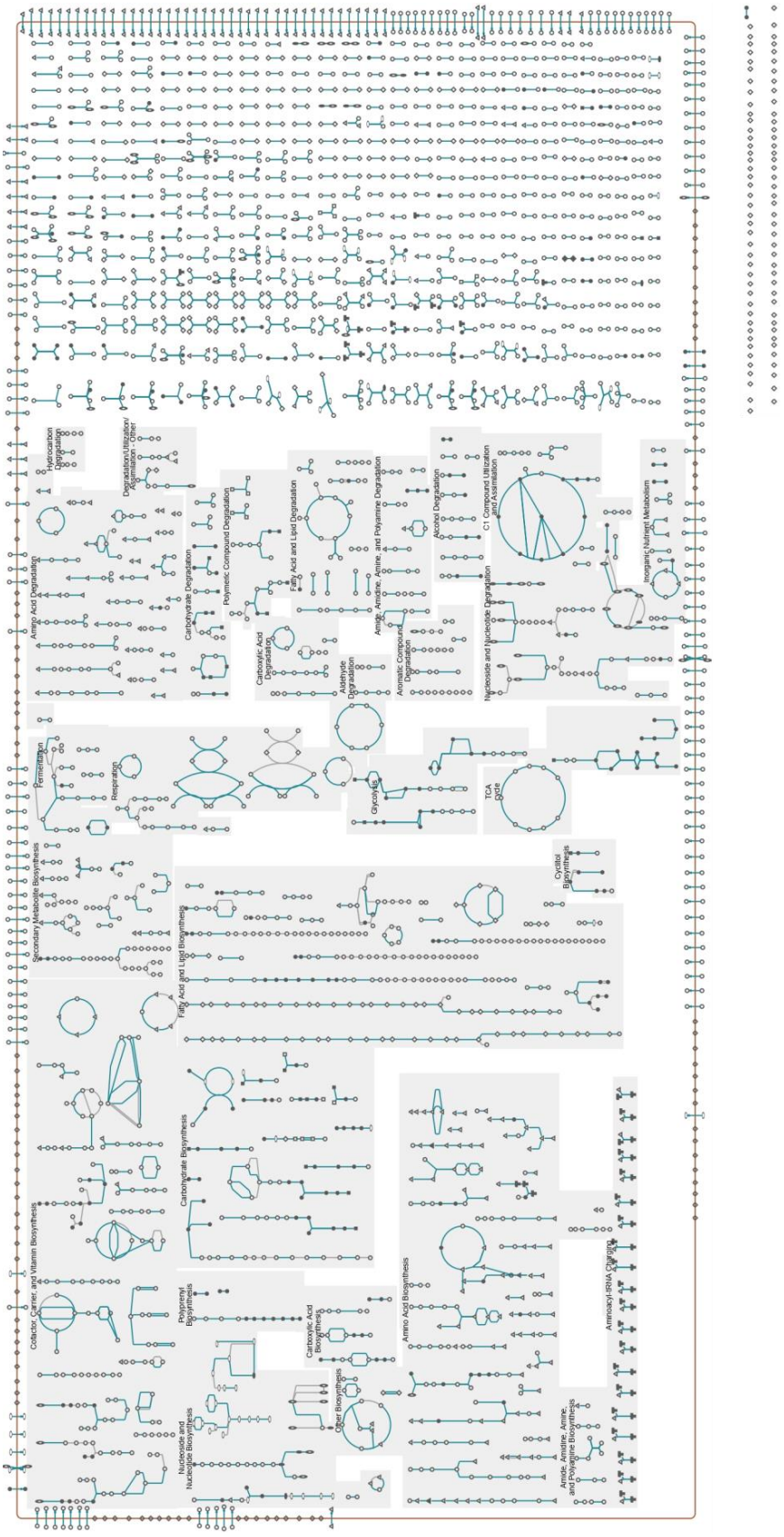


Figure 3-9 : Diagramme de la carte métabolique de PchCyc extrait de BioCyc. *Une membrane cellulaire unique sépare le milieu extracellulaire du milieu intracellulaire, intégrant les transporteurs avec des flèches indiquant le sens du transport. Chaque ligne bleue représente une réaction biochimique unique, tandis que chaque nœud symbolise un métabolite. Les triangles représentent les acides aminés, les carrés les glucides et leurs dérivés, les losanges les protéines, les ellipses les bases azotées, les « T » les ARNt, et les cercles toute autre catégorie de composé. Les formes remplies indiquent des composés phosphorylés.*

Enfin, en interrogeant la base de données UniProt avec l'expression clé « *P. rubens* Wisconsin 54-1255 » et en vérifiant la concordance de l'identifiant taxonomique du NCBI, nous dénombrons 12 789 entrées sur UniProt dont 12 512 comportent la référence directe de l'ORF (*i.e.* `Pc[0-9]{2}[0-9]{5}`) à laquelle ils appartiennent. Parmi les 277 séquences sans référence directe aux ORFs, et par conséquent non représentées en **Figure 3-8**, nous retrouvons, entre autres, les enzymes de la voie de biosynthèse de quelques métabolites spécialisés dont celles de la voie des roquefortines et de la méléagrine, des chrysogines, de la PR-toxine et de l'andrastine.

De prime abord, les différences entre les chiffres présentés en **Figure 3-8** peuvent sembler anecdotiques, d'autant plus que seulement 0,3 % et 2,1 % des séquences présentes dans *i*AL1006 et Prubens sont absentes des données d'EnsemblFungi. Ces différences peuvent s'expliquer soit par l'absence ou une certaine latence dans le maintien et l'évolution des ressources accessibles *via* les différents portails web, soit par des processus d'annotations propres à chaque institution aboutissant à des résultats légèrement différents, soit, par une confusion entre les données des espèces *P. rubens* et *P. chrysogenum*. Aujourd'hui, il paraît peu probable d'obtenir des annotations suffisamment détaillées et fines pour reconstruire des **GSMNs** hautement sensibles jusqu'au niveau de l'espèce. Néanmoins, la clarté et l'accessibilité des données restent essentielles. Les points évoqués jusqu'à présent témoignent des problématiques scientifiques générales qui ont émergé au cours des dernières années et illustrent l'importance de la mise en place et du suivi des principes directeurs FAIR. Dans le cadre de la reconstruction des **GSMNs**, une transparence exhaustive est certes complexe voire illusoire, mais tout au long des travaux présentés dans ce manuscrit, nous avons cherché à respecter au mieux ces principes.

Logiquement, le dernier point de cette section concerne l'impact potentiel que ces différences ont pu avoir sur la topologie d'***i*Prub22**. Au sein de notre **GSMN**, nous dénombrons 31 séquences génomiques qui n'appartiennent pas à l'ensemble des données extraites d'EnsemblFungi. Parmi ces séquences, 25 ont été obtenues à partir des sources externes *i*AL1006, Prubens et PchCyc, et elles sont impliquées dans un total de 29 réactions. Nous avons également ajouté manuellement à la GPR correspondante, la séquence *Pc21g21390* dont le produit génique est l'enzyme responsable de la catalyse de la première réaction de la voie de biosynthèse de l'isopenicillin N. Enfin, lors de l'enrichissement de l'annotation réalisé avec l'extraction des données présentes sur la base de données KEGG (cf. Annexe 1 : *Reconstruction d'iPrub22 – sous-réseau issu de l'annotation fonctionnelle*, page 425), 5 gènes supplémentaires ont été récupérés. Fait intéressant, ces gènes interviennent dans 18 réactions, dont un quart sont impliquées, selon la classification des pathways de MetaCyc, dans la biosynthèse de métabolites spécialisés. Il convient toutefois de noter que ces réactions correspondent aux étapes initiales de ces voies. Néanmoins, parmi ces réactions, nous relevons la participation de deux réactions dans la biosynthèse de la guadinomine B, deux autres dans celle de la griséofulvine, ainsi qu'une réaction pour chacun des composés suivants : la fusicoccine A, l'ergothionéine, la canavanine, la calonectrine, l'anditomine et la lovastatine A.

2.1.2. LA RÉCONCILIATION DES DONNÉES, UNE APPROCHE PERTINENTE ?

Dans le domaine de la biologie des systèmes, la réconciliation des données est le processus de comparaison, d'ajustement et de fusion de données provenant de différentes sources pour garantir leur cohérence et leur intégrité. La réconciliation vise à homogénéiser des données acquises à diverses échelles : génomique, transcriptomique, protéomique et métabolomique. Dans le cadre de la modélisation de réseaux métaboliques, en identifiant les gènes codants pour des enzymes métaboliques (*i.e.* génomique), en évaluant leur activité et leur régulation (*i.e.* transcriptomique), en complétant ces informations avec les données d'expression et d'activité des protéines (*i.e.* protéomique) et en analysant les concentrations, les profils et les flux de métabolites (*i.e.* métabolomique), la génération de modèles, au sens large, qui en découle vise à décrire avec précision le métabolisme de l'organisme étudié. La cartographie de ces informations permet alors de comprendre les interactions entre gènes, protéines et métabolites. Néanmoins, dans la pratique, l'aspect le plus délicat à gérer est lié à la qualité et l'hétérogénéité des données. En effet, comme elles-possèdent des origines variées, leur intégration est souvent complexe en raison de techniques, de format et de temporalité d'acquisition différentes.

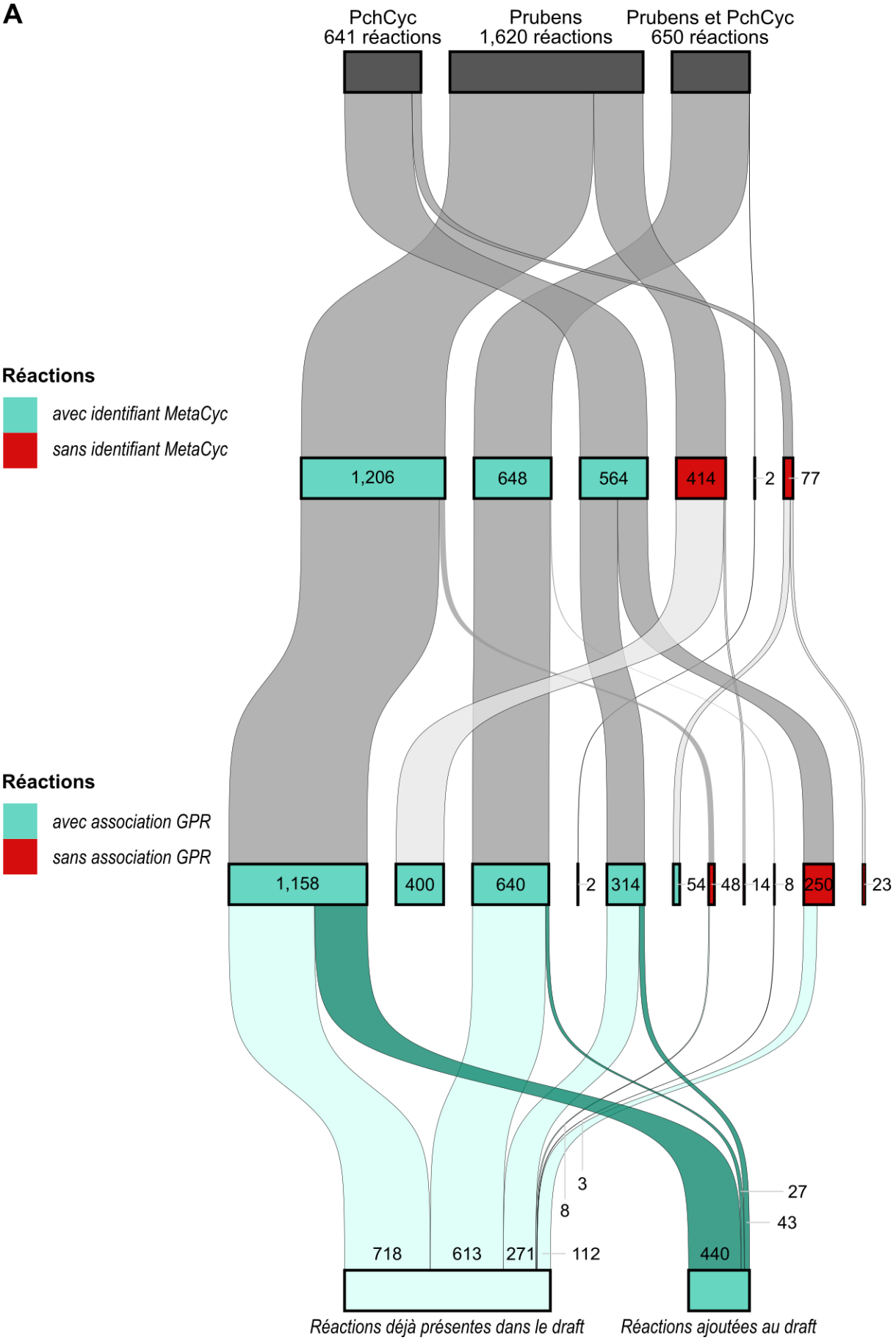
La réconciliation des données est une problématique récurrente dans le traitement de l'information. Le processus sous-jacent s'avère être souvent long et complexe, expliquant pourquoi nous nous sommes concentrés exclusivement sur l'échelle génomique pour la génération d'***i*Prub22**. Ainsi, dans cette section, nous avons choisi de détailler les résultats et les étapes suivies pour enrichir les données des sous-réseaux d'annotation fonctionnelle et de recherche d'orthologie en intégrant les informations des reconstructions précédemment publiées. Notons à ce titre que la diversification des approches adoptées nous a permis d'obtenir une complémentarité des résultats, malgré leur fondement commun basé sur la recherche de similarité entre séquences génomiques. Les résultats présentés en Annexe 3 *: Reconstruction d'iPrub22 – fusion des données et première ébauche de réseau (draft)* (page 445) illustrent cette variabilité de données. Enfin, d'autres exemples, tels que le choix des *templates* utilisés lors de la recherche d'orthologie ou la conception des listes de métabolites servant à vérifier la qualité des modèles, auraient pu nous servir pour illustrer les enjeux de cette thématique de réconciliation de données. Cependant, ces considérations dépassent le cadre des objectifs de cette thèse.

Comme énoncé lors du Chapitre 1, section *3.3 Penicillium rubens Wisconsin 54-1255* (page 72), *P. rubens* est un organisme modèle pour lequel de nombreuses études ont été menées, générant ainsi une abondance de données. Pour enrichir notre draft, outre le PGDB PchCyc présenté dans la section précédente, nous avons utilisé deux **GSMNs**, *i*AL1006 *(Agren et al. 2013)* et Prubens *(Prigent et al. 2018)*, qui comprennent respectivement 1 632 réactions soutenues par 1 006 gènes et 2 574 réactions appuyées par 1 793 gènes. Ainsi, et en faisant abstraction de la compartimentation, nous disposions d'un ensemble de 4 543 réactions à interroger.

Dans notre approche pour générer le draft de notre champignon, nous avons axé nos efforts sur l'exploitation des informations présentes dans son génome. Ainsi, notre principal critère d'intégration des données des reconstructions antérieures repose sur la présence d'une association GPR. Ensuite, le second filtre que nous avons appliqué concerne l'interopérabilité des identifiants. Seuls les éléments référencés dans MetaCyc ont été sélectionnés pour être ajoutés au draft, ce qui inclut les réactions avec un identifiant MetaCyc ainsi que celles sans identifiant MetaCyc mais dont l'ensemble des substrats possèdent un identifiant de métabolites MetaCyc. Toutefois, pour des raisons pratiques, ces filtres ont été inversés pour sélectionner les données. Afin de visualiser les apports, les pertes et les perspectives d'amélioration, nous avons subdivisé le *pool* de réactions à examiner en trois sections distinctes. Un résumé synthétique de ces éléments est présenté en **Figure 3-10**.

- **Réactions spécifiques et communes à PchCyc et Prubens (Figure 3-10.A)**

Le processus initial de discrétisation se fonde sur la présence d'un identifiant MetaCyc actif, c'est-à-dire référencé et non obsolète au moment de la recherche. Les premières données examinées concernent les ensembles de réactions communes et spécifiques à Prubens et PchCyc. Parmi les 2 911 réactions à notre disposition, 83 % sont en accord avec ce premier critère. Dans l'étape suivante, nous classons l'ensemble des réactions en fonction de la présence ou de l'absence d'une association GPR. Nous constatons que sur les 493 réactions écartées lors du premier filtrage, 92 % sont perdues en raison d'une incompatibilité d'identifiant. Enfin, seules les réactions non détectées lors de la génération des sous-réseaux d'annotation et d'orthologie ont été ajoutées au draft. Ainsi, sur les 2 235 réactions d'intérêt, 23 % ont été intégrées pour enrichir le draft, indiquant que les 77 % restants ont été récupérés par annotation fonctionnelle et recherche d'orthologie, renforçant par conséquent le poids du draft. Il est également à noter que 123 réactions, provenant principalement de PchCyc et dépourvues de soutien génomique, sont déjà incluses dans le draft, suggérant que depuis leur dépôt sur PchCyc, l'enzyme responsable de leur catalyse et son gène associé ont été caractérisés.


A
PchCyc
641 réactions
Prubens
1,620 réactions
Prubens et PchCyc
650 réactions
Réactions
avec identifiant MetaCyc
sans identifiant MetaCyc
1,206
648
564
414
2
77
Réactions
avec association GPR
sans association GPR
1,158
400
640
2
314
54
48
14
8
250
23
3
8
27
43
718
613
271
112
440
Réactions déjà présentes dans le draft
Réactions ajoutées au draft

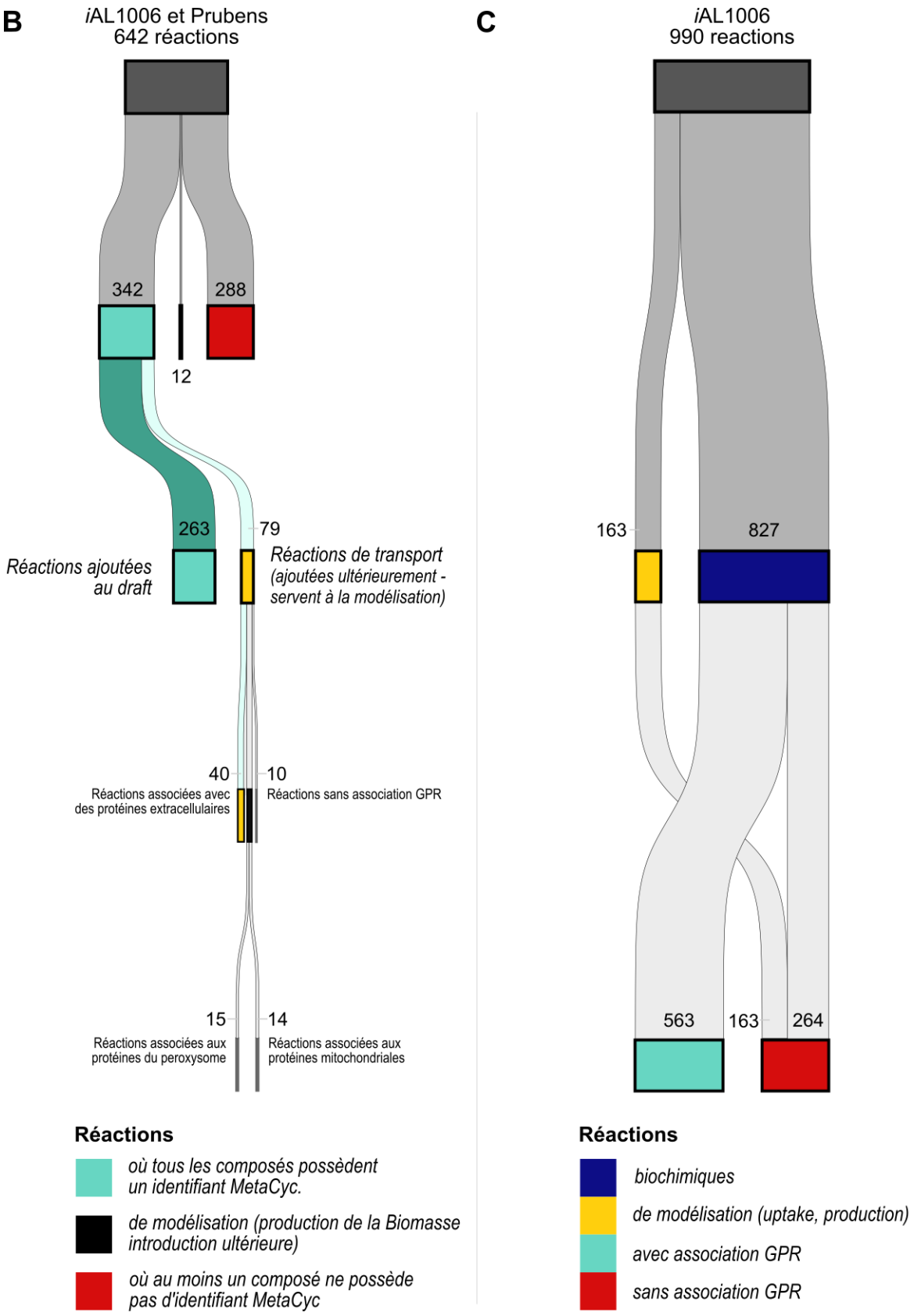


Figure 3-10 : Diagramme de Sankey illustrant le processus de sélection et d'intégration des réactions provenant des sources externes. *En fonction de la nature des données (i.e. présence d'identifiants MetaCyc et d'associations GPR) et de leurs traitements, les réactions sont réparties en trois ensembles distincts:* ***(A)*** *réactions propres et communes à Prubens et PchCyc,* ***(B)*** *réactions communes à iAL1006 et Prubens, et* ***(C)*** *réactions spécifiques à iAL1006.*

- **Réactions communes à Prubens et *i*AL1006 (Figure 3-10.B)**

Pour faciliter la sélection et améliorer la lisibilité, les réactions communes à Prubens et *i*AL1006 ont été traitées séparément. Le **GSMN** Prubens, issu des travaux de Prigent *et al.* *(2018)*, fut reconstruit en fusionnant deux drafts résultant de la comparaison entre le protéome de *P. rubens* Wisconsin 54-1255 et les données de MetaCyc et d'*i*AL1006. Le travail de réconciliation des données avait alors été effectué lors de ces travaux puisque les deux ensembles furent ensuite fusionnés à l'aide d'une liste d'identifiants de réactions et de métabolites établie manuellement entre ceux de *i*AL1006 et de MetaCyc. En l'absence de correspondance, les identifiants de *i*AL1006 étaient privilégiés. De notre côté, afin de maximiser à la fois, le nombre d'informations provenant de *i*AL1006, le **GSMN** d'origine bénéficiant d'une intense curation manuelle, et l'interopérabilité des données, nous avons choisi de sélectionner les réactions dont tous les produits et réactants possédaient un identifiant MetaCyc valide. Après avoir vérifié qu'au moins un gène était associé à chaque réaction, 53 % des réactions partagées entre *i*AL1006 et Prubens ont été incorporées dans le réseau. Afin d'assurer la traçabilité des données, nous avons également préservé l'identifiant provenant d'*i*AL1006. Enfin, les réactions spécifiques à la modélisation, telles que les réactions représentant les sous-systèmes de la réaction de biomasse ou les réactions de transport et d'échange, ont été intégrées lors d'étapes ultérieures (cf. section *2.2.2 Modélisation de différentes conditions de culture*, page 256 et section *2.2.3 Modélisation de la croissance de l'organisme, la réaction de biomasse*, page 277).

- **Réactions exclusives à *i*AL1006 (Figure 3-10.C)**

Le **GSMN** : *i*AL1006 est composé de 849 métabolites uniques, dont 17 % sont dépourvus d'annotations permettant leur identification précise (*i.e.* outre le nom du métabolite, au mieux, un identifiant InChI ou une formule brute leur est associé). Toutefois, même si 703 métabolites sont associés à un identifiant standard (*i.e.* 7 composés avec un identifiant PubChem, 70 avec un identifiant KEGG et 626 avec un identifiant ChEBI), ces informations ne sont pas toujours suffisantes pour trouver une correspondance sur MetaCyc. Ainsi, comme nous avons souhaité intégrer uniquement des données référencées dans cette base de données, nous avons écartés 77 % des réactions biochimiques d'*i*AL1006 de notre draft. Néanmoins, les réactions de modélisation (*i.e. uptake* et *production*) ont été incluses dans le draft lors des étapes ultérieures de curation manuelle.

En résumé, l'exploitation des données provenant des sources externes vient enrichir les 4 143 réactions obtenues lors de la fusion des sous-réseaux d'annotation et d'orthologie, en y ajoutant 773 nouvelles réactions. La contribution des sources externes représente ainsi 17 % de l'ensemble des réactions du draft. Cependant, se pose la question de savoir si ces réactions sont véritablement uniques ou si elles ont contribué à complexifier davantage la reconstruction.

Les systèmes biologiques sont intrinsèquement résistants aux perturbations *(Kitano 2004)*. Ce caractère, attribuable en partie à la redondance génétique, résulte de la duplication des gènes, un processus évolutif courant consistant en la reproduction d'un gène, engendrant des copies supplémentaires dans le génome. Cette capacité d'adaptation est renforcée par la variabilité d'activité des gènes dupliqués en fonction des conditions environnementales, ainsi que par la présence de mécanismes de compensation tels que les voies métaboliques alternatives ou les réseaux de régulation *(Gu et al. 2003)*. Ainsi, la redondance génétique contribue à la robustesse biologique de deux manières principales : par la diversification fonctionnelle, où les copies dupliquées peuvent acquérir de nouvelles fonctions, facilitant ainsi l'adaptation de l'organisme à des environnements variés, et par la redondance fonctionnelle, où la présence de multiples copies d'un même gène permet à l'organisme de tolérer les mutations délétères dans l'une des copies, préservant ainsi les fonctions biologiques essentielles. Un moyen de quantifier et d'évaluer la redondance biologique peut s'effectuer à travers l'examen des réactions létales essentielles et synthétiques dans divers environnements *(Sambamoorthy et Raman 2018)*. Les réactions létales essentielles sont des réactions biochimiques ou processus cellulaires dont la perturbation ou la suppression entraîne la mort de l'organisme dans des conditions dites standards. Les réactions synthétiques, quant à elle, deviennent létales ou compromettent la viabilité de l'organisme lorsqu'elles sont combinées avec la suppression ou la perturbation d'une autre réaction spécifique. En d'autres termes, ces réactions sont synthétiquement létales lorsqu'elles interagissent avec d'autres processus ou gènes, révélant ainsi des interactions fonctionnelles entre différents composants cellulaires.

En conséquence, l'existence de réactions dupliquées et redondantes dans un réseau métabolique est un phénomène naturel reflétant l'évolution. Au sein de la modélisation de réseau métabolique, la redondance devrait donc s'exprimer *via* les associations GPR avec la présence de plusieurs gènes et non par la multiplication de réactions similaires. Or, lors de l'analyse des données de la reconstruction, nous avons identifié des groupes de réactions partageant des équations et associations GPR similaires ou identiques, témoignant ainsi d'une redondance que nous qualifierons d'informatique. À cet égard, nous faisons la distinction entre les réactions dupliquées au sens strict, c'est-à-dire présentant des équations identiques, et les réactions redondantes en raison de leur encodage. Les détails des éléments et des modifications résultants sont présentés dans la version complète du fichier Livescript accompagnant la publication d'***i*Prub22**. Un extrait de ce fichier est également fourni en Annexe 4 : *LiveScript MATLAB détaillant les caractéristiques d'iPrub22*, page 451 du présent document.

Premièrement, comme les réactions dupliquées sont définies par des équations identiques, elles sont aisément identifiables à partir de la matrice stœchiométrique du réseau (*i.e.* recherche de colonnes similaires). La fonction `checkDuplicateRxn()` de la COBRA Toolbox effectue automatiquement cette opération en comparant des paires de réactions, puis en conservant une seule réaction et en supprimant la seconde. Nous l'avons utilisée sans tenir compte de la réversibilité des réactions et jusqu'à ce qu'il n'y ait plus aucune réaction dupliquée dans le modèle évalué. Ces réactions n'ont cependant pas été supprimées, mais uniquement bloquées en ajoutant une contrainte sur les bornes inférieure et supérieure, fixées toutes deux à zéro. Parmi

les groupes de réactions dupliquées, nous observons deux cas de figures distincts : soit les réactions proviennent toutes deux de MetaCyc, auquel cas la redondance est simplement l'expression d'une répétition d'information contenues dans la base de données (cf. le couple de réactions `1.1.1.39-RXN` et `RXN-19748`), soit la duplication résulte du processus de réconciliation où nous confrontons des identifiants référencés contre des identifiants dits maison (cf. le couple de réactions `r0157` et `1.2.1.2-RXN`).

Exemples de réactions dupliquées

ID de réaction : `1.1.1.39-RXN`
Équation de la réaction : `1 MAL + 1 NAD => 1 PYRUVATE + 1 CARBON-DIOXIDE + 1 NADH`
Association GPR : *Pc13g04510* or (*Pc22g15710* and *Pc15g01410*) or *Pc21g20250* or (*Pc22g06830* and *Pc16g05780*) or (*Pc16g03630* and *Pc12g00890* and *Pc21g22770* and *Pc13g15950* and *Pc16g11650* and *Pc20g06810* and *Pc22g12240* and *Pc21g09400* and *Pc21g06640*)

ID de réaction : `RXN-19748`
Équation de la réaction : `1 MAL + 1 NAD => 1 PYRUVATE + 1 CARBON-DIOXIDE + 1 NADH`
Association GPR : *Pc13g04510* or *Pc21g20250*

ID de réaction : `r0157`
Équation de la réaction : `1 FORMATE + 1 NAD => 1 NADH + 1 CARBON-DIOXIDE`
Association GPR : *Pc12g04310*

ID de réaction : `1.2.1.2-RXN`
Équation de la réaction : `1 FORMATE + 1 NAD <=> 1 CARBON-DIOXIDE + 1 NADH`
Association GPR : *Pc12g05480* or *Pc20g14820* or *Pc12g02680* or *Pc21g03190* or *Pc21g23650* or (*Pc12g04310* and *Pc20g10430*)

Deuxièmement, nous sommes confrontés à des groupes de réactions, souvent des paires, qui, bien que représentant la même entité biologique, diffèrent dans leur encodage (*i.e.* identifiants et types de substrats conduisant à des équations différentes). Ce problème, difficilement traitable automatiquement, a émergé suite à l'inclusion des réactions provenant de *i*AL1006. Outre le fait qu'elles possèdent des identifiants propres, c'est-à-dire non universels, nous avons observé deux évènements différents. D'une part, nous retrouvons des réactions où les substrats sont équivalents d'un point de vue biologique mais dont les identifiants imposent une différence dans le traitement des informations par le modèle (cf. couple de réactions `r0027` et `GLU6PDEHYDROG-RXN`). L'utilisation conjointe de termes spécifiques et génériques relatifs à la classification et à la dénomination des métabolites sera discutée et expliquée plus longuement en section *2.2.2 Modélisation de différentes conditions de culture*, page 256 et notamment via l'encart : *Un « même » composé mais des identifiants différents, lequel choisir ? Illustration avec l'entité glucose*, page 259. D'autre part, nous constatons la présence de « pseudos doublons » de réactions où l'une des réactions ne partage pas exactement la même équation que l'autre en raison d'un déséquilibre stœchiométrique. De manière relativement récurrente, c'est le signe d'une absence de proton dans la réaction provenant de *i*AL1006 (cf. couples de réactions `r0027` et `GLU6PDEHYDROG-RXN` ainsi que `r0106` et `GLYCOLATE-REDUCTASE-RXN`). Ainsi, à l'instar de ce que nous avons fait avec les réactions dupliquées, et après nous être assurés que les associations GPR étaient identiques ou que tous les gènes d'une réaction étaient inclus dans l'autre, nous avons retenu la réaction la plus équilibrée, généralement celle provenant de MetaCyc, et bloqué la seconde, celle de *i*AL1006.

Exemples de réactions redondantes

ID de réaction : r0027
Équation de la réaction : ALPHA-GLC-6-P + NADP -> NADPH + D-6-P-GLUCONO-DELTA-LACTONE
Association GPR : *Pc20g03310* or *Pc20g03330*

ID de réaction : GLU6PDEHYDROG-RXN
Équation de la réaction : D-glucopyranose-6-phosphate + NADP -> NADPH + D-6-P-GLUCONO-DELTA-LACTONE + PROTON
Association GPR : *Pc20g03330*

ID de réaction : r0106
Équation de la réaction : NAD + GLYCOLLATE -> NADH + GLYOX
Association GPR : *Pc21g22080*

ID de réaction : GLYCOLATE-REDUCTASE-RXN
Équation de la réaction : NAD + GLYCOLLATE <=> NADH + GLYOX + PROTON
Association GPR : (*Pc21g22040* and *Pc20g15620*) or *Pc20g14820* or (*Pc16g12980* and *Pc21g22080* and *Pc22g04940*) or *Pc22g04940* or *Pc21g23650* or *Pc20g10430*

La redondance est donc un terme, qui en fonction de son domaine d'utilisation peut revêtir une certaine ambigüité. Il est donc nécessaire de faire la différence entre redondance biologique et redondance informatique. Notons enfin que la réconciliation de données externes est la cause principale mais pas unique à ce phénomène.

En fin de compte, même si ces informations de redondance demeurent pertinentes dans le cadre de la reconstruction, elles complexifient considérablement les modèles. Sur la base de ces éléments, nous avons bloqué 84 réactions en nous appuyant sur les résultats de la fonction `checkduplicateRXN()` et 43 autres suite à la comparaison des réactions issues de *i*AL1006 avec celles qui possèdent un identifiant MetaCyc.

Au vu du nombre de réactions dans ***i*Prub22**, ces données peuvent sembler de prime abord minimes. Cependant, la modélisation de flux peut-être hautement sensible à ce type de contraintes. En effet, ce sont précisément ces ajustements, effectués lors des phases de débogage des modèles, qui ont permis de proposer un modèle pleinement fonctionnel. Ainsi, ces modifications, combinées à d'autres éléments détaillés dans le Livescript, ont probablement permis de briser des cycles futiles, libérant et débloquant graduellement les modèles (*e.g.* sensibilité à la nature et à la quantité des sources d'azote, sensibilité à la quantité d'oxygène ou de phosphate inorganique, *etc.*). De plus, la reconstruction paramétrée ***i*Prub22** est composée de 280 réactions provenant d'*i*AL1006, parmi lesquelles 212, soit 75 %, ont été bloquées. Elles représentent 14 % de l'ensemble des 1 489 réactions ne portant aucun flux dans des conditions standard de modélisation.

Ainsi, l'ensemble des informations présentées ci-dessus, confirme la difficulté à bien des égards de réaliser la réconciliation des données provenant de sources externes, et ce quand bien même les bases de données interrogées sont identiques. De plus, ces résultats soulèvent des interrogations quant à la valeur ajoutée de telles approches, au regard du ratio entre l'amélioration significative de la reconstruction et des modèles par rapport au temps investi. Il s'avère donc crucial de trouver un juste équilibre entre le temps consacré à l'intégration de données préexistantes et celui nécessaire à la génération de nouvelles données.

2.1.3. UN GSMN DE HAUTE QUALITÉ ?

2.1.3.1. Cadre général et principes fondamentaux

Dans l'ordre chronologique de nos travaux, les questionnements relatifs à la qualité d'un modèle se sont imposés dès que nous nous sommes éloignés de l'utilisation d'AuReMe, c'est-à-dire dès que la topologie d'***i*Prub22** fut fixée et que les premiers tests de fonctionnalités furent concluants. Pour approfondir nos analyses, nous avons naturellement utilisé la suite COBRA Toolbox proposée dans MATLAB. À ce stade, nous avons constaté une importante perte d'informations, notamment au niveau des annotations. En effet, la plupart des annotations initialement présentes dans les fichiers de reconstruction ne figuraient plus ou n'étaient pas reconnues dans le format **SBML**. Ce constat a été confirmé ultérieurement avec l'utilisation de MEMOTE. Il s'est avéré ensuite que ces observations étaient le reflet de symptômes bien plus importants liés à des questions de format et d'encodage des informations, limitant alors grandement la réutilisabilité de notre modèle, et de ce fait, sa qualité. Cependant, avant d'approfondir ces aspects, il est crucial de se poser une question fondamentale : que signifie exactement l'expression « **GSMN** de haute qualité » ?

> **QUALITÉ N. F.**
> *XIIe siècle. Emprunté du latin qualitas, « qualité, manière d'être », lui-même dérivé de qualis, « quel, de quelle sorte ». […] Ce qui appartient en propre à une chose et la distingue d'une autre ; caractère particulier, propriété. […] Par extension, degré d'excellence relative, valeur que l'on attribue à une chose et qui permet de la juger, de la classer par rapport à une norme de référence ou par rapport à des choses analogues.* Dictionnaire de l'Académie Française, 9e édition (2011)

La qualité pour reconstruire un réseau métabolique se réfère à la fiabilité et à la précision des données utilisées ainsi qu'à la robustesse des méthodes employées pour générer le modèle. Cela inclut la qualité des annotations des gènes et des protéines, la précision des réactions biochimiques associées et la cohérence des données expérimentales utilisées pour informer le modèle. En outre, la qualité d'un réseau métabolique reconstruit dépend de son aptitude à capturer fidèlement les interactions entre les métabolites et les enzymes, ainsi que sa capacité à prédire avec précision le comportement du système biologique sous diverses conditions. Néanmoins, la question de l'évaluation de la qualité d'un **GSMN** est ambiguë. Selon la thématique explorée et en l'absence de référencement solide, la réponse peut être variable. Ainsi, pour limiter et retarder les phénomènes d'obsolescence des réseaux, le modèle proposé répond à certains nombres de recommandations établies par la communauté *(Carey et al. 2020)*.

La question relative à la qualité d'un **GSMN** a déjà été examinée dans la section appropriée de l'article, et nous rappelons à cet égard les éléments suivants. La conception et l'adoption systématiques de critères uniformes pour évaluer la qualité d'un modèle dans l'histoire de la reconstruction des **GSMNs** sont relativement récents. À notre connaissance, à l'exception de la suite de tests MEMOTE *(Lieven et al. 2020)*, dont nous détaillerons ci-après le mode de fonctionnement, il existe peu d'outils ou de mesures permettant d'évaluer la qualité des reconstructions et modèles publiés. Notre attention s'est alors portée sur les trois critères minimaux définis par les travaux de Monk et *al.* *(2014)* :

- **La couverture métabolique** : quel est l'espace métabolique, c'est-à-dire la proportion du génome, recouvert par les données présentes dans le **GSMN** ? Cet aspect a été abordé précédemment au sein de l'article, notamment dans la section « *1.3.2.1 Connectivity and metabolic coverage* » (page 151) et en **Figure 3-5** (page 153). Pour rappel, la reconstruction ***i*Prub22** présente une couverture métabolique de 45 % contre 8 % et 14 % pour *i*AL1006 et Prubens.

- **L'étendue de la curation manuelle** : quels sont les aspects et les moyens mis en œuvre pour affiner le modèle ? Outre les informations d'ores et déjà traitées, nous reviendrons en section *2.2. Les points sujets à question* (page 237) du présent chapitre, sur trois aspects primordiaux : le *gap-filling*, la conception des divers *media* et l'élaboration de la réaction de biomasse.

- **L'utilisation de procédures opérationnelles standards établies** : quels sont les moyens et normes d'écriture à utiliser pour encoder les informations ? Cette étape est indépendante des connaissances biologiques et permet d'élargir le champ d'application des **GSMNs**. Ainsi, nous nous intéresserons ici préférentiellement au format de distribution majoritaire des réseaux métaboliques, le **SBML**, et à la cohérence des annotations des diverses entités d'un modèle. Ce point, centré sur la structure des modèles, teste le format d'échange au sens strict et non pas la qualité du contenu.

S'ajoutent à ces questions les points relatifs aux capacités prédictives des modèles, à la transparence dans l'acquisition des données et à la génération des modèles, à la documentation jointe et plus largement à l'ensemble des éléments liés aux principes FAIR. Il est essentiel de retenir que la reconstruction et les modèles qui lui sont liés se nourrissent mutuellement, c'est-à-dire que divers ajustements ont eu lieu entre ces deux entités avant d'aboutir à la génération et à la version publiée d'***i*Prub22**. Ainsi, concilier le contenu biologique et les approches techniques induit des prises de décisions, parfois arbitraires de la part des concepteurs, entrainant alors des biais implicites et explicites qui nécessitent d'être détaillés pour la compréhension globale et l'utilisation des modèles par autrui.

L'une des clés pour garantir la visibilité, l'utilité et l'accessibilité des modèles réside dans l'utilisation de systèmes standards et de ressources bien établies. Certains de ces dispositifs sont spécifiques à la modélisation de **GSMNs** (*e.g.* méthodes COBRA, package FBC), tandis que d'autres sont applicables à une échelle plus large et à tout type de données informatiques (*e.g.* directives MIRIAM et MIASE). Néanmoins, l'utilisation cohérente et régulière de ces normes n'est pas systématiquement la règle. Les récents travaux de Carey et al. *(2020)*, en s'appuyant sur une enquête auprès de la communauté de la modélisation métabolique, illustrent ce phénomène. À ce titre, leur étude présente une liste de vérifications de méthodes et d'outils à privilégier pour tendre vers des **GSMNs** accessibles et de haute qualité.

Soulignons que, outre le papier de référence sur la reconstruction des **GSMNs** « *A protocol for generating a high-quality genome-scale metabolic reconstruction* » (*i.e.* procédure de 96 étapes réparties en quatre sections : 1-*Creating a draft reconstruction*, 2-*Manual reconstruction refinement*, 3-*Conversion from reconstruction to mathematical model*, 4-*Network evaluation* = *'Debugging mode*) *(Thiele et Palsson 2010a)*, la publication de Carey et *al. (2020)* a été centrale pour légitimer les efforts entrepris visant à rendre notre réseau le plus interopérable et accessible possible. Ainsi, nous présentons en **Tableau 3-3** l'ensemble des ressources standards que nous avons utilisées pour générer ***i*Prub22**. De surcroît, nous nous sommes fortement inspirés de leurs recommandations pour élaborer le fichier Livescript d'***i*Prub22** (*i.e.* documentation récapitulant les caractéristiques de la reconstruction et retraçant les modifications apportées pour générer les modèles).

Tableau 3-3 : Référentiel des ressources standards utilisées pour la génération de la reconstruction paramétrée *i*Prub22. En plus de la documentation des curations automatiques et manuelles fournies, nous nous sommes appuyés sur ces normes, bases de données et outils spécifiquement développés pour la modélisation métabolique ou largement utilisés en biologie des systèmes, pour élaborer un modèle s'approchant au mieux du "gold standard".

■ Bases de données et ressources

BioCyc Ensemble de bases de données et site web offrant une gamme complète d'outils bio-informatiques pour la modélisation du métabolisme.	Base de données métaboliques	(Karp *et al.* 2019)
MetaNetX Plateforme en ligne permettant d'accéder, d'analyser et de manipuler des réseaux métaboliques à l'échelle du génome en fournissant un accès centralisé aux données (*i.e.* réconciliation d'identifiants).	Interopérabilité entre bases de données	(Moretti *et al.* 2021)
Identifier.org Ressource en ligne qui fournit des identifiants persistants et uniques pour divers types d'entités biologiques.	Service de résolution d'identifiants	https://identifiers.org/
BioModels Base de données qui héberge et offre un accès centralisé à une collection de modèles mathématiques de systèmes biologiques.	Diffusion de modèles	(Malik-Sheriff *et al.* 2020)

■ Normes et standards

MIRIAM *Minimum Information Requested In the Annotation of Models* Ensemble de directives visant à standardiser l'annotation des modèles biologiques en fournissant des recommandations pour l'utilisation d'identifiants persistants, de descriptions détaillées, *etc.*	Annotation des modèles	(Novère *et al.* 2005)
MIASE *Minimum Information About a Simulation Experiment* Ensemble de directives qui standardisent la manière dont les expériences de simulation sont décrites.	Documentation expérimentale	(Waltemath *et al.* 2011)
GO **Gene Ontology** Classification hiérarchique des termes normalisés utilisés pour annoter les fonctions biologiques, les processus cellulaires et les localisations moléculaires.	Annotation fonctionnelle	(Ashburner *et al.* 2000)

Tableau 3-3 (suite et fin)

SBO *Systems Biology Ontology* Classification qui fournit un ensemble de termes standardisés pour décrire les concepts et les entités utilisés dans la modélisation des systèmes biologiques.	Annotation des modèles	(Bernasconi et Masseroli 2019)
SBML *Systems Biology Markup Language* Langage de balisage utilisé préférentiellement pour représenter les modèles de réseaux métaboliques.	Format d'échange	(Hucka *et al.* 2018)
FBC *Flux Balance Constraints* Extension du format **SBML** qui vise à étendre les éléments nécessaires à l'encodage des contraintes de flux métaboliques dans les modèles de réseaux métaboliques.	Modélisation métabolique	(Olivier et Bergmann 2018)
COMBINE *COmputational Modeling in BIology Network* Initiative collaborative visant à coordonner le développement de standards, de formats de données et d'outils logiciels pour la modélisation computationnelle en biologie des systèmes.	Standardisation	https://co.mbine.org/

■ Outils et logiciels

COBRA *Constraint-Based Reconstruction and Analysis* Collection de méthodes de modélisation qui utilisent une approche basée sur les contraintes pour prédire quantitativement et analyser les réseaux métaboliques, offrant à la fois des techniques de modélisation de bases et avancées.	Analyses de réseaux	(Heirendt *et al.* 2019)
LibSBML Bibliothèque logicielle open-source permettant de lire, écrire et manipuler, convertir et valider des modèles biologiques au format **SBML**.	Traitement des modèles	(Bornstein *et al.* 2008)
MEMOTE *Model Examination and Model Optimization Tool* Outil de validation et d'optimisation de modèles qui évalue la qualité et la conformité des modèles biologiques aux standards et aux bonnes pratiques en biologie des systèmes.	Validation de modèles	(Lieven *et al.* 2020)
FROG *Flux variability analysis, Reaction deletion, Objective function values, and Gene deletion fluxes* Ensemble d'analyses pour les modèles basés sur les contraintes afin de générer des données de référence standardisées et numériquement reproductibles (utilisé actuellement dans le flux de travail de BioModels).	Validation de modèles	https://www.ebi.ac.uk/biomodels/curation/fbc
Fluxer *FLUX and pathways viewER* Application web pour le calcul et la visualisation interactive des graphes de flux à partir de modèles métaboliques à l'échelle du génome.	Visualisation	(Hari et Lobo 2020)

2.1.3.2. Évaluation de la qualité selon MEMOTE : détails et analyse du rapport

Parmi les éléments mentionnés ci-dessus, nous allons à présent nous attarder sur MEMOTE. Cet outil est une suite de tests visant à qualifier un modèle en fournissant un résumé de ses caractéristiques, tant en termes de topologie que de flux, et en attribuant un score de qualité qui varie de 0 à 100 %. Ce dernier, comme tout score, est largement discutable et sujet à débat, puisqu'il considère principalement la teneur en annotations du modèle. À titre d'exemple, en se basant sur l'équation du calcul du score du rapport et de la documentation, le simple ajout de termes SBO aux entités monitorées par MEMOTE garantit un score minimal de 25 %. De surcroît, à travers l'exemple suivant, nous tenons à insister sur le fait que MEMOTE met en évidence la qualité de la structure des réseaux métaboliques plutôt que leur contenu. Le principal défi rencontré lors de la reconstruction a été la fonctionnalité du modèle. Pendant une période de plus de six mois, bien que les modèles testés aient été capables de produire de la biomasse, ils se sont révélés insensibles aux variations des sources de carbone, tant en termes de nature que de quantité. La résolution de ce problème majeur n'a toutefois eu aucun impact sur le score MEMOTE.

En revanche, les détails du rapport nous ont permis d'effectuer divers diagnostics, d'identifier et de cibler les points à améliorer. Ainsi, cet outil constitue une ressource facilement accessible pour tendre vers l'uniformisation des modèles, pour détecter certains points bloquants ou anomalies et pour aider la curation manuelle. En conséquence, même s'il faut retenir que la mise en avant d'un score MEMOTE élevé ne constitue pas un critère suffisant de la qualité intrinsèque d'un **GSMN** (*i.e.* le score que nous affichons ne reflète que l'interopérabilité du format basée sur l'écriture des informations et la reconnaissance des entités entre différentes bases de données), il offre une base solide pour répondre aux besoins de standardisation.

D'un point de vue plus technique, MEMOTE est un outil conçu et optimisé pour évaluer des modèles répondant préférentiellement au format SBML3 FBC2 *(Hucka et al. 2018; Olivier et Bergmann 2018)*. Les réseaux antérieurs *i*AL1006 et Prubens ont respectivement obtenu un score de 31 % et 16 %. Cependant, ces résultats sont à relativiser puisque, les modèles n'étant pas encodés avec le package FBC, certaines caractéristiques ne peuvent être évaluées convenablement. Ce point concerne notamment les analyses liées aux gènes et à leurs produits puisque cette notation est introduite à partir de la version 2 du package FBC. De ce fait, et de manière générale, il est nécessaire d'évaluer les résultats de comparaison avec précaution.

Par ailleurs, les résultats d'***i*Prub22** sont également discutables et doivent être contextualisés selon les partis pris lors de la reconstruction, telles que la création de gènes artificiels pour aider à leur identification, la conservation systématique de toute donnée supportée par une information génomique ou l'annotation de certains temes SBO non monitorés. Un autre aspect non-négligeable à prendre en considération concerne directement le comportement de l'outil influencé, entre autres, par l'utilisation de *parsers* plus ou moins spécifiques. Nous noterons à cet effet que des discordances existent dans les catégories *Biomass* (*e.g.* identifiants MetaNetX non reconnus pour l'ATP et plus généralement ceux impliqués dans les tests de détection des cycles de production d'énergie erronés) et *Energy Metabolism* (*e.g.* réactions NGAM et GAM non retrouvées – cf. section *2.2.3 Modélisation de la croissance de l'organisme, la réaction de biomasse*, page 277). Ainsi les caractéristiques de la section qui couvre les statistiques générales et les aspects spécifiques d'un réseau métabolique (*i.e.* non référencées dans le calcul du score) sont à évaluer avec prudence.

Selon les différents tests évalués par MEMOTE, nous rappelons également que notre **GSMN** présente un score de 74 % (**Figure 3-11.A**), suggérant des possibilités d'amélioration. Le rapport détaillé de MEMOTE sur ***i*Prub22** est disponible en Annexe 5 : *rapport MEMOTE*, page 479, et ci-dessous, nous examinons les critères et suggestions, si elles existent ou si elles sont pertinentes, pour l'amélioration du **SBML**, et par conséquent, du score MEMOTE.

▲ CRITÈRES ÉVALUANT LA COHÉRENCE ET L'HOMOGÉNÉITÉ DU MODÈLE

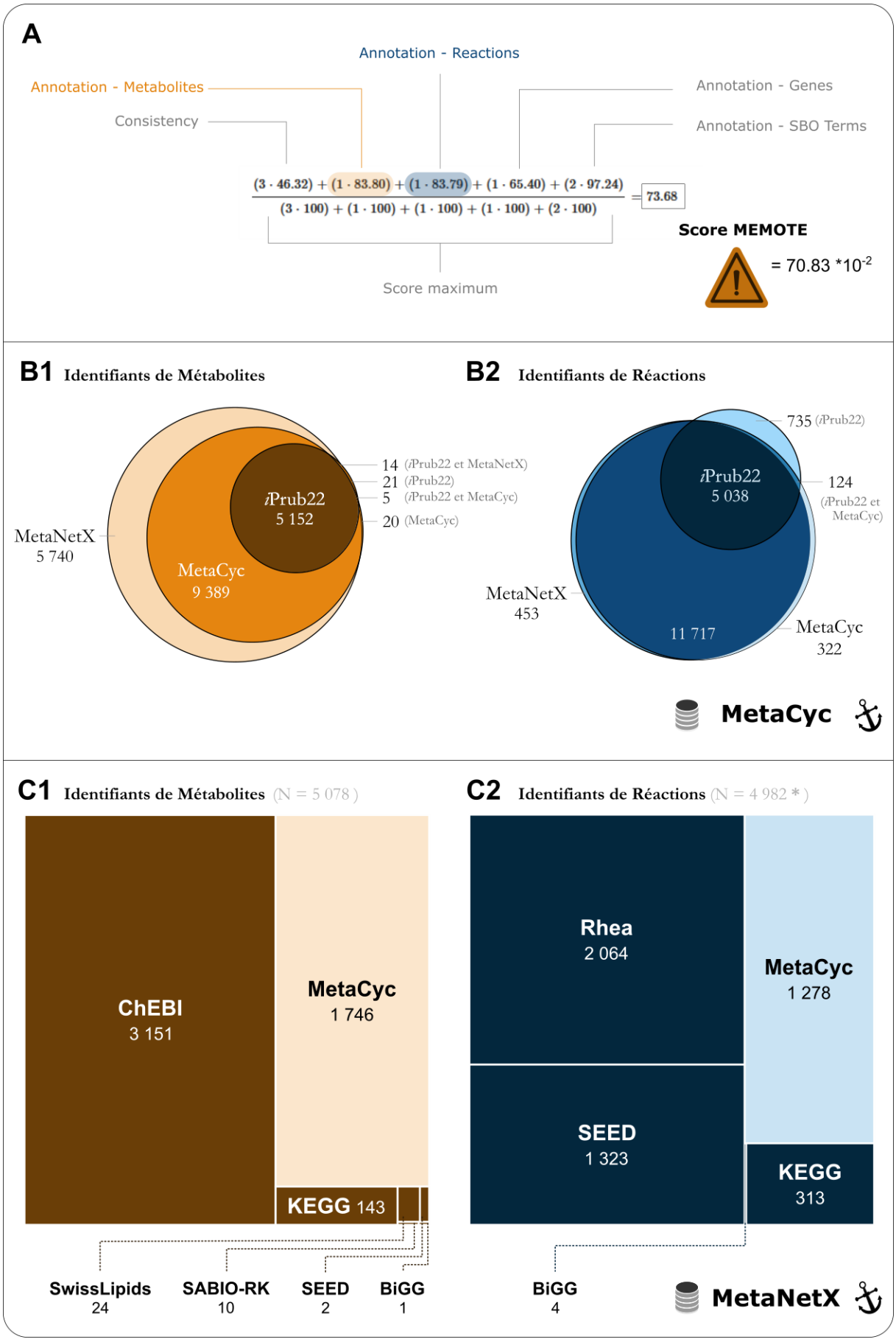
Avec un score affiché de seulement 46 %, la section du rapport présentant la plus grande marge de progression pour la qualité d'***i*Prub22** se caractérise par les items suivants :

Stoichiometric Consistency 0%

La cohérence stœchiométrique est évaluée par une approche de notation binaire, où un score de 0 est attribué dès qu'un métabolite est détecté dans les sections "*unconserved metabolites*" ou "*inconsistent minimal net stoichiometry*". Étant donné que nous n'avons pas nettoyé les données extraites de MetaCyc, les informations présentées dans la section "*consistency*" reflètent fidèlement les caractéristiques intrinsèques des éléments provenant de cette base de données. De surcroît, notons que nous évaluons l'intégralité de la reconstruction qui est, certes, paramétrée, mais qui possède également ces « incohérences » propres. Pour analyser le modèle strictement, nous pourrions nous concentrer exclusivement sur les réactions porteuses d'un flux (*i.e.* environ un tiers des réactions totales) et ainsi éliminer l'ensemble des éléments bloqués (*i.e.* réactions ne portant aucun flux de par la modélisation et réactions que nous avons délibérément bloquées en raison essentiellement d'un équilibre stœchiométrique non respecté).

Mass Balance 79.7% et Charge Balance 84.9%

Les réactions dont les équilibres de masse ou de charge ne sont pas respectés se caractérisent par deux cas : soit la somme des masses et des charges des substrats diffère de 0, soit l'une des entités de l'équation manque de données. Dans notre reconstruction, nous avons identifié 4 réactions avec des déséquilibres de charge, 299 avec des déséquilibres de masse, et 852 présentant à la fois des déséquilibres de charge et de masse. Pour garantir la fonctionnalité des modèles, nous avons fermé divers types de réactions telles que les réactions avec incohérences stœchiométriques, les réactions redondantes et les réactions dupliquées (cf. détails disponibles dans le LiveScript) et enfin les réactions présentant un déséquilibre de masse et/ou de charge. Cependant, bloquer l'ensemble des réactions non équilibrées est une contrainte forte, et pour assurer à la fois la production de biomasse et de métabolites spécialisés, 56 réactions présentant un défaut de charge et/ou de masse n'ont pas été contraintes (*e.g.* réactions générant les sous-systèmes de biomasse, réaction NGAM, *etc.*). Enfin, notons que, contrairement aux indications du rapport MEMOTE, sur les 5 464 métabolites contenus dans ***i*Prub22** (*i.e.* nombre tenant compte des éléments dupliqués dû à la compartimentation), 188 métabolites ne sont pas associés à une charge et 194 à une masse.


A
Annotation - Reactions
Annotation - Metabolites
Consistency
Annotation - Genes
Annotation - SBO Terms
(3 · 46.32) + (1 · 83.80) + (1 · 83.79) + (1 · 65.40) + (2 · 97.24) / (3 · 100) + (1 · 100) + (1 · 100) + (1 · 100) + (2 · 100) = 73.68
Score maximum
Score MEMOTE
= 70.83 *10-2
B1 Identifiants de Métabolites
MetaNetX 5 740
iPrub22 5 152
MetaCyc 9 389
14 (iPrub22 et MetaNetX)
21 (iPrub22)
5 (iPrub22 et MetaCyc)
20 (MetaCyc)
B2 Identifiants de Réactions
735 (iPrub22)
iPrub22 5 038
124 (iPrub22 et MetaCyc)
MetaNetX 453
11 717
MetaCyc 322
MetaCyc
C1 Identifiants de Métabolites (N = 5 078)
ChEBI 3 151
MetaCyc 1 746
KEGG 143
SwissLipids 24
SABIO-RK 10
SEED 2
BiGG 1
C2 Identifiants de Réactions (N = 4 982 *)
Rhea 2 064
SEED 1 323
MetaCyc 1 278
KEGG 313
BiGG 4
MetaNetX

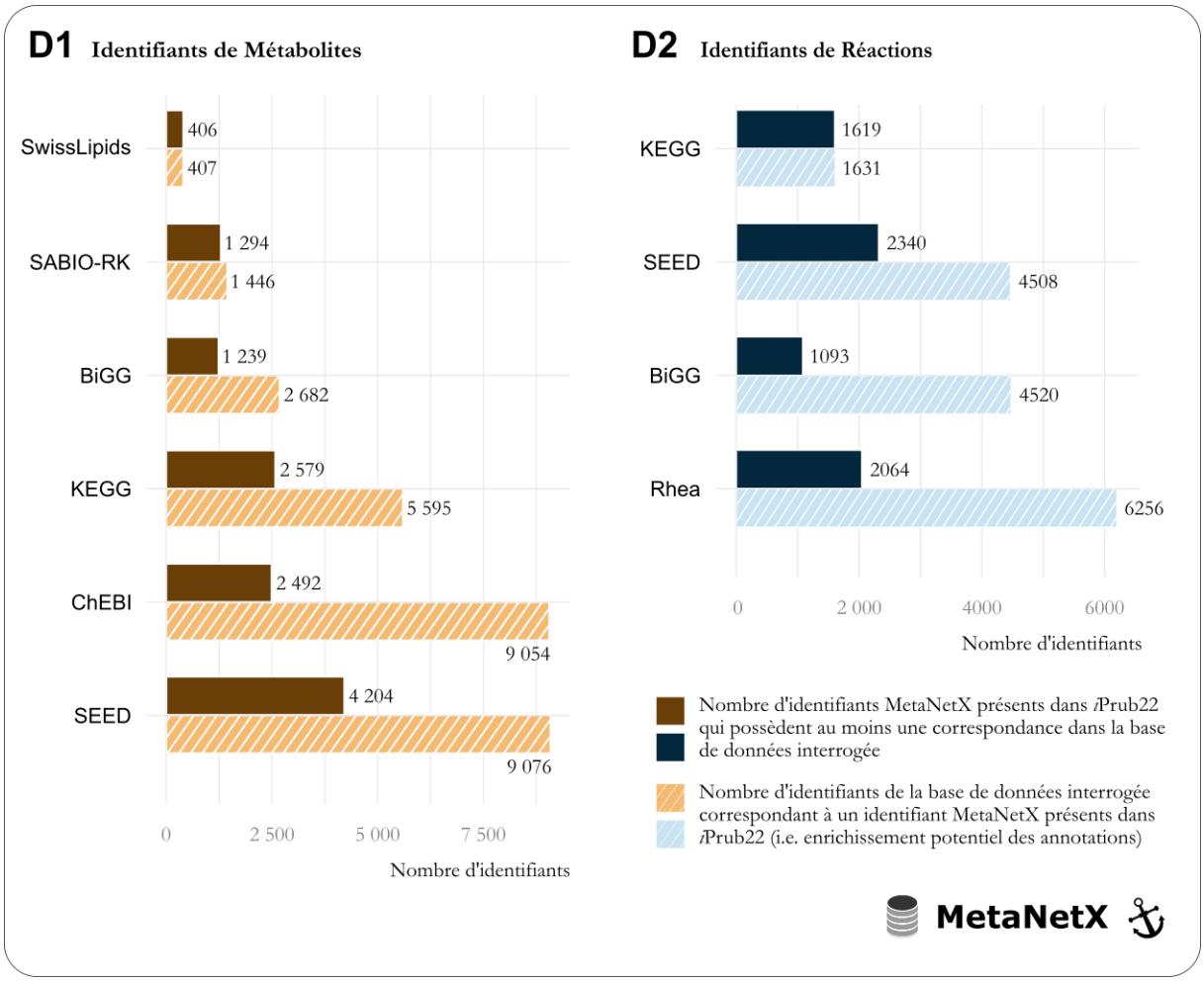


Figure 3-11 : Score de qualité MEMOTE et focus sur les annotions des idendifiants de métabolites et de réactions entre les diverses bases de données métaboliques. (A) *Détails du calcul du score affiché sur le rapport MEMOTE. Cinq catégories, concernant essentiellement la qualité des annotations contenues dans le modèle sont évaluées. Fondamentalement, il existe un léger écart entre le score total affiché et celui calculé. Le fait que cette constatation soit applicable à tous les sous-scores et ne soit pas limitée à* ***iPrub22****, puisque nous l'avons remarquée dans plusieurs rapports MEMOTE de* ***GSMNs*** *récemment publiés, peut être attribuable à une erreur dans l'affichage des résultats ou des calculs (e.g. un facteur de pondération non explicité). Toutefois, une différence de plus de 18 points de pourcentage pour le sous-score évaluant la cohérence du modèle est observée (i.e. score affiché : 46,32 %, score calculé : 64,84 %).* ***(B)*** *Diagrammes d'Euler illustrant la cohérence et l'uniformité des identifiants utilisés pour caractériser les métabolites et les réactions d'****iPrub22****. MetaCyc, base de données de référence pour la reconstruction, sert ici d'ancrage pour la recherche de correspondance d'identifiants (i.e. mapping). Ainsi, ces deux diagrammes ont été générés à partir de la liste de l'ensemble des identifiants contenus dans* ***iPrub22****, de la liste des ensemble de métabolites et de réactions de MétaCyC v23.0 et de l'ensemble des identifiants associés à cette même base de données dans les fichiers MetaNetX v4.1. Nous constatons que la majorité des entités d'****iPrub22*** *possèdent une correspondance MetaNetX, assurant ainsi une connexion vers des références externes. Notons que les éléments dépourvus d'une telle annotation correspondent essentiellement à des éléments de modélisation (e.g. réactions de transport, d'échange ou entités liées à la modélisation de la biomasse). Par ailleurs, nous observons que l'espace d'identifiant de métabolites MetaNetX associé à un identifiant MetaCyc est nettement supérieur au contenu des identifiants de MetaCyc. Ce phénomène, moins prononcé pour les réactions, peut s'expliquer vraisemblablement par la pérennité et le suivi des identifiants MetaCyc au fil des versions.* ***(C)*** *Carte en tranches représentant la proportion et la nature des identifiants de métabolites (C1) et de réactions (C2) extraits de MetaNetX et intégrés à* ***iPrub22*** *(et). Données extraites des fichiers chem_prop.tsv et reac_prop.tsv de la version 4.1 qui contiennent les identifiants MNXref associés à un identifiant représentant au mieux cette entrée dans une ressource externe. Nous disposons de 5 078 et 4 982 identifiants uniques MetaNetX caractérisant respectivement les métabolites et les réactions. Notons que le nombre total d'identifiants MetaNetX associés à des réactions est de 5 037 et nous avons alors observé 36 identifiants MetaNetX associé à plus d'une réaction. Enfin, 34 % et 26 % des identifiants MetaNetX de composés ( ) et de réactions ( ) pointent vers la base de données MetaCyc et n'ont donc pas contribué à l'enrichissement vers des ressources externes.* ***(D)*** *Diagramme en barres représentant l'espace d'identifiants pouvant être enrichi, mais potentiellement au détriment de la qualité. Ces données ont été extraites des fichiers MetaNetX recensant l'ensemble des références externes. Nous constatons que, dans le cadre des données d'****iPrub22*** *et à l'exception des bases de données SwissLipids et SABIO-RK pour les métabolites, ainsi que KEGG pour les réactions, les identifiants MetaNetX sont généralement liés à plus d'un identifiant par ressource externe.*

Metabolite Connectivity 99.8%

De par la méthode de reconstruction, la présence de métabolites déconnectés est impossible, pourtant MEMOTE en détecte 12. Cette quantité minime met en lumière 17 réactions non équilibrées sur un total de 56, dans lesquelles un même substrat agit à la fois comme réactant et comme produit (*i.e.* phénomène se traduisant par une suppression de l'entité, coefficiants nuls, dans la matrice stœchiométrique). Notons que pour la modélisation, ces réactions ont été bloquées (cf. section *2.1.2 du LiveScript* – non présentée dans ce document).

Unbounded Flux In Default *Medium* 59.9%

Lors d'une FVA, les réactions dont le flux est équivalent au flux maximal ou minimal du modèle sont des indicateurs potentiels de problèmes dans la modélisation. Une large fraction des réactions de modèle est caractérisée par ce phénomène (*i.e.* 516 réactions, soit plus 40 % des réactions actives) et l'amélioration intrinsèque de la qualité d'***i*Prub22** dépendra de la résolution de ces incohérences dans le futur. Ces réactions témoignent d'anomalies liées, par exemple, à la définition des contraintes thermodynamiques aboutissant à la présence de cycles thermodynamiquement infaisables (*i.e.* appelés également pathways extrême de cycle III ou *Erroneous energy-producing cycles*). Ces cycles peuvent survenir en raison de réactions irréversibles, de boucles redondantes, ou de déséquilibres énergétiques dans le réseau métabolique. Leur identification et leur élimination sont donc essentielles pour garantir la cohérence et la validité des modèles, car ils contredisent les lois de la thermodynamique (*i.e.* violation des principes de conservation de l'énergie et de l'augmentation de l'entropie) *(Price et al. 2002; Fritzemeier et al. 2017)*. Ces investigations pourront se poursuivre avec CycleFreeFlux *(Desouki et al. 2015)*, qui, jusqu'à présent, nous a permis de confirmer que les prérequis énergétiques d'***i*Prub22** étaient valides ou en testant de nouveaux algorithmes, tel que LooplessFluxSampler *(Saa et al. 2024)* qui met en œuvre un échantillonnage dirigé.

La présence d'*Erroneous energy-producing cycles* dans les versions antérieures d'***i*Prub22** s'est traduit par des phénomènes incohérents observés lors de simulations de modèles où aucun *uptake* n'était autorisé : les flux des sous-systèmes de biomasse ainsi que les flux de production d'éléments énergétiques (*e.g.* ATP, GTP, CTP, UTP, ITP) ou de cofacteurs (*i.e.* NADH, NADPH, $FADH_2$, $FMNH_2$ et Acetyl-CoA) étaient non nuls (*i.e.* vérification à l'aide de la création de réactions de dissipation servant tout à tour d'objectif en FBA). Le blocage des réactions dont les masses et les charges n'étaient pas équilibrées a permis de résoudre l'ensemble de ces incohérences. De surcroît, cela a rendu les modèles désormais sensibles à la nature des sources de carbone qui lui étaient proposées. Le choix de bloquer ces réactions a été motivé par les données de MEMOTE, puisqu'une large proportion des réactions concernées par la catégorie présente était commune aux catégories précédentes.

▲ CRITÈRES ÉVALUANT LES ANNOTATIONS DES MÉTABOLITES ET DES RÉACTIONS

Les travaux de Pham *et al.* *(2019)*, intitulés « *Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling* », résument et démontrent par des mesures basiques de comparaisons l'hétérogénéité des noms et des identifiants trouvés dans les **GSMNs**. Pour pallier ce problème couramment rencontré dans la communauté de modélisation, nous avons choisi d'ajouter des références croisées aux identifiants présents dans notre réseau, provenant à la fois de base de données spécifiques à la modélisation du métabolisme ou indépendantds de ces bases, à l'instar des identifiants de structures InChI, InChIKey ou SMILES pour les métabolites. Une fois n'est pas coutume, le défi majeur de cette approche est de concilier le temps passé à la réalisation de cette tache et la valeur ajoutée à la reconstruction. Les mécanismes nous ayant conduit au peuplement des identifiants seront détaillés ultérieurement dans cette section, et les espaces que nous avons interrogés sont regroupés dans le **Tableau 3-4**. En attendant, les scores du rapport MEMOTE donnent un aperçu des résultats d'interopérabilité d'***i*Prub22**, et notons que les identifiants MetaNetX nous ont servi d'ancrage pour la recherche de correspondance d'identifiants.

Tableau 3-4 : Espaces interrogés pour l'enrichissement des annotations des métabolites et des réactions d'*i*Prub22. Le peuplement des annotations s'est effectué en *parsant* les données du fichier `*.padmet` du réseau puis les fichiers de la base de données MetaNetX (v4.1). Les informations sont issues : soit exclusivement du fichier `*.padmet` ■, soit exclusivement de la base de données MetaNetX ★, soit des deux sources ▲. Pour l'enrichissement des identifiants à partir de MetaNetX, nous avons extrait uniquement l'identifiant de la « meilleure ressource » associé à cet identifiant, c'est-à-dire un élément parmi ceux marqué par une astérisque *. Les informations relatives aux charges et formules brutes des métabolites (non mentionnées dans ce tableau) proviennent de MetaNetX.

RÉACTIONS

Bases de données métaboliques

★* BiGG	Base de données intégrant les modèles génomiques et métaboliques	*(King et al. 2016)*
▲ BRENDA -EC	Base de données d'enzymes et classification enzymatique	*(Chang et al. 2021)*
▲*KEGG	*Kyoto Encyclopedia of Genes and Genomes* – base de données de gènes et de métabolites	*(Kanehisa et al. 2021)*
★* ModelSEED	Génération et optimisation de modèles métaboliques à l'échelle génomique	*(Henry et al. 2010)*
▲*Rhea	Base de données de réactions biochimiques curées manuellement	*(Bansal et al. 2021)*
■ UniProt	Base de données d'informations sur les protéines, incluant des données d'activité enzymatique	*(The UniProt Consortium 2023)*

Tableau 3-4 (suite et fin)

MÉTABOLITES

Ressources chimiques génériques

■ CAS	*Chemical Abstracts Service* - organisation de l'*American Chemical Society*	*https://commonchemistry.cas.org/*
▲*ChEBI	*Chemical Entities of Biological Interest* – dictionnaire des entités moléculaires	*(Hastings et al. 2016)*
■ ChemSpider	Collection de données sur les molécules regroupant leur structures chimiques	*https://www.chemspider.com/*
▲ InChI ▲ InChIKey	*International Chemical Identifier* – identifiant pour les substances chimiques Représentation condensée et numérique de l'InChI	*(Heller et al. 2015)*
■ PubChem	Base de données chimique des substances et de leurs activités biologiques	*(Kim et al 2023)*
★ SMILES	*Simplified Molecular Input Line Entry System* – notation pour représenter les structures chimiques	*(Weininger et al 1988)*

Bases de données métaboliques

▲*BiGG	Base de données intégrant les modèles génomiques et métaboliques	*(King et al. 2016)*
▲* KEGG (compound/drug/glycan)	*Kyoto Encyclopedia of Genes and Genomes* – base de données de gènes et de métabolites	*(Kanehisa et al. 2021)*
★ MetaNetX	Plateforme de réconciliation des réseaux métaboliques à l'échelle génomique	*(Moretti et al. 2021)*
▲*ModelSEED	Génération et optimisation de modèles métaboliques à l'échelle génomique	*(Henry et al. 2010)*

Bases de données métaboliques spécifiques

■ DrugBank	Base de données sur les médicaments, incluant interactions et cibles biologiques	*(Wishart et al. 2017)*
■ HMDB	*The Human Metabolome Database* – base de données métabolomique humaine	*(Wishart et al. 2022)*
■ KNApSAcK	Base de données sur les relations espèce-métabolite, utiles en métabolomique végétale et pharmacognosie	*(Shinbo et al. 2006)*
■ LIPID MAPS	Données sur les lipides, incluant classification et propriétés chimiques	*(Conroy et al. 2023)*
■ MetaboLights	Archives de données expérimentales en métabolomique	*(Yurekten et al. 2023)*
▲*SABIO-RK	Données cinétiques et enzymatiques des réactions biochimiques	*(Wittig et al 2018)*
▲*SwissLipids	Base de données spécialisée dans les lipides pour divers contextes biologiques	*(Aimo et al. 2015)*
■ UM-BBD	*The University of Minnesota Biocatalysis/Biodegradation Database* – Base de données qui recense les voies de biocatalyse et dégradation pour les produits chimiques environnementaux	*(Gao et al. 2010)*

Les différentes conventions de dénomination (*i.e.* espaces de noms) contenues entre les **GSMNs**, voire au sein d'un seul modèle, complexifient la comparaison ou la combinaison de **GSMNs** pré-existants, limitent leur réutilisation et entravent leur expansion. En effet, l'une des applications principales des **GSMNs** réside dans leur capacité à intégrer et à contextualiser une multitude de données issues des diverses sciences omiques *(Gu et al. 2019)* pour lesquelles un ancrage cohérent est nécessaire. Ainsi, un aspect crucial de la qualité d'un **GSMN** réside dans l'utilisation uniforme d'un espace de nom pour l'identification des entités du modèle *(Pham et al. 2019)*.

Uniform Metabolite Identifier Namespace (Metabolites 84.4% ; Reaction 87.5%)

Selon les données fournies par le rapport MEMOTE, nous constatons que 852 identifiants de métabolites et 758 identifiants de réactions divergent de l'espace de noms le plus étendu, à savoir BioCyc. Cependant, une analyse plus approfondie révèle des incohérences dans les entités identifiées comme ne relevant pas de cet espace de nom. Certaines entités, dont les identifiants présentent une forte similitude avec ceux déjà répertoriés dans MetaCyc, ne sont pas considérées comme appartenant à cet espace de nom, ce qui soulève des interrogations. Le nombre de réactions non-conforme à cette directive est quant à lui relativement cohérent, puisque la plupart d'entre elles sont principalement liées à des éléments de modélisation, tels que des réactions de transport, d'échange ou des composants liés à la modélisation de la biomasse. En revanche, en ce qui concerne les métabolites et après avoir éliminé les informations sur les compartiments de ces identifiants, nous identifions 720 métabolites uniques, dont 645 semblent être associés à un identifiant de MetaCyc. Les 75 restants, ne correspondant pas à l'espace de nom BioCyc, proviennent soit de la fonction biomasse, soit de composés dont les identifiants sont probablement obsolètes.

Pour illustrer cette observation, la **Figure 3-11.B** (B1 pour les métabolites et B2 pour les réactions) présente les espaces de noms des identifiants contenus dans ***i*Prub22**, MetaCyc v23.0 et les entrées MetaCyc contenues dans les fichiers MetaNetX v4.1. En nous basant sur la proportion d'identifiants ***i*Prub22** possédant à la fois un identifiant MetaNetX et un identifiant MetaCyc reconnu, nous pouvons conclure que le critère de qualité concernant l'uniformité de l'utilisation d'un espace de nom cohérent pour nos identifiants est pleinement validé.

Metabolite/Reaction Annotations Per Database

Ces sections du rapport représentent le nombre d'entités possédant des références croisées vers diverses ressources externes (*e.g.* base de données de reconstructions métaboliques générales et spécifiques, bases de données de structure de molécules, *etc.*). Notons que, même si cela peut sembler de prime abord contre-intuitif, rechercher un score de 100 % pour chacune de ces catégories est illusoire en raison de l'état actuel des bases de données et de leurs connexions. Nous pourrions même remettre en question la pertinence et la qualité intrinsèque d'un modèle qui présenterait un score « parfait » au sein de cette section.

En revanche, multiplier le nombre de références vers diverses sources permet de maximiser l'espace de couverture. Le **Tableau 3-4** présente donc l'ensemble des ressources utilisées pour enrichir ***i*Prub22** en références croisées.

L'un des défis majeurs rencontrés lors de la recherche de mise en correspondance des identifiants est la cohérence et l'exactitude des associations effectuées. Les espaces de noms sont loin d'être des ensembles bijectifs et un identifiant d'une base de données peut pointer vers divers autres identifiants dans une autre base de données. C'est pourquoi nous avons choisi de réaliser notre enrichissement à partir de l'identifiant de référence associé à un identifiant MetaNetX. Pour approfondir les éléments abordés dans cette partie, nous présentons ces résultats en **Figure 3-11.C**, tandis que la **Figure 3-11.D** illustre l'espace qui pourrait potentiellement être recouvert si nous étions moins stricts.

Enfin, les identifiants garantissant une identification claire et précise des entités possèdent des scores satisfaisants, avec une couverture supérieure à 65 % pour les informations provenant des identifiants MetaNetX de métabolites et des annotations InChI, InChIKey, ChEBI et PubChem, ainsi qu'une couverture supérieure à 70 % pour les identifiants MetaNetX de réactions et d'annotations en numéro EC. Toutefois, une marge de progression conséquente existe pour assurer les connexions entre les bases historiques du métabolisme, notamment BiGG, ModelSEED et, dans une moindre mesure, KEGG. Bien que notre reconstruction gagnerait en réutilisabilité potentielle avec une interopérabilité accrue, cet aspect est néanmoins contrebalancé par l'existence des identifiants MetaNetX.

▲ CRITÈRES ÉVALUANT LES ANNOTATIONS DES GÈNES

En préambule de cette catégorie, soulignons qu'une mise à jour du format (*i.e.* FBC2) a été requise pour évaluer cette section, nous détaillerons cet aspect ultérieurement.

Le peuplement des annotations des gènes a été effectué d'une part, en respectant les espaces des noms définis *via* identifier.org, et d'autre part, sur les bases de données qui étaient pertinentes pour notre organisme, à savoir :

- **RefSeq** : collection de séquences de référence non redondantes, maintenue par le NCBI, représentant une compilation diversifiée de séquences génomiques, transcrites et protéiques.
- **NCBI Gene** : base de données du NCBI fournissant des informations spécifiques aux gènes, telles que leur localisation, leurs annotations fonctionnelles, leurs variations et leurs relations avec d'autres éléments moléculaires.
- **NCBI Protein** : compilation de séquences protéiques provenant de diverses sources (*e.g.* GenBank, RefSeq, SwissProt, *etc.*) offrant une ressource exhaustive pour l'étude et l'analyse des protéines.
- **KEGG Gene** : collection de catalogues de gènes établie à partir de ressources accessibles au public, proposée par la base de données KEGG, offrant une vue globale des gènes impliqués dans divers processus biologiques et voies métaboliques.

- **UniProt** : ressource majeure pour les séquences de protéines et leurs informations fonctionnelles, offrant une plateforme complète pour l'accès aux séquences, aux annotations, aux structures et aux relations des protéines.

Il convient de noter que, dans son calcul du score, MEMOTE semble également analyser des informations issues de bases de données spécialisées telles que EcoGene (*i.e.* base de donnés spécifique à *E. coli*), Consensus CDS (*i.e.* projet collaboratif centré sur l'humain et la souris), *The Human Protein Reference Database* (*i.e.* plateforme axée sur le protéome humain) et ASAP (*i.e. a systematic annotation package for community analysis of genome*, base de données stockant des informations sur les génomes bactériens). Dans le contexte de notre champignon, ces catégories sont logiquement vides et entraînent alors une sous-estimation du score.

Total Genes 6,171 et Gene Annotations Per Database at 92.4%.

> Dans notre reconstruction, afin de faciliter l'exploration du réseau, nous avons associé des gènes artificiels aux réactions de modélisation et spontanées afin de les différencier aisément des réactions de *gap-filling*. Plus précisément, les gènes de *P. rubens* suivent le schéma `Pc\d{2}\g\d{5}`, tandis que les gènes artificiels sont désignés par `[s|u|p|d|t|sk]\d{3}`, associés respectivement aux 86 réactions spontanées, aux 185 réactions d'absorption, aux 42 réactions de production, aux 37 réactions de demande, à 117 réactions de transport et à une réaction puits. En conséquence, le nombre total de gènes est surestimé, ce qui conduit à des pourcentages sous-estimés de présence d'annotations de gènes dans chaque base de données.

▲ CRITÈRES ÉVALUANT LA TENEUR EN ANNOTATION SBO

La définition d'une syntaxe commune permettant de clarifier la sémantique des modèles constitue un défi persistant dans le domaine de la biologie des systèmes. L'ontologie de la biologie des systèmes (SBO) représente un ensemble de vocabulaires relationnels contrôlés de termes couramment utilisés dans ce domaine, et en particulier dans la modélisation informatique. Son objectif est de normaliser la représentation des éléments dans les modèles informatiques afin de faciliter l'échange et l'intégration des données. Les termes SBO permettent de décrire avec précision les composants moléculaires, les interactions entre ces composants, ainsi que les processus biologiques régissant le fonctionnement des systèmes biologiques. Organisés hiérarchiquement, ces termes sont employés pour annoter les modèles biologiques à différents niveaux d'abstraction. Ainsi, l'ontologie comprend diverses branches définissant les expressions mathématiques décrivant le système, la représentation des métadonnées, le cadre de modélisation utilisé (*e.g.* "*constraint-based framework*"), les types d'entités physiques (*e.g.* "*simple chemical*", "*biomass*", "*gene*") et se produisant (*e.g.* "*biochemical reaction*", "*biomass reaction*") ainsi que les rôles des participants à la réaction et les paramètres de description des systèmes (*e.g.* "*flux bound*"). MEMOTE apporte une importance non-négligeable à l'annotation du modèle en termes SBO. Dans l'absolu, cette étiquette, qui n'est pas spécifique à la modélisation des réseaux métaboliques à l'échelle du génome, permet de retrouver aisément chaque entité du modèle en fonction de sa nature.

Notre reconstruction paramétrée ***i*Prub22** recense un total de 22 termes SBO différents utilisés pour caractériser les métabolites, les réactions, les gènes, mais également le modèle (SBO:0000624 - flux balance framework), les flux (SBO:0000625 - flux bound et SBO:0000626 - default flux bound) et les compartiments (SBO:0000290 - physical compartment ; SBO:0000410 - implicit compartment). Cependant, parmi les éléments évalués par MEMOTE, les quatre catégories suivantes n'atteignent pas le score maximal et en tenant compte du fait que nous avons attribué un terme SBO à chaque entité du modèle, nous expliquons ci-après pourquoi cela se produit.

Metabolic Reaction SBO:0000176 (98.3%)

Parmi les 91 réactions considérées comme purement métaboliques par MEMOTE et devant être annotées avec le terme SBO:0000176, nous avons choisi d'associer à chacune d'elles des termes SBO plus spécifiques représentant au mieux leur nature respective. Ainsi, nous comptons 86 réactions spontanées (SBO:0000672), 1 réaction de demande, 3 réactions annotées par le terme parent (SBO:0000167 : *biochemical or transport reaction*), ainsi que la réaction NGAM (SBO:0000630 - *ATP maintenance*).

Transport Reaction SBO:0000185 (88.7 %)

Le rapport précise qu'un total de 36 réactions métaboliques manquent d'annotation. Cependant, ces 36 réactions sont des réactions de demande (`Demand_\d{3}`) et, en conséquence, elles ont été associées au terme SBO approprié : SBO:0000628.

***Demand* Reaction SBO:0000628 (90.2%)**

Les quatre réactions mentionnées par MEMOTE comme manquant de l'annotation SBO:0000628 ne sont pas des réactions de demande, mais des réactions non équilibrées associées à la décomposition d'unités de glycanes. Elles ont été annotées par le terme SBO approprié pour les réactions métaboliques. Rappelons qu'***i*Prub22** comprend un total de 37 réactions de demande, toutes associées avec le terme SBO correspondant.

Gene SBO:0000243 (92.4%)

Le rapport indique qu'un total de 468 gènes (soit 7.58 % de tous les gènes) manquent d'annotation. Ces 468 gènes correspondent aux « gènes artificiels » créés dans le but de faciliter la discrétisation des réactions résultant du *gap-filling*, pour lesquelles aucune association GPR n'a été établie, des réactions de modélisation. Par conséquent, nous avons étiquetés les faux gènes associés aux réactions de transport, d'échange ou spontanées avec le terme SBO:0000291 – *empty set* (*i.e.* entité définie par l'absence de tout objet réel).

Avant de conclure sur les apports de MEMOTE, nous souhaitons faire un aparté sur un autre type de score : le score de confiance attribué aux réactions dans une reconstruction. Cette problématique, loin d'être nouvelle dans la reconstruction de **GSMNs**, a été au cœur de nos préoccupations lors de la reconstruction d'***i*Prub22**, comme en témoignent la génération du draft à partir de diverses sources et nos réflexions autour du processus de « *gap-filling* raisonné » (cf. Annexe 1 : *Reconstruction d'iPrub22 – sous-réseau issu de l'annotation fonctionnelle*, page 425, Annexe 2 : *Reconstruction d'iPrub22 – sous-réseau issu des recherches d'orthologie*, page 433, Annexe 3 : *Reconstruction d'iPrub22 – fusion des données et première ébauche de réseau (draft)*, page 445 et section *2.2.1 Un gap-filling « raisonné »*, page 239). Néanmoins, hormis une information vague sur l'origine du ou des gènes associés à une réaction (*i.e.* informations encodées dans la balise note des réactions sous la forme `<p> CATEGORIES: PREUVE </p>`, où `PREUVE` correspond aux mots-clés `ANNOTATION`, et/ou `ORTHOLOGY` ou `MANUAL`), ou l'absence d'association GPR pour les réactions issues du *gap-filling*, le fichier `*.sbml` d'***i*Prub22** ne fait pas mention d'informations plus précises et explicites sur la pertinence de leur présence dans la reconstruction. Pour assurer la traçabilité nous tenons à souligner que ces informations sont toutefois disponibles dans le fichier `*.padmet` du modèle et dans des fichiers annexes rendus disponibles lors de la publication d'***i*Prub22**. Cependant, ce mode de fonctionnement complique grandement la praticité du modèle.

Thiele et Palsson proposaient dans leur protocole de reconstruction d'inclure un système de notation de confiance pour les réactions. Selon ce système, une réaction avec un score de 0 n'avait pas encore été évaluée, un score de 1 indiquait des réactions nécessitant une validation expérimentale et provenant des données de modélisation. Les réactions basées sur des données physiologiques et des annotations génomiques recevaient un score de 2, tandis que les réactions avec un score de 3 étaient appuyées par des preuves directes ou indirectes de la fonction des gènes (*e.g.* surexpression, *knock-out*, *etc.*). Enfin, les réactions avec un score de 4, représentant le degré de confiance le plus élevé, étaient étayées par des données biochimiques issues de preuves directes (*e.g.* purification de protéines, essais biochimiques, *etc.*) *(Thiele et Palsson 2010a)*. Il est à noter cependant que ces scores, étant cumulatifs, pouvaient parfois devenir difficiles à interpréter et perdre en lisibilité.

Aujourd'hui, pour accomplir cette étape, il serait opportun d'enrichir ***i*Prub22** en termes issus de l'ontologie des preuves et des conclusions (ECO - *Evidence and Conclusion Ontology*) *(Nadendla et al. 2022)*. Cette ontologie est régulièrement utilisée dans les étapes d'annotation et de biocuration - processus par lequel les données biologiques sont collectées, organisées, annotées et présentées de manière à être utilisées de façon efficace par la communauté scientifique garantissant la qualité, l'intégrité et l'accessibilité des données biologiques. De nombreuses ressources biologiques dont la Gene Ontology (*i.e.* historiquement ECO a été développée et soutenue conjointement pour soutenir les annotations de produits génétiques) et la base de données UniProt intègrent ce vocabulaire. L'ontologie ECO fournit un ensemble de termes servant à décrire les types de preuves et les méthodes d'assertion qui justifient la présence d'une donnée dans le modèle, facilitant ainsi le suivi de leur provenance.

Concrètement, pour illustrer ces propos, concentrons-nous sur la biosynthèse de la pénicilline. Notons qu'un extrait de l'encodage en **SBML**, de la réaction de production de la pénicilline `RXN-17062`, du composé lui-même, et du gène PenDE sont disponibles en page 232. Il est admis que la réaction de production de la pénicilline à partir de l'aminopénicillate est catalysée par l'enzyme Isopénicilline-N *N*-acyltransférase, produit génique de la séquence PenDE - *Pc21g21370*. La production de pénicilline étant largement documentée, il serait tout à fait envisageable d'associer à cette réaction et à ce gène le terme ECO:0000006 - *experimental evidence*, voire un terme enfant plus spécifique. En revanche, dans ***i*Prub22**, la réaction de biosynthèse de la pénicilline, `RXN-17062`, est également associée à la séquence *Pc13g09140*, détectée par annotation fonctionnelle. Dans notre reconstruction, nous tenons à souligner une fois de plus que nous avons délibérément choisi de conserver l'intégralité des informations disponibles. Ainsi, pour accroître la qualité de cette association GPR et souligner la similitude des séquences, ce gène pourrait être associé, par exemple, au terme ECO:0007672 - *computational evidence*. En généralisant et en associant les termes ECO les plus pertinents en priorité aux associations GPR, tels que ECO:0000201 *sequence orthology evidence*, ECO:0000205 *curator inference*, ECO:0000218 *manual assertion*, ECO:0000366 *evidence based on logical inference from automatic annotation used in automatic assertion*, ou encore ECO:0007661 *combinatorial computational and experimental evidence*, notre reconstruction gagnerait en précision

Enfin, nous savons que la nécessité d'échanger et d'intégrer les modèles au niveau syntaxique est une préoccupation prépondérante dans la modélisation, résolue par le recours aux ontologies. Ces dernières se définissent par l'utilisation d'un vocabulaire contrôlé, développé et soutenu par le biais de la collaboration de la communauté scientifique. La capture de ces informations permet de suivre la provenance des annotations afin d'établir des mesures de contrôle de la qualité et de faciliter leur interrogation. Toutefois, bien que ces couches sémantiques améliorent la compréhension et l'analyse du modèle d'un point de vue programmatique, elles ne sont pas nécessaires à la fonctionnalité du réseau et peuvent donc, à ce titre, être ajoutées à la modélisation à n'importe quel moment.

Comme stipulé à diverses reprises, notre reconstruction ***i*Prub22** a obtenu un score de qualité de 74 %. Néanmoins avant d'aboutir à ce résultat que nous qualifierons de satisfaisant, plusieurs ajustements de format ont été nécessaires. Nous avons fait mention au début de cette section de problèmes de compatibilité et de reconnaissance au niveau des identifiants du modèle. À cette étape de nos travaux, la version de notre réseau était dotée d'un score MEMOTE de 4 %. Le processus d'enrichissement d'identifiant lié aux questions d'interopérabilité a été l'un des leviers principaux pour l'amélioration du score. Sur cette base et en nous guidant des informations du rapport MEMOTE, nous avons pu améliorer significativement la cohérence et l'interopérabilité de notre modèle.

La pérennité d'un réseau dépend de sa facilité d'utilisation et une annotation des entités du modèle axée sur diverses bases de données revêt un critère indéniable pour accroître la qualité d'un réseau. De plus, lorsque ces annotations appartiennent à des bases de données spécifiques telles que KNApSAcK, base recensant les métabolites spécialisés, elles constituent un point d'entrée pour l'exploration des modèles.

Soulignons toutefois que le cœur d'un réseau métabolique et, à fortiori, du modèle réside dans sa matrice stœchiométrique. Les ajustements de format que nous avons apportés se sont limités à la structure du **SBML**, en mettant l'accent sur l'interopérabilité entre bases données, et *de facto*, entre outils. Comme de nombreux termes présentés jusqu'alors, le mot interopérabilité revêt en fonction de son domaine d'application diverses significations. Nous nous concentrerons ici sur deux aspects majeurs : la connexion entre bases de données avec le peuplement des annotations et la compatibilité entre outils en nous axant sur les standards des formats **SBML** puisque, utiliser de telles normes, est un gage d'exportabilité. Pour atteindre une interopérabilité optimale tant entre bases de données et outils, la méthode utilisée se résume en une simple expression : effectuer le *parsing* des fichiers. Cet anglicisme informatique désigne l'action d'explorer la syntaxe d'un fichier afin d'en vérifier l'écriture et d'en extraire les informations d'intérêt.

Pour comprendre la teneur des scores affichés dans les sections Annotations des métabolites et des réactions, il est nécessaire de comprendre le mode de fonctionnement d'AuReMe. Nous nous intéressons alors plus particulièrement aux fonctionnalité du script `sbmlGenerator.py` qui permet, à partir d'un fichier `*.padmet` d'obtenir le réseau et ses métadonnées au format **SBML**. De manière synthétisée, les informations contenues dans le fichier `*.padmet` sont stockées dans des dictionnaires qui sont ensuite parcourus pour écrire le fichier **SBML**. Notons qu'AuReMe est un outil déployé avec docker ce qui nous permet d'éditer son contenu à notre convenance et ainsi d'effectuer aisément toutes les modifications et ajustements que nous jugeons nécessaires. Nous avons pu ainsi utiliser une version MetaNetX actualisée (Version 4.1) et, combiné à l'utilisation de *parser* adaptés, nous avons pu intégrer, selon les normes de la ressource MIRIAM, les annotations de références croisées à l'aide d'URI (*Uniform Ressource Identifier*) respectant ainsi une syntaxe générique.

L'enrichissement des annotations en identifiants de métabolites et de réactions s'est effectué en deux étapes. Dans un premier temps, nous avons extrait du fichier `*.padmet` les identifiants provenant de plus d'une quinzaine de bases de données différentes pour les métabolites et les identifiants de KEGG, Rhea et UniProt pour les réactions. Dans un second temps, les fichiers MetaNetX ont été parcouru afin de récupérer, s'il existe, l'identifiant MetaNetX des entités qui servent ensuite d'ancrage pour retrouver l'identifiant de référence qui leur est associé (*i.e.* *« the best external resouce identifier »* provenant soit de BiGG, ChEBI, KEGG, ModelSEED, SABIO-RK ou SwissLipids pour les métabolites, soit de BiGG, KEGG, ModelSEED ou Rhea pour les réactions). Ces informations ont été présentées en **Tableau 3-4** (page 219).

Outre les problèmes de connexions inter-bases de données, parmi les problèmes intra-bases de donnés évoqués dans les travaux de Pham *et al.* *(2019)* décrivant ces (in)cohérences, deux principes illustrent ce phénomène. Le premier aspect concerne l'ambiguïté des noms (*i.e.* nombre moyen d'identifiants associés par nom de composé) et nous noterons que MetaCyc appartient aux bases de données de leur étude les moins concernées par ce problème. En revanche, concernant le second critère évalué, la multiplicité des identifiants et l'utilisation de synonymes (*i.e.* nombre moyen de noms de composé associés par identifiant), MetaCyc se trouve parmi les bases de données les plus problématiques. Un extrait centré sur MetaCyc de leurs résultats est présenté en **Tableau 3-5**.

Ainsi, la comparaison directe et naïve des noms des entités à l'aide d'algorithmes de chaînes de caractères est souvent insuffisante *(Bertrand et al. 2017)*. Il n'est pas rare que le nom d'une entité dans une base de données soit associés à plusieurs entrées dans une autre, que le même nom soit donné à divers composés, et ce, parfois même au sein de la même base de données ou que le même nom pointe sur des identifiants de molécules différents selon la base de données considérée. De surcroît, les composés génériques, à l'instar des lipides ou des glycanes possèdent des noms ambigus. Ainsi, et à bien des égards, mapper à partir des noms des entités est une mauvaise idée, c'est pourquoi nous nous sommes concentrés exclusivement sur la mise en correspondance à partir d'identifiants MetaNetX pour effectuer un pont entre toutes les références croisées.

Tableau 3-5 : Pourcentage d'identifiants de métabolites de la base de données MetaCyc qui trouve une correspondance dans les autres bases de données en utilisant comme point d'ancrage les noms des métabolites ou les identifiants MetaNetX. Résultats extraits des travaux de Pham *et al.* *(2019)*. Les versions de MetaCyc et MetaNetX utilisées dans cette étude sont légèrement antérieures à celles que nous avons utilisées. De manière générale, outre une confiance accrue dans les résultats du *mapping*, la mise en correspondance des identifiants à partir d'identifiants MetaNetX couvre une plus grande surface de l'espace des identifiants.

Base de données	*Mapping* à partir des Noms des entités	*Mapping* à partir des Identifiants MetaNetX
BiGG	4,1 %	33,9 %
ChEBI	16,3 %	9,3 %
HMBD	4,6 %	7,8 %
KEGG	2,1 %	18,6 %
Lipid Maps	1,6 %	5,1 %
Recatome	23,7 %	42,6 %
SABIO-RK	4,4 %	33,8 %
ModelSEED	27,8 %	42,8 %
SLM	0,1 %	0,1 %

Nous venons de constater que le premier verrou est lié à l'interconnexion des identifiants entre ressources. Le résoudre permet de s'affranchir des contraintes liées à l'utilisation d'une base de données unique, favorisant ainsi l'exploitation de l'ensemble des résultats en générant une plateforme de ressources uniformes. Le second aspect informatique à considérer concerne l'interopérabilité entre logiciels pour viser une utilisation et une réutilisation optimales des modèles.

En réalité le véritable critère de qualité pour évaluer la forme d'un modèle réside dans l'utilisation de standards, et à ce titre, MEMOTE constitue une porte d'entrée vers cette problématique. En travaillant sur l'interopérabilité et le peuplement des annotations, nous nous sommes intéressés *de facto* à la manière dont nous devions les encoder. Utiliser un format standard bien qu'évident et essentiel n'est pourtant pas trivial et les motivations qui nous ont poussé à explorer cette thématique résident dans la viabilité ultérieure d'***i*Prub22**.

2.1.3.3. Encodage au format SBML : conformité et standards

Un format uniforme assure l'exportabilité de la reconstruction, favorisant et garantissant sa compréhension par une multitude d'outils et de ressources existantes et à venir, facilitant par conséquent sa réutilisation future. Cependant, pour atteindre cet objectif, la condition *sine qua non* est d'avoir recours à des standards uniformes et uniques. Nous souhaitons insister fortement sur cette notion de « standards uniformes et uniques », car même si aujourd'hui une tendance forte semble se diriger vers l'emploi de **SBML** (version 3) *(Keating et al. 2020)* couplé à l'utilisation du package FBC2 *(Olivier et Bergmann 2018)*, il n'en a pas toujours été ainsi *(Dräger et Palsson 2014)*.

Concrètement, un fichier XML (*eXtensible Markup Language*) est un format de fichier texte utilisé pour stocker et échanger des données de manière structurée. Il utilise des balises pour définir les éléments et les attributs des données qu'il contient. Dans le domaine de la biologie des systèmes, un fichier **SBML** (*Systems Biology Markup Language*) est un type spécifique de fichier XML. Il est utilisé pour formaliser des modèles mathématiques de réseaux biologiques, décrivant les réactions biochimiques, les processus cellulaires et les interactions moléculaires à l'aide d'une syntaxe XML standardisée. Les composants d'un modèle **SBML** sont décrits à l'aide de classes, caractérisées par des attributs et pouvant contenir des classes imbriquées. Les attributs de données d'une classe mère sont alors hérités par les classes enfants. Par exemple, la classe `ListOfReactants` et la classe `ListOfProducts` sont des classes enfants de `ListOfReactions`. Dans le cadre d'***i*Prub22**, nous dénombrons sept classes principales :

- listOfFluxObjectives — *Liste du ou des objectifs de flux dans le modèle (i.e. biomass)*
- listOfGeneProducts — *Liste des produits géniques dans le modèle* **SBML**
- listOfUnitDefinitions — *Liste des unités de mesure utilisées dans le modèle* **SBML**
- listOfCompartments — *Liste des compartiments biologiques ou spatiaux du modèle* **SBML**
- listOfSpecies — *Liste des espèces chimiques dans le modèle* **SBML**
- listOfParameters — *Liste des paramètres du modèle* **SBML**
- listOfReactions — *Liste des réactions biochimiques ou biologiques dans le modèle* **SBML**

Bref historique des versions SBML et du package FBC

Le format **SBML** (*Systems Biology Markup Language*) est le résultat d'un effort collaboratif et évolutif de la communauté de la modélisation des systèmes biologiques pour fournir un langage de modélisation robuste et flexible pour la représentation des processus biologiques complexes. La première version de ce langage de balisage qui établissait un cadre standard en introduisant les concepts fondamentaux de **SBML**, tels que les espèces, les réactions et les compartiments, a été publiée en 2000. En 2010, le **SBML** Level 3 a été une refonte majeure de la spécification **SBML**, en introduisant une architecture modulaire permettant une extension facilitée avec de nouveaux packages. L'introduction des termes SBO est apparue lors de cette mise à jour et des attibuts tels que `id` et `names` sont devenus obligatoires pour décrire et identifier précisément les entités du modèle. Au fil des versions, l'encodage des annotations vers les références externes s'est tourné vers l'utilisation d'identifiants URI MIRIAM. Enfin, la dernière mise à jour du format date de 2020 (**SBML** Level 3 Version 2 Core Release 3).

Le package FBC (*Flux Balance Constraints*), déployé en février 2013, a été conçu pour résoudre les limitations rencontrées dans l'encodage des contraintes de flux, notamment par l'ajout des instances `<ListOfFluxBounds>` et `<ListOfObjectives>`. La version 2 de ce package, distribuée en septembre 2015, enrichit les possibilités offertes par une révision approfondie des caractéristiques liées aux produits géniques.

Les ajustements entrepris sur la modification du format de notre réseau (**Tableau 3-6**), résultants soit de la conversion du format FBC1 en FBC2, réalisée *via* un script disponible sur l'interface de programmation d'application LibSBML *(Bornstein et al. 2008)*, soit de nos propres modifications ont été intégralement contrôlés à l'aide d'un script de vérification rendu accessible à la communauté sur l'interface de programmation Python pour libSBML (*i.e.* `validateSBML.py` - test la syntaxe et la cohérence interne d'un fichier **SBML**). Procéder ainsi nous a permis de tracker et d'éliminer les erreurs de notre modèle ainsi que les messages d'alertes (*i.e.* warning). Désormais, le seul message qui subsiste concerne la détection d'un terme SBO non monitoré encore par l'outil, celui des réactions spontanées. De ce point de vue, notre réseau présente donc un format valide.

Tableau 3-6 : Ajustements et améliorations apportés au format d'*i*Prub22. Ce tableau répertorie les ajustements et améliorations réalisés pour correspondre aux critères **SBML** lvl 3 et FBC 2. Ces changements sont le résultat de l'exécution du script de conversion de FBC1 vers FBC2 disponible sur l'API libSBML Python (`conversion_fbc1_to_fbc2.py`), de modifications que nous avons apportées au script AuReMe (`sbmlGenerator.py`), ou de révisions manuelles.

Classe	Modifications	Conteneur	Références

Model	Ajout des attibuts `id`, `name`, `metaid` et `volumeUnits`	`<model>`	`SBML – section 3.2.1`
			`SBML – section 3.2.2`
			`SBML – section 3.2.3`
			`SBML – section 4.2.4`

	Ajout du terme `SBO:0000624`	`<model>`	`SBML – section 3.2.4`
			`SBML – section 4.2.1`
			`SBML – section 5`

	Ajout de la description du modèle	`<notes>`	`SBML – section 3.2.5`
	Ajout de l'orgine des données (références vers la taxonomie de l'organisme et l'assemblage)	`<annotation>`	`SBML – section 3.2.6` `SBML – section 6`
Objective	Ajout des classes `FBC Objective` et `FluxObjective`	-	`FBC1 – section 3.6` `FBC1 – section 3.7`
GeneProduct	Ajout de la classe `GeneProduct`	-	`FBC2 – section 3.5`
	Ajout du terme `SBO:0000243`	`<geneProduct>`	`SBML – section 3.2.4` `SBML – section 5`
	Ajout d'informations de localisation subcellulaire des produits géniques (`SignalP`, `TMHMM`, `DeepLoc`)	`<notes>`	`SBML – section 3.2.5`
	Ajout de références croisées	`<annotation>`	`SBML – section 3.2.6` `SBML – section 6`

Tableau 3-6 (suite)

UnitDefinitions	Ajout de la classe `UnitDefinitions` Défintion de l'unité mmol.gDW^{-1}.h^{-1}	-	SBML – section 4.4
Compartment	Ajout des attributs `metaid`, `spatialDimension` et `units`	<compartment>	SBML – section 3.2.3 SBML – section 4.5.2 SBML – section 4.5.4
	Ajout des termes `SBO:0000290` ou `SBO:0000410`	<compartment>	SBML – section 3.2.4 SBML – section 4.5.6 SBML – section 5 SBML – section 8.2.2
	Ajout de références croisées	<annotation>	SBML – section 3.2.6 SBML – section 6
Species	Ajout du terme `SBO:0000247`	<species>	SBML – section 3.2.4 SBML – section 4.6.2 SBML – section 5
	Ajout des attributs `charge` et `formule` ▲	<notes>	FBC1 – section 3.4 SBML – section 3.2.5
	Ajout de références croisées	<annotations>	SBML – section 3.2.6 SBML – section 6
▲ Initialement, lors du chargement du modèle sur MATLAB les informations relatives à la formule brute et à la charge des métabolites, bien qu'encodées convenablement comme en témoigne la section 3.4 de la documentation du package FBC, n'étaient pas reconnues, c'est pourquoi nous avons décidé de les stocker dans la balise `<notes>` pour pallier ce problème (*i.e.* patch).			
Parameters ■	Ajout de la classe `Parameters`	-	SBML – section 4.7
	Ajout des attributs `id`, `value`, `units`, `constant` et `name`	<parameter>	SBML – section 4.7.1 SBML – section 4.7.2 SBML – section 4.7.3 SBML – section 4.7.4 SBML – section 3.2.2
	Ajout des termes `SBO:0000625` ou `SBO:0000626`	<parameter>	SBML – section 3.2.4 SBML – section 4.7.5 SBML – section 5
■ Classe établissant la liste des paramètres équivalants aux contraintes définies dans le modèle lors de son chargement. Ces étiquettes sont utilisées pour la définition des attributs `lowerFluxBound` et `upperFluxBound` caractérisant les réactions. L'explication de la définition du paramètre est portée dans l'attribut `name` et de plus amples détails sur les étiquettes sont présentés dans le Livescript. Outre l'initiation des *uptakes* et l'ouverture des réactions de production (*i.e.* définition d'un environnement nutritif minimal permettant la production de biomasse et de métabolites spécialisés), diverses bornes inférieures et supérieures sont caractérisées, telles que les bornes par défaut pour les réactions non contraines réversibles ou non - `default_lb`, `default_ub`, `irrLow`, les bornes de la réaction NGAM - `NGAM_BOUND` et les bornes de la réaction d'échange de biomasse - `Exchange_Biomass_lb` et `Exchange_Biomass_ub`. Exemple d'encodage sur la réaction d'*uptake*_001 : `<parameter constant="true" id="ub_uptake_001" name="Uptake of (S)-lactate" sboTerm="SBO:0000625" units="mmol_per_gDW_per_hr" value="0"/>`			

Tableau 3-6 (suite et fin)

Reactions	Ajout du terme SBO:0000176 ●	<reaction>	SBML – section 3.2.4 SBML – section 4.11.1 SBML – section 5
	Ajout des attributs lowerFluxBound et upperFluxBound ■ et metaid	<reaction>	FBC2 – section 3.8 SBML – section 3.2.3
	Ajout de la classe geneProductAssociation	-	FBC2 – section 3.9
	Ajout de référence croisées	<annotations>	SBML – section 3.2.6 SBML – section 6

● Lors de la génération du **SBML**, un terme SBO par défaut est associé aux entités geneProduct, species et reactions. Il sert alors d'ancrage lors du *parsing* du fichier pour des modifications ultérieures (*e.g.* affinement des termes, réalisé actuellement uniquement sur les réactions et les gènes).

Sources :

SBML : http://SBML.org/specifications/SBML-level-3/version-2/core/release-2/
FBC1 : http://identifiers.org/combine.specifications/SBML.level-3.version-1.fbc.version-1.release-1
FBC2 : http://identifiers.org/combine.specifications/SBML.level-3.version-1.fbc.version-2.release-1

Nous proposons de visualiser, sur les trois pages suivantes (**Figure 3-12**), les résultats des ajustements évoqués en **Tableau 3-6** en présentant des extraits du fichier **SBML** avant et après les modifications pour l'encodage de la pénicilline, de la réaction conduisant à sa production et de son gène associé.

Encodage du gène PenDE à l'origine de la catalyse de la réaction de biosynthèse de la Pénicilline G

```
<fbc:geneProduct fbc:id="gp_Pc21g21370" fbc:label="Pc21g21370" metaid="gp_Pc21g21370"
sboTerm="SBO:0000243">
  <notes>
    <body xmlns="http://www.w3.org/1999/xhtml">
      <p>DeepLoc: Cytoplasm</p>
      <p>TMHMM: Topology=o</p>
    </body>
  </notes>
  <annotation>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"xmlns:bqbiol="http://
    biomodels.net/biology-qualifiers/">
      <rdf:Description rdf:about="#gp_Pc21g21370">
        <bqbiol:is>
          <rdf:Bag>
            <rdf:li rdf:resource="https://identifiers.org/ncbigene/8304967"/>
            <rdf:li rdf:resource="https://identifiers.org/ncbiprotein/XP_002569112"/>
            <rdf:li rdf:resource="https://identifiers.org/KEGG.genes/pcs:Pc21g21370"/>
            <rdf:li rdf:resource="https://identifiers.org/UniProt/B6HLT9"/>
            <rdf:li rdf:resource="https://identifiers.org/RefSeq/XP_002569112"/>
          </rdf:Bag>
        </bqbiol:is>
      </rdf:Description>
    </rdf:RDF>
  </annotation>
</fbc:geneProduct>
```

Encodage du métabolite Pénicilline G avant et après modifications du SBML

```
<species metaid="M_PENICILLIN__45__G_c" id="M_PENICILLIN__45__G_c" name="penicillin G"
compartment="c" initialAmount="0" hasOnlySubstanceUnits="false" boundaryCondition="false"
constant="false" fbc:charge="-1" fbc:chemicalFormula="C16H17N2O4S">
 <notes>
  <body xmlns="http://www.w3.org/1999/xhtml">
   <p>mass: 333.383</p>
   <p>smiles:[H][C@]12SC(C)(C)[C@@H](N1C(=O)[C@H]2NC(=O)Cc1ccccc1)C([O-])=O</p>
   <p>InChIKey: JGSARLDLIJGVTE-MBNYWOFBSA-M</p>
  </body>
 </notes>
 <annotation>
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dcterms="http://purl.org
  /dc/terms/" xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#" xmlns:vCard4="http://www.w3.org
  /2006/vcard/ns#" xmlns:bqbiol="http://biomodels.net/biology-qualifiers/" xmlns:bqmodel=
  "http://biomodels.net/model-qualifiers/">
   <rdf:Description rdf:about="#M_PENICILLIN__45__G_c">
    <bqbiol:isVersionOf>
     <rdf:Bag>
      <rdf:li rdf:resource="http://identifiers.org/InChI/InChI=1S/C16H18N2O4S/c1-16(2)12
      (15(21)22)18-13(20)11(14(18)23-16)17-10(19)8-9-6-4-3-5-7-9/h3-7,11-12,14H,8H2,1-2H3,
      (H,17,19)(H,21,22)/p-1/t11-,12+,14-/m1/s1"/>
     </rdf:Bag>
    </bqbiol:isVersionOf>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

```
<species boundaryCondition="false" compartment="c" constant="false" fbc:charge="-1"
fbc:chemicalFormula="C16H17N2O4S" hasOnlySubstanceUnits="false" id="M_PENICILLIN__45__G_c"
initialAmount="0" metaid="M_PENICILLIN__45__G_c" name="penicillin G" sboTerm="SBO:0000247">
 <notes>
  <body xmlns="http://www.w3.org/1999/xhtml">
   <p>formula: C16H17N2O4S</p>
   <p>charge: -1</p>
   <p>mass: 333.38326</p>
   <p>smiles: CC1(C)S[C@@H]2[C@H](NC(=O)Cc3ccccc3)C(=O)N2[C@H]1C([O-])=O</p>
  </body>
 </notes>
 <annotation>
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/" xmlns:vCard=http://www.w3.org/2001/vcard-rdf/3.0#
  xmlns:vCard4="http://www.w3.org/2006/vcard/ns#" xmlns:bqbiol="http://biomodels.net/biology-
  qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
   <rdf:Description rdf:about="#M_PENICILLIN__45__G_c">
    <bqbiol:is>
     <rdf:Bag>
      <rdf:li rdf:resource="https://identifiers.org/InChI/InChI=1S/C16H18N2O4S/c1-
      16(2)12(15(21)22)18-13(20)11(14(18)23-16)17-10(19)8-9-6-4-3-5-7-9/h3-7,11-12,14H,8H2,1-
      2H3,(H,17,19)(H,21,22)/p-1/t11-,12+,14-/m1/s1"/>
      <rdf:li rdf:resource="https://identifiers.org/InChIKey/JGSARLDLIJGVTE-MBNYWOFBSA-M"/>
      <rdf:li rdf:resource="https://identifiers.org/MetaNetX.chemical/MNXM1812"/>
      <rdf:li rdf:resource="https://identifiers.org/chebi/CHEBI:51354"/>
      <rdf:li rdf:resource="https://identifiers.org/BioCyc/META:PENICILLIN-G"/>
      <rdf:li rdf:resource="https://identifiers.org/KEGG.compound/C05551"/>
      <rdf:li rdf:resource="https://identifiers.org/chemspider/555781"/>
      <rdf:li rdf:resource="https://identifiers.org/hmdb/HMDB15186"/>
      <rdf:li rdf:resource="https://identifiers.org/metabolights/MTBLS51354"/>
      <rdf:li rdf:resource="https://identifiers.org/PubChem.compound/640429"/>
     </rdf:Bag>
    </bqbiol:is>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Encodage de la réaction de biosynthèse de la Pénicilline G avant et après modifications du SBML

```
<reaction id="R_RXN__45__17062" name="Pc21g21370_product" reversible="false" fast="false">
 <notes>
  <body xmlns="http://www.w3.org/1999/xhtml">
   <p>GENE_ASSOCIATION: (Pc13g09140) or (Pc21g21370)</p>
   <p>CATEGORIES: ANNOTATION</p>
   <p>SUBSYSTEM: PWY-7716</p>
  </body>
 </notes>
 <listOfReactants>
  <speciesReference species="M_CPD__45__207_c" stoichiometry="1" constant="false"/>
  <speciesReference species="M_6__45__AMINOPENICILLANATE_c" stoichiometry="1" constant="false"/>
 </listOfReactants>
 <listOfProducts>
  <speciesReference species="M_PENICILLIN__45__G_c" stoichiometry="1" constant="false"/>
  <speciesReference species="M_CO__45__A_c" stoichiometry="1" constant="false"/>
  <speciesReference species="M_PROTON_c" stoichiometry="1" constant="false"/>
 </listOfProducts>
</reaction>
```

```
<reaction fast="false" fbc:lowerFluxBound="irrLow" fbc:upperFluxBound="default_ub"
id="R_RXN__45__17062" metaid="R_RXN__45__17062" name="Pc21g21370_product" reversible="false"
sboTerm="SBO:0000176">
 <notes>
  <body xmlns="http://www.w3.org/1999/xhtml">
   <p>GENE_ASSOCIATION: (Pc13g09140) or (Pc21g21370)</p>
   <p>CATEGORIES: ANNOTATION</p>
   <p>SUBSYSTEM: PWY-7716</p>
  </body>
 </notes>
 <annotation>
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/" xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#"
  xmlns:vCard4="http://www.w3.org/2006/vcard/ns#" xmlns:bqbiol="http://biomodels.net/biology-
  qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
   <rdf:Description rdf:about="#R_RXN__45__17062">
    <bqbiol:is>
     <rdf:Bag>
      <rdf:li rdf:resource="https://identifiers.org/MetaNetX.reaction/MNXR120626"/>
      <rdf:li rdf:resource="https://identifiers.org/BioCyc/META:RXN-17062"/>
      <rdf:li rdf:resource="https://identifiers.org/seed.reaction/rxn47608"/>
     </rdf:Bag>
    </bqbiol:is>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
 <fbc:geneProductAssociation xmlns:fbc="http://www.SBML.org/SBML/level3/version1/fbc/version2">
  </fc:or>
   <fbc:geneProductRef fbc:geneProduct="gp_Pc13g09140" sboTerm="SBO:0000243"/>
   <fbc:geneProductRef fbc:geneProduct="gp_Pc21g21370" sboTerm="SBO:0000243"/>
  </fc:or>
 </fbc:geneProductAssociation>
 <listOfReactants>
  <speciesReference constant="true" species="M_CPD__45__207_c" stoichiometry="1"/>
  <speciesReference constant="true" species="M_6__45__AMINOPENICILLANATE_c" stoichiometry="1"/>
 </listOfReactants>
 <listOfProducts>
  <speciesReference constant="true" species="M_PENICILLIN__45__G_c" stoichiometry="1"/>
  <speciesReference constant="true" species="M_CO__45__A_c" stoichiometry="1"/>
  <speciesReference constant="true" species="M_PROTON_c" stoichiometry="1"/>
 </listOfProducts>
</reaction>
```

Figure 3-12 : Encodage des entités au sein du SBML avant et après modifications, illustration avec le produit génique PenDE, le composé pénicilline et sa réaction de biosynthèse. *Les numéros de lignes grisées indiquent les éléments ajoutés lors des modifications. Les balises « gene product » ont été intégrées lors de la conversion en FBC2. Elles ont été ensuite peuplées, tout comme celles relatives aux « species » et « reaction » à l'aide des diverses annotations en notre possession.*

Avec l'ensemble des éléments présentés jusqu'alors, nous pouvons ainsi considérer que le format de notre réseau est de qualité. En revanche, pour étendre cette définition à son contenu, il sera nécessaire de comparer ses capacités prédictives aux informations expérimentales (cf. sections *2.2.3 Modélisation de la croissance de l'organisme, la réaction de biomasse* page 277, et *3.4 Une production de métabolites spécialisés sensible aux conditions environnementales ?*, page 337).

2.1.3.4. Accessibilité et partage du modèle *i*Prub22

Avant de conclure, nous souhaitons poursuivre cette section consacrée à la « qualité » sur les notions d'accessibilité. Comme nous l'avons mentionné précédemment, ***i*Prub22** est disponible sur la base de données BioModels *(Malik-Sheriff et al. 2020)* sous le numéro d'accès : `MODEL2306150001` (https://www.ebi.ac.uk/biomodels/MODEL2306150001). Outre le fichier **SBML** de la reconstruction paramétrée ***i*Prub22** et le lien direct vers sa publication, nous avons également fourni le rapport MEMOTE, le LiveScript décrivant ses caractéristiques, tout deux au format HTML, ainsi qu'une analyse FROG intégrée dans un fichier OMEX (*i.e. Open Modeling EXchange format* – format standardisé facilitant l'échange des informations nécessaires à une expérience de modélisation et de simulation en biologie et favorisant ainsi la reproductibilité et l'interopérabilité des modèles *(Bergmann et al. 2014)* avec un résumé de cette analyse dans un fichier XLS. Procéder ainsi nous permet d'être conformes aux préconisations MIASE *(Waltemath et al. 2011)*. Les lignes directrices de ce concept illustrent la liste des « bonnes pratiques » à mettre en œuvre pour proposer une expérience de simulation qui soit correctement et facilement interprétable et reproductible. Enfin, le recours à un dépôt central présente de nombreux avantages : il facilite l'accès au modèle par autrui, simplifie sa diffusion et garantit sa pérennité grâce à un système de gestion de versions.

Le dernier aspect concernant la notion d'accessibilité que nous souhaitons évoquer concerne la potentielle mise en place d'un wiki pour notre reconstruction. La Toolbox AuReMe intègre la technologie MediaWiki et nous offre ainsi la possibilité de générer et de déployer des pages wiki pour notre réseau. Cette solution de visualisation, une thématique délicate dans la gestion des **GSMNs**, permet de regrouper les données selon une structure de navigation cohérente en fournissant, entre autres, des détails de la reconstruction sur l'origine des données et des liens hypertextes vers des ressources externes telles que les pages des réactions ou des métabolites sur MetaCyc, la base de données de référence de la reconstruction, ou sur d'autres bases de données. Ainsi, en hébergeant ces fichiers sur un serveur web, nous pourrions offrir une plateforme collaborative permettant le partage, la création et la modification des données d'***i*Prub22**, favorisant alors une construction collective et évolutive.

2.1.3.5. Conclusion : point clé à retenir

L'ensemble des informations que nous venons de présenter avait pour objectif de mettre en lumière les aspects clés et les subtilités sous-jacentes à l'expression trop souvent utilisée de « **GSMN** de haute qualité ». Même si des pistes d'amélioration subsistent, nous avons cherché, tout au long du travail présenté dans ce document, à nous conformer aux normes établies par la communauté de modélisation et à être aussi transparents que possible. Enfin, nous désirons conclure nos réflexions sur une illustration (**Figure 3-13**) issue du travail de Kim *et al.* *(2018)*. « *Experimenting with reproducibility: a case study of robustness in bioinformatics* » synthétisant et résumant une fois de plus les enjeux que nous venons d'aborder.

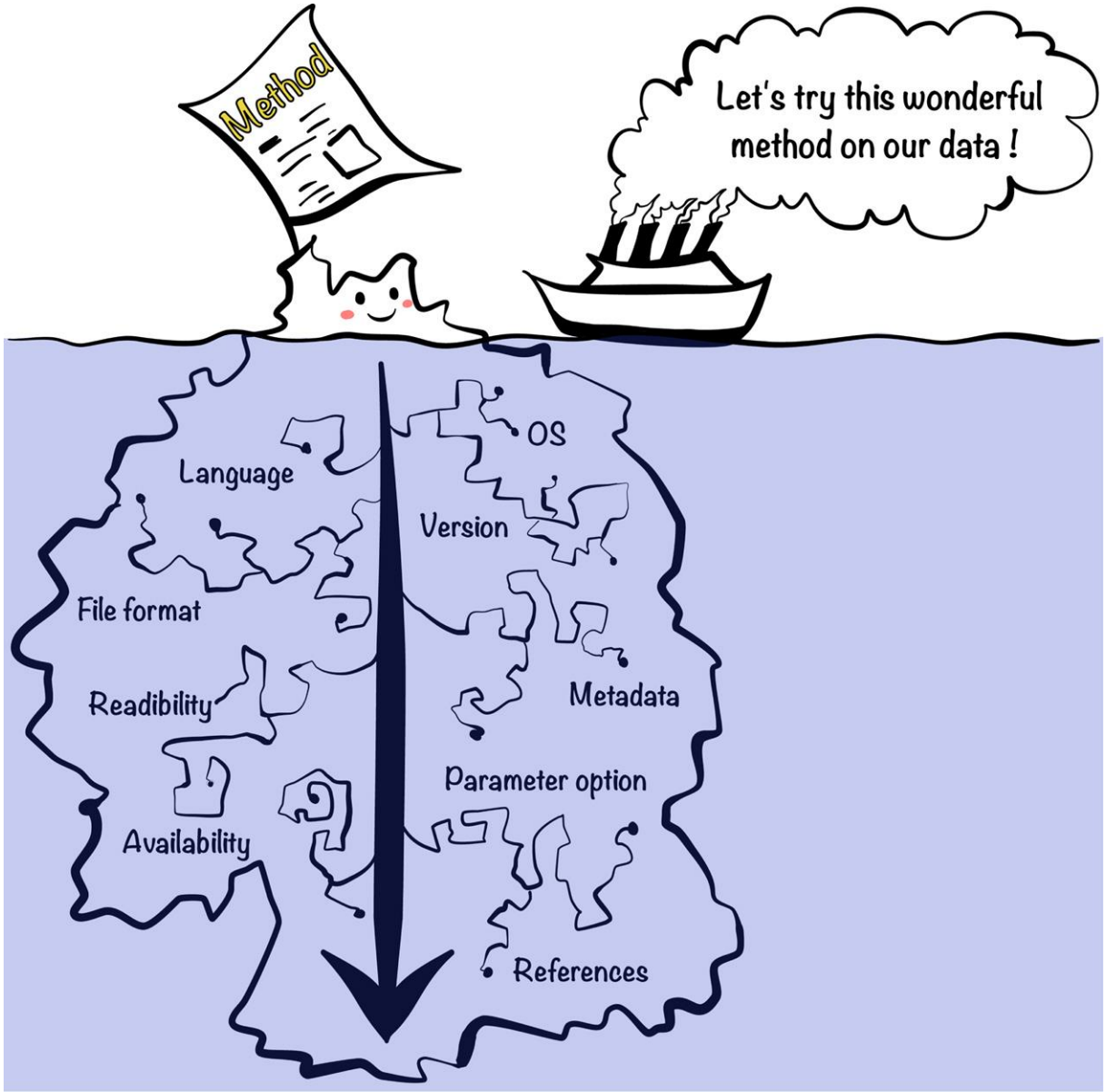


Figure 3-13 : « Les problèmes de reproductibilité cachés sont comme un iceberg sous-marin. *Les lecteurs de journaux scientifiques ont l'impression de pouvoir presque visualiser l'ensemble du travail impliqué dans la méthode. En réalité, les articles ne tiennent pas compte des ajustements et de la configuration nécessaires pour une réplication significative dans la plupart des cas. Par conséquent, il existe un écart significatif entre le travail exécutable apparent (i.e. la partie émergée de l'iceberg) et l'effort nécessaire en pratique (i.e. l'ensemble de l'iceberg).» Figure et légende issues et traduite des travaux de (Kim et al. 2018).*

2.2. Les points sujets à question

Nous venons d'évoquer certains des principes clés du concept FAIR. Néanmoins, la reproductibilité intégrale d'un **GSMN** est illusoire. Outre le fait que ces modèles représentent les connaissances à un instant donné – nous verrons ultérieurement que la mise à jour de certaines ressources utilisées complique grandement la génération de nouveaux **GSMNs** – la génération d'une reconstruction puis des modèles qui en découlent résulte de choix parfois peu ou pas documentés de la part du constructeur. Nous nous sommes efforcés de documenter au mieux ces partis pris, et c'est pourquoi, afin d'assurer la transparence de notre reconstruction, nous souhaitons revenir en détail sur trois étapes cruciales qui nous ont conduits à transformer notre draft initial en **GSMN** de haute qualité : le *gap-filling*, la mise au point de différents *scenarii* de nutrition et la conception de la réaction de biomasse.

Par ailleurs, ces trois points, indiqués en gras ci-dessous, font partie des cinq sources de questionnement définies par Bernstein *et al. (2021)* dans leurs travaux sur les incertitudes liées à la reconstruction et à l'analyse des **GSMNs** (**Figure 3-14**) :

- Annotation du génome
- **Spécification de l'environnement**
- **Formulation de la réaction de biomasse**
- **Comblement des lacunes du réseau**
- Simulation des flux

Notons que les réflexions portant sur l'annotation du génome, et plus généralement sur l'extraction des données à partir des séquences génomiques, sont présentées en Annexe 1 : *Reconstruction d'iPrub22 – sous-réseau issu de l'annotation fonctionnelle*, page 425, Annexe 2 : *Reconstruction d'iPrub22 – sous-réseau issu des recherches d'orthologie*, page 433 et en Annexe 3 : *Reconstruction d'iPrub22 – fusion des données et première ébauche de réseau (draft)*, page 445. Enfin, nous aurons l'occasion de revenir sur les choix liés à la simulation des flux dans la section *3 - Focus sur le métabolisme spécialisé : descriptions de diverses voies de biosynthèse de métabolites spécialisés*, page 308.

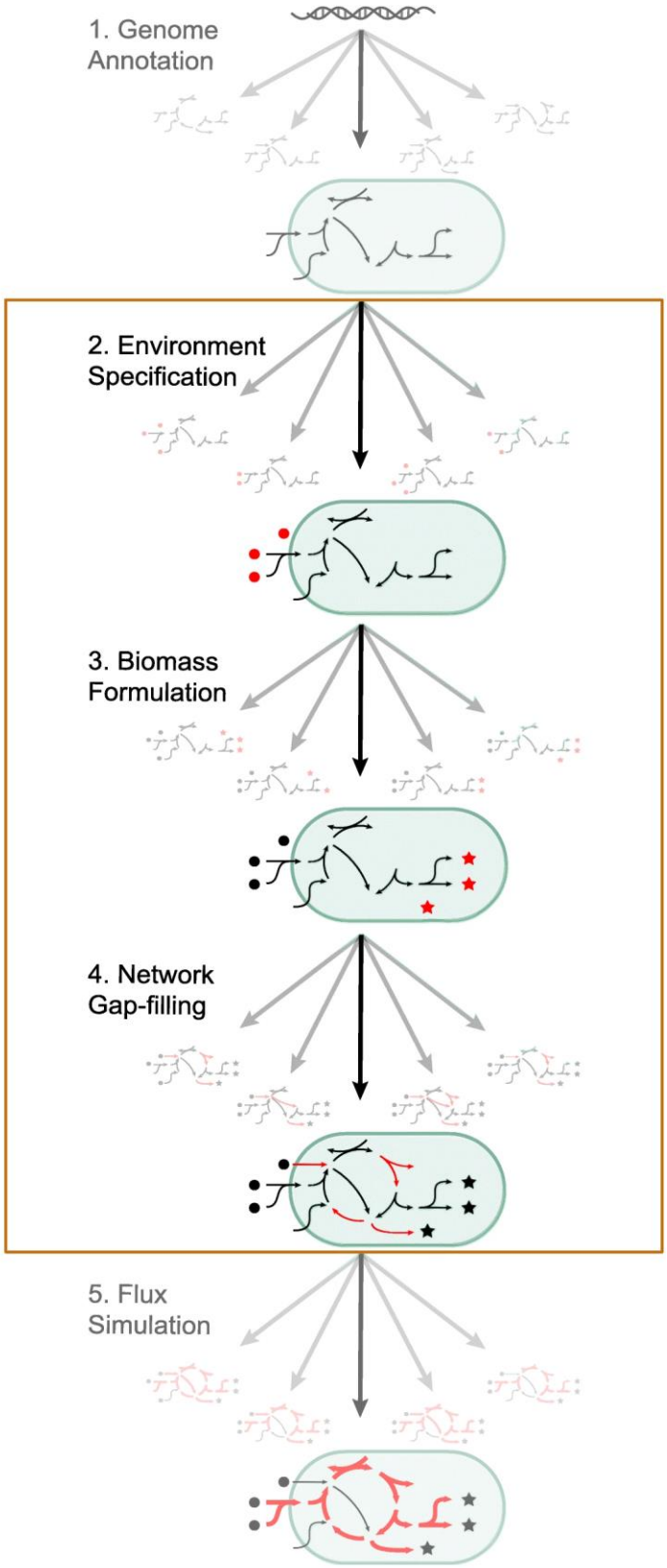


Figure 3-14 : « La progression générale de la reconstruction et de l'analyse d'un modèle métabolique à l'échelle du génome est représentée par cinq étapes principales. *Les flèches noires centrales illustrent une approche standard, qui produit un seul résultat à chaque étape. Les flèches grises représentent l'incertitude dans ce processus, le résultat de chaque étape étant un ensemble de résultats possibles. Les nouvelles additions au modèle à chaque étape sont indiquées en rouge, les cercles représentent les métabolites, les étoiles représentent les composants de la biomasse, les flèches représentent les réactions métabolique, et les flèches en gras représentent une distribution de flux spécifiques. » Figure et légende issues et traduite de (Bernstein et al. 2021). Dans cette section, nous nous concentrons exclusivement sur les trois étapes encadrées en* ***orange****.*

2.2.1. UN *GAP-FILLING* « RAISONNÉ »

2.2.1.1. Principes et données d'entrée

Au sein du draft, les capacités métaboliques du modèle sont inférées, soit par l'annotation fonctionnelle, soit par les liens d'orthologie, soit par les informations présentes dans les anciennes reconstructions. En conséquence, à ce stade de la reconstruction et en raison de ses propriétés inhérentes, seules les réactions catalysées par des protéines enzymatiques, et donc soutenues par au moins un gène, sont présentes dans le réseau.

En raison de notre connaissance partielle de chaque organisme, les reconstructions de réseaux métaboliques présentent un nombre variable de lacunes. Ces lacunes, généralement indicatrices de réactions manquantes, se manifestent souvent par la présence d'impasses métaboliques, couramment désignées sous le terme de *dead-end* *(Orth et Palsson 2010)* (cf. **Figure 3-15** et encart : *Réduction du nombre de dead-ends et réactions bloquées*, page 240). Des réactions métaboliques ou des réactions de transport sont alors ajoutées à travers des étapes de *gap-filling* afin de connecter les métabolites isolés à des réactions existantes ou d'introduire de nouvelles voies métaboliques.

La réalisation du *gap-filling* dans notre réseau métabolique a été effectuée à l'aide de Meneco *(Prigent et al. 2017)*. Les résultats fournis par cet outil découlent d'une heuristique, une méthode de résolution non fondée sur un modèle formel et guidée par des règles empiriques. Il est important de souligner que les solutions proposées par Meneco ne sont pas garanties comme optimales et sont fortement influencées par les paramètres d'entrée.

Théoriquement, il est crucial de minimiser le nombre de réactions issues du *gap-filling* afin d'éviter un risque de sur-ajustement du modèle (*i.e.* « *over-fitting* ») et l'introduction de faux positifs. Il convient de noter qu'à l'exception d'une étude manuelle et systématique de chaque réaction, il n'existe pas de critères universels permettant de déterminer la pertinence des résultats. De surcroît, et par définition, l'ajout de telles réactions à la reconstruction n'est pas étayé par une séquence génomique. Par conséquent, la traçabilité de l'origine de ces réactions est essentielle pour garantir la qualité de la reconstruction.

Sur ce dernier point, lors de l'intégration de nouvelles réactions, AuReMe offre la possibilité d'ajouter des commentaires aux opérations effectuées sur le réseau. Ainsi, afin d'assurer la traçabilité des réactions ajoutées par *gap-filling* et de justifier leur présence, nous avons mentionné l'origine de chaque réaction en fournissant le nom de l'outil employé (*i.e.* Meneco), la liste de cibles, le réseau de réparation utilisé et, lorsqu'il était clairement identifié, l'impact de la réaction sur la topologie du draft. Cependant, ces commentaires ne sont pas inclus dans la version finale d'***i*Prub22** (*i.e.* fichier **SBML**). Ils sont toutefois conservés et disponibles en tant qu'informations supplémentaires de l'article (S5_file.xls) *(Nègre et al. 2023)*. Pour des raisons de lisibilité, les réactions présentes dans le classeur Excel susmentionné sont classées par listes de cibles.

Réduction du nombre de *dead-ends* et réactions bloquées

Une impasse métabolique se définit comme un métabolite présentant une discontinuité topologique résultant de l'absence de réactions métaboliques ou de transport associées. Elle diminue la fidélité du modèle en influençant d'une part, la capacité du système métabolique à équilibrer les flux de métabolites, et d'autre part, à optimiser l'utilisation des substrats. Ainsi, le processus de *gap-filling* s'intègre pleinement dans l'effort de curation visant à améliorer la qualité du modèle. Cependant, une vigilance accrue doit être exercée lors des ajouts car biologiquement une lacune peut également signifier, par exemple, une perte de fonctionnalité au cours de l'évolution ou l'existence d'un pathway alternatif.

Les *dead-ends* se divisent en deux catégories : d'une part, les métabolites uniquement consommés, également appelés « *root no-production metabolites* » (*i.e.* correspondant dans un graphe aux sources) et d'autre part, les métabolites uniquement produits, également nommés « *root no-consumption metabolites* » (*i.e.* correspondant dans un graphe aux puits).

Lors de la modélisation des flux dans un modèle, (*e.g.* FBA, FVA), l'intégralité des voies consommant ou produisant une *dead-end* sont dites bloquées. Cette condition découle de l'état d'équilibre qui interdit l'accumulation de composés au sein d'un système. Ainsi, réduire le nombre de ces impasses constitue un objectif essentiel lors de la curation manuelle et représente également une piste d'amélioration continue. La détection de ces impasses s'opère généralement à travers l'analyse de la connectivité des métabolites dans la matrice stœchiométrique du réseau métabolique, en recherchant, par exemple, les métabolites pour lesquels la somme des degrés entrant ou sortant est nulle.

La résolution des impasses s'effectue majoritairement par *gap-filling*, un processus de comblement de lacunes qui vise à améliorer la reconstruction en agissant sur la connectivité topologique, induisant une diminution du nombre de *dead-end* et débloquant ainsi quelques portions du modèle. Le *gap-filling* permet de connecter des voies métaboliques auparavant disjointes en raison, par exemple, de la présence d'une réaction orpheline.

Ainsi, lors de l'ajout de réactions au cours des phases de curation manuelle, nous ciblons préférentiellement les réactions biochimiques, les réactions de transport (*e.g.* communication entre les milieux intracellulaire et extracellulaire) puis les réactions de modélisation telles que les réactions d'échange (*e.g.* import d'une source de nutriment ou sécrétion d'un produit) et les réactions appelées *demand* ou *sink* réactions.

Ces réactions de modélisation utilisent et recyclent les métabolites qui s'accumulent dans le modèle. Elles établissent donc une connexion entre le pool de métabolites intracellulaires et l'extérieur du système (*i.e.* simulation de la membrane plasmique cellulaire). Par exemple, si le métabolite considéré est effectivement le produit final d'une voie métabolique, l'ajout d'une réaction d'export est nécessaire pour débloquer le modèle.

De manière similaire, si un métabolite est uniquement consommé, soit il appartient au milieu nutritionnel et auquel cas, nécessite l'ajout d'une réaction d'*uptake*, soit sa réaction de production demeure inconnue et une réaction de *demand* ou de *sink* devra alors être intégrée. Les *demand* réactions, ajoutées de manière irréversible au modèle, sont destinées à consommer les métabolites déposés dans le système. Enfin, les métabolites produits dans des réactions inconnues ou ceux définis en dehors du cadre du système sont introduits dans le réseau *via* des réactions réversibles d'absorption, également appelées *sink* réactions. Toutefois, leur nombre est à limiter et leur utilisation est considérée comme une option de dernier recours en raison de son caractère artificiel dans la modélisation.

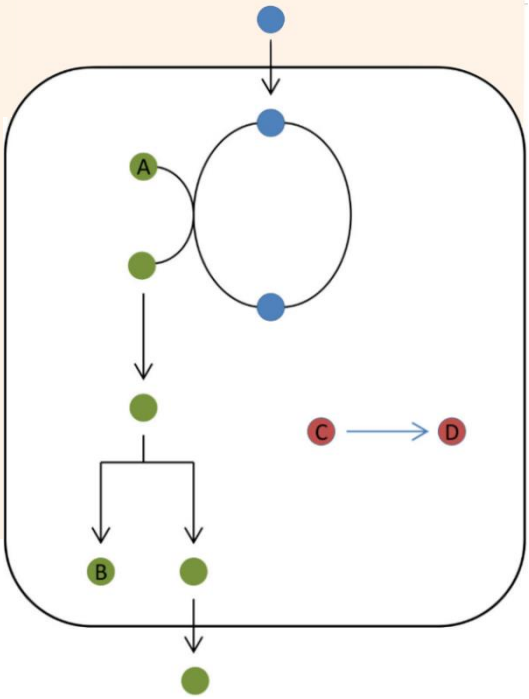


Figure 3-15 : « Représentation générique des métabolites dead-end au sein d'un réseau métabolique. *Les composés étiquetés A et C ne sont ni produits ni transportés par aucune autre réaction dans le réseau, tandis que les composés B et D ne sont ni consommés ni transportés par aucune autre réaction dans le réseau ». Figure extraite de (Mackie et al. 2013).*

En tenant compte des différents points mentionnés ci-dessus et en accordant une attention particulière à la nature et à la composition des données fournies à Meneco - quatre types d'inputs décrits ci-après - nous avons effectué un *gap-filling* que nous qualifions de « raisonné ».

- **Le réseau à compléter** (DRAFT)

Le *gap-filling* constitue un processus répétitif où, à chaque cycle, notre draft évolue et est actualisé. La transition du draft vers le **GSMN** est un processus complexe et chronophage. Pour apporter de la clarté aux multiples itérations du *gap-filling* et pour évaluer l'impact des différents traitements améliorant la connectivité, nous avons linéarisé le processus en six phases distinctes. Les résultats de ces itérations sont synthétisés en **Tableau 3-1** (page 141) et en **Figure 3-16**.

Figure 3-16 : Linéarisation du protocole de gap-filling (de a à q). *À chaque modification du draft (i.e. résultat positif du gap-filling conduisant à l'ajout d'au moins une réaction), nous recommençons le proces8sus à partir de la première étape (a). L'ordre d'exécution est défini en fonction de la confiance et de la spécificité décroissante des listes de cibles et des réseaux de réparation.*

- **Les composés d'initiation** (*SEEDS*)

Lors de la sélection de ces composés, notre objectif principal était de modéliser l'environnement nutritionnel de l'organisme. Ainsi, pour établir la liste des *seeds*, nous nous sommes essentiellement focalisés sur les métabolites présents dans le compartiment extracellulaire afin de refléter au mieux la réalité et les contraintes biologiques. En outre, pour maintenir une cohérence entre les analyses topologiques et les analyses de flux, la liste des composés d'initiation est similaire à la liste des composés liés à une réaction d'*uptake* (cf. *2.2.2 Modélisation de différentes conditions de culture*, page 256). À noter que pour répondre aux besoins de la modélisation, notamment pour l'activation des cycles, quelques artefacts, tels que des cofacteurs et des molécules énergétiques, ont été ajoutés à cette liste de *seeds*. Enfin, bien que l'établissement de cette liste soit déterminant pour les résultats du *gap-filling*, nous y avons, à tort, ajouté trois composés (cf. encart : *Les risques de la cuaration manuelle - Des intrus se sont glissés dans la liste des composés d'initiation*, page 245).

- **Les cibles** (TARGETS)

Comme mentionné dans la section « 1.3.1.6 *Target selection for model enhancement* » (page 139), nous avons constitué différents ensembles de cibles que nous avons classés en fonction de leur pertinence :

> 🎯 **243 Métabolites** issus de la recherche Bibliographique - Targets 1
>
> Ensemble de métabolites constitutifs ou spécialisés, dont la présence chez *P. rubens* est établie soit par des données bibliographiques provenant d'approches expérimentales, soit par leur nécessité dans le processus de production de biomasse, soit par leur étude approfondie chez cet organisme.
>
> 🎯 **35 Métabolites** sélectionnés dans les Bases de Données spécialisées - Targets 2
>
> Ensemble de métabolites issus de LOTUS *(Rutz et al. 2022)*, une banque de données classant les métabolites spécialisés par taxonomie. Seuls les métabolites associés avec certitude à la souche *P. rubens* Wisconsin 54-1255 ont été sélectionnés. La correspondance entre les identifiants LOTUS et MetaCyc a ensuite été réalisée en utilisant la première section des InChIKey.
>
> 🎯 **400 Métabolites** spécialisés selon les Ontologies de MetaCyc - Targets 3
>
> Ensemble de métabolites possédant par extension une annotation d'intérêt sous MetaCyc. Nous avons commencé par extraire l'ensemble des réactants et produits des réactions appartenant à des voies métaboliques annotées avec le terme "*Secondary*" dans MetaCyc (*i.e.* 742 pathways étiquetés "*Secondary Metabolite Biosynthesis*" et 63 pathways désignés "*Secondary Metabolite Degradation*"). Ensuite, nous avons réalisé un *mapping* entre ces composés et les métabolites du draft initial. La liste résultante a finalement été filtrée pour exclure les molécules énergétiques, les cofacteurs et autres "*hubs*" (*i.e.* composés hautement connectés tels que l'eau, l'oxygène, ou le phosphate inorganique).
>
> 🎯 **4 959 Métabolites** issus du draft – Targets 4
>
> Ensemble des métabolites contenus dans le draft initial. Cependant, compte tenu de l'objectif de minimiser l'ajout de réactions issues du *gap-filling*, la question se pose : pourquoi utiliser tous les métabolites du draft comme cibles ? À première vue, cette approche peut sembler contre-intuitive, voire brutale. Entreprendre un *gap-filling* complet, c'est-à-dire utiliser l'ensemble des métabolites supposés présents dans notre reconstruction par rapport à l'intégralité de la base de données MetaCyc est, d'un point de vue biologique et évolutif, absurde. Cependant, les étiquettes « seulement consommé », « seulement produit » ou « consommé et produit » ne tiennent compte que des voisins directs du métabolite étudié. Ainsi, un métabolite annoté « consommé et produit » peut nécessiter des étapes de *gap-filling* pour assurer sa productibilité topologique. Pour ces raisons, nous avons appliqué cette approche uniquement sur des réseaux de réparation réduits, impliquant des réactions spontanées et/ou des réactions orphelines.

- **Les réseaux de réparation** (*REPAIRNET*)

Pour réaliser les étapes de *gap-filling*, **MENECO** utilise un réseau de réparation afin de compléter les lacunes détectées dans le réseau interrogé. En forçant l'outil à utiliser préférentiellement des réactions jugées plus pertinentes que d'autres, cela nous permet d'accroître le taux de confiance dans les réactions que nous ajoutons au réseau. Ainsi, avant d'utiliser l'intégralité de MetaCyc (v23.0) nous créons nos propres réseaux de réparation à partir de la topologie de réseau évalué à l'instant t, auquel nous ajoutons un ou plusieurs sous-ensemble de MetaCyc d'intérêt pour la recherche de nouvelle cibles à débloquer (**Figure 3-16**).

Dans un premier temps, nous contraignons et réduisons l'espace des solutions à un ensemble de réactions indépendantes des informations génomiques en ciblant les réactions spontanées (cf encart *Exemple d'application sur les réactions spontanées*, page 244) et les réactions orphelines (*i.e.* enzyme inconnues). À l'inverse d'une réaction catalysée par une enzyme qui consomme de l'énergie (*i.e.* réactions endergoniques), une réaction spontanée libère de l'énergie (*i.e.* réactions exergoniques) et à ce titre peut se produire naturellement (*i.e.* évolution vers un état de plus grande stabilité ou de moindre énergie, se produit généralement lorsque les produits de la réaction sont moins énergétiques que les réactifs). Une réaction orpheline, quant à elle, est une réaction biochimique dont la présence dans l'organisme étudié est avérée, mais qui est catalysée par un produit génique dont l'origine demeure inconnue. La comparaison entre les capacités de production du réseau et des phénotypes caractéristiques de l'organisme, telles que l'absorption ou la sécrétion d'un substrat particulier, ou la présence d'une interruption dans une voie métabolique conservée, constitue des indicateurs de l'existence de telles réactions au sein des reconstructions. Par définition, ces deux types de réaction ne peuvent être présents dans notre draft.

Nous utilisons ensuite, des ensembles de réactions MetaCyc qui ont lieu chez d'autres champignons au sens large. Nous disposons à ce titre d'un méta-réseau métabolique de champignons *(Belcour et al. 2022)* fourni par l'équipe Dyliss à Rennes et de trois PGDBs disponibles sur BioCyc. Enfin, en dernier recours et uniquement pour l'utilisation de cibles de haut degré de confiance (*i.e.* Targets 1 et Targets 2), nous interrogeons l'ensemble de la base MetaCyc.

Exemple d'application sur les réactions spontanées

L'étude et la modélisation du métabolisme ont longtemps privilégié la caractérisation des réactions enzymatiques, négligeant souvent les réactions non-enzymatiques, pour lesquelles il existe peu de ressources. Pourtant, le rôle de ces réactions spontanées est essentiel. Elles permettent, entre autres, de renseigner sur l'évolution des voies métaboliques, de détecter de nouveaux composés et d'identifier les sources potentielles de perte de carbone et de toxicité *(Keller et al. 2015; Jeffryes et al. 2021)*. Ainsi, afin de compléter notre reconstruction, nous avons adopté deux approches distinctes pour détecter les réactions spontanées à intégrer. Tout d'abord, nous avons extrait les réactions annotées comme spontanées à partir des réseaux métaboliques antérieurs **(1)**. Ensuite, lors du processus de *gap-filling*, nous avons accordé la priorité à l'intégration de ce type de réactions en interrogeant un sous-ensemble de la base de données MetaCyc **(2)**.

(1) En faisant abstraction de la compartimentation, l'analyse de *i*AL1006 a révélé la présence de 17 réactions spontanées. Suite à une recherche de correspondances, soit *via* un *mapping* avec MetaNetX lorsqu'un identifiant KEGG était disponible, soit par l'identification des métabolites constituant la réaction, nous avons identifié 9 réactions sur MetaCyc. Parallèlement, l'exploration du PGDB PchCyc a indiqué l'existence de 18 réactions étiquetées comme spontanées. Même si cela peut sembler anecdotique, il est à noter que l'une des réactions issue de PchCyc possède un identifiant obsolète sur MetaCyc et qu'un total de six réactions sont de fausses réactions spontanées. En effet, elles étaient d'ores et déjà présentes dans le draft signifiant de ce fait qu'elles sont associées à des gènes. Concernant Prubens, bien que nous ne disposions pas d'annotations permettant de classifier et d'isoler facilement les réactions spontanées, sur les 2 574 réactions appartenant à ce réseau, seules 113 d'entre elles ne sont pas associées à une séquence génomique. En comparant ces données avec les réactions annotées comme spontanées sur MetaCyc, nous avons pu en isoler 15. En conclusion, ces trois sources externes ont permis d'identifier 19 réactions d'intérêt qui ont été ajoutées à la reconstruction.

(2) La seconde option utilisée pour détecter les réactions spontanées à ajouter à la reconstruction métabolique reposait sur l'utilisation de l'outil de *gap-filling* Meneco et des « réseaux de réparation restreints ». Ces derniers ont été reconstruits à partir de la topologie des drafts successifs et des réactions de MetaCyc détectées à partir des options de filtration disponibles (*i.e.* filtration sur l'expression « *Spontaneous reactions for which no enzyme is required* »). En octobre 2021, 705 réactions correspondaient à ce critère. Comme la sélection des targets et du réseau de réparation nous assurait que les réactions soulevées étaient automatiquement liées au draft, nous avons ciblé spécifiquement et exclusivement les réactions spontanées qui ont amélioré la connectivité du draft. Cette procédure semi-automatique a permis de sélectionner 67 réactions en un minimum de temps de traitement.

Conformément aux recommandations de Palsson *et al. (2010)*, seules les réactions dont au moins une des entités (*i.e.* réactif ou produit) est déjà présente dans le draft ont été ajoutées à la reconstruction. Notons également que, contrairement à Prubens, *i*AL1006 et PchCyc fournissent des annotations sur les réactions qui aident à leur classification. Ce point met en exergue la nécessité de labeliser les réactions en fonction de leur type pour faciliter l'exploration des données dans ce type de modélisation. Ainsi, afin de discrétiser plus aisément nos réactions spontanées de celles provenant d'un *gap-filling* classique, un gène artificiel de la forme s001 à s086 a été attribué pour chacune d'entre elles ainsi que le terme SBO correspondant aux réactions spontanées. En principe, pour également discriminer les réactions orphelines, nous aurions pu appliquer le même raisonnement. Cependant, à notre connaissance, il n'existe pas de termes pour désigner ce type de réaction.

In fine, l'agrégation des données sélectionnées permet d'ajouter un total de 86 réactions spontanées à la reconstruction, représentant ainsi 17 % des 510 réactions ajoutées lors du *gap-filling*.

2.2.1.2. Exemple d'illustration avec la voie de biosynthèse de l'ergotamine

Pour illustrer le processus semi-automatique de *gap-filling* et accompagner les réflexions liées à la curation manuelle, nous examinons de près la voie de biosynthèse de l'ergotamine (**Figure 3-17**). Bien que cet exemple ne soit pas le plus représentatif des résultats du *gap-filling*, il nous permet néanmoins d'illustrer son fonctionnement et de mettre en lumière les défis fondamentaux dans ce type de démarche : le choix des données d'entrée et le traitement des résultats. Enfin, la question qui sous-tend un résultat obtenu par *gap-filling* est simple : devons-nous, ou non, conserver cette information ? En revanche, nous verrons que la réponse à apporter à cette question est bien plus complexe.

Les risques de la cuaration manuelle - Des intrus se sont glissés dans la liste des composés d'initiation

Comparé aux processus automatiques, le risque majeur de la curation manuelle réside dans l'introduction d'incohérences liées au facteur humain, un paramètre d'autant moins négligeable lorsque les étapes sont longues et répétitives. Néanmoins, la transparence sur l'acquisition et l'intégration des données permet de détecter ces erreurs, car la traçabilité autorise le *back-tracking* des informations.

Cette affirmation est illustrée ci-dessous par l'analyse de l'incidence et de la portée d'une erreur détectée, *a posteriori*, lors de l'étude des voies de biosynthèse de métabolites spécialisés dans l'élaboration de la liste des *seeds* (*i.e.* voies de biosynthèse de la versicolorine B et de l'aflatoxine - cf. **Figure 3-40**, page 333).

Avant de fixer notre liste de composés d'initiation, divers tests de *gap-filling* et d'analyses topologiques avec la suite d'outils Mene Tools (*e.g.* recherche des réactions actives et des métabolites productibles) ont été effectués. Ils visaient à affiner la réflexion autour de la notion de *gap-filling* raisonné et à rechercher, par exemple, des impasses uniquement consommées à débloquer en priorité pour améliorer la connectivité globale du réseau. Cependant, sur les 185 composés d'initiation que nous avons retenus et utilisés, trois d'entre eux ont été malencontreusement conservés lors des phases de nettoyage et sont donc, à ce titre, des intrus.

En conséquence, cette erreur a eu des répercussions sur les réactions de modélisation puisque les réactions d'importation de métabolites ajoutées à la reconstruction pour la simulation des modèles sont issues de cette liste de *seeds*. Ainsi, nous avons constaté que la dioscine, le 6-Déméthylstérigmatocystine et le lysergate ont été indûment ajoutés à la liste des *seeds*, rendant respectivement inappropriés les couples de réactions `Uptake_053/Demand_007`, `Uptake_008/Demand_001`, et `Uptake_115/Demand_005`. Au sein du draft, les trois composés décrits ci-dessous sont des impasses uniquement consommées.

La dioscine est une saponine stéroïde impliquée dans la biosynthèse des stéroïdes et étudiée pour ses propriétés pharmacologiques potentielles. Elle est essentiellement produite chez les *Dioscoreaceae* (*e.g.* ignames), une famille de plantes monocotylédones. Sa présence, de prime abord surprenante, découle d'une réaction (`RXN-14574`) issue de l'annotation fonctionnelle du génome qui est soutenue par quatre gènes. Étant donné que ce composé ne possède pas de réaction de production répertoriée dans MetaCyc et que le 3-O-*β*-D-glucosyl-diosgenin, produit de cette réaction, est également une impasse métabolique (*i.e.* absence de réaction de consommation sous MetaCyc), l'ajout de la dioscine dans la liste des *seeds* n'a eu aucune incidence sur les résultats du *gap-filling*.

En revanche, concernant le 6-Déméthylstérigmatocystine et le lysergate, les résultats du *gap-filling* ont été quelque peu biaisés, car ces composés sont tous deux impliqués dans la biosynthèse de composés appartenant à nos listes de cibles. Le lysergate est un alcaloïde « précurseur » de l'ergotamine (Targets 1), tandis que le 6-Déméthylstérigmatocystine est une mycotoxine impliquée dans la voie de la stérigmatocystine (Targets 1) elle-même impliquée dans celle de l'aflatoxine (Targets 3). Ainsi, les avoir utilisés à tort comme composés d'initiation implique que ces impasses ne peuvent être résolues et qu'il existe des chaînons manquants pour la biosynthèse de chacun de ces composés.

Afin de visualiser les conséquences de l'ajout de *seeds* non pertinentes sur les résultats du *gap-filling*, la voie de biosynthèse de l'ergotamine est présentée en **Figure 3-17**. Celle de l'aflatoxine sera discutée dans la section *3.2 Reconstruire les voies de biosynthèse des métabolites spécialisés*, page 318.

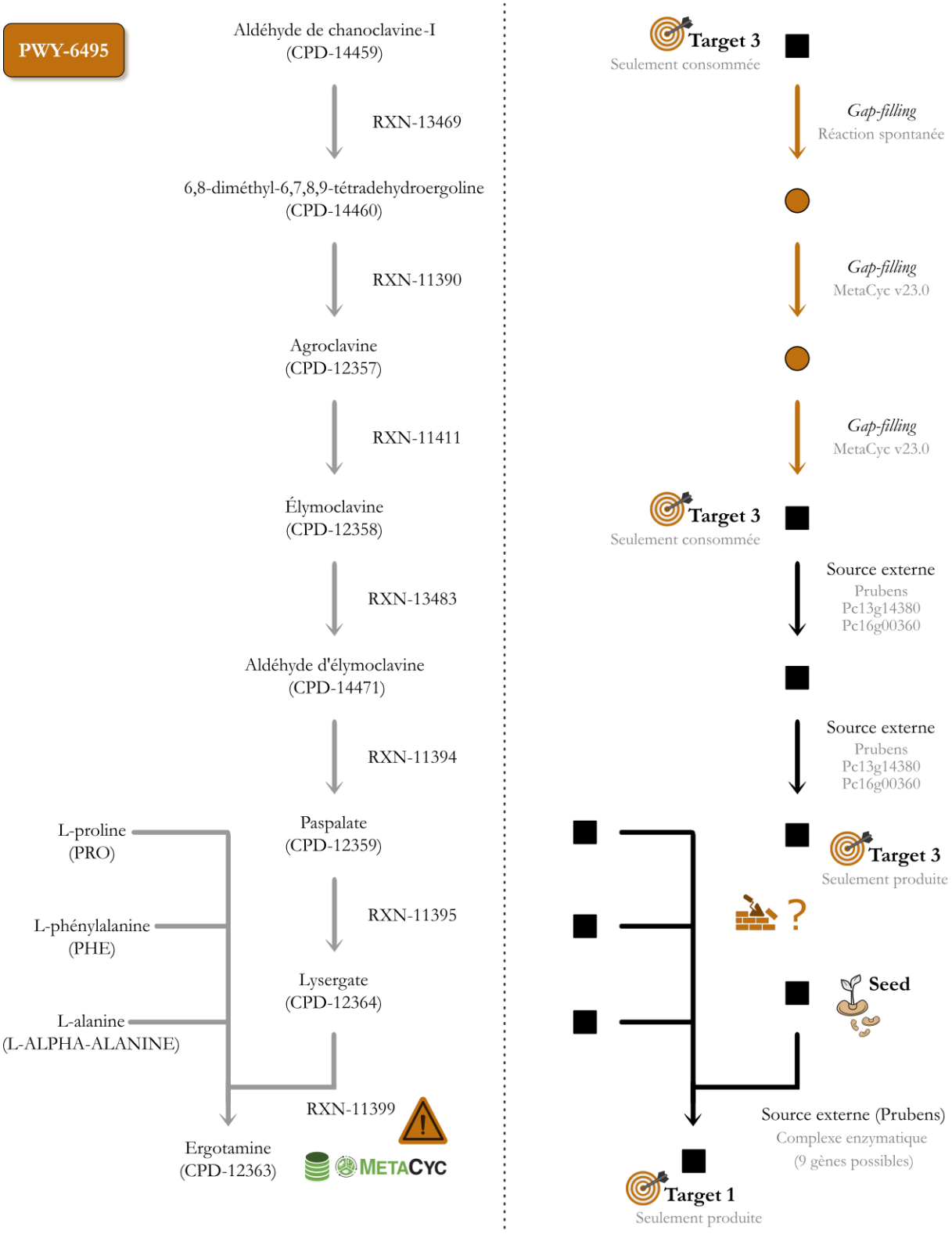


Figure 3-17 : Illustration de l'impact des paramètres d'entrée de Meneco sur le gap-filling : étude de la voie de biosynthèse de l'ergotamine. *À gauche, le schéma de la voie de biosynthèse de l'ergotamine (PWY-6495) telle que référencée dans la version 23.0 de MetaCyc. Cette version est néanmoins obsolète suite à une mise à jour datant de 2023, notamment concernant la réaction RXN-11399 – indiquée par le panneau d'avertissement. À droite, la reconstruction et l'intégration du pathway dans* ***iPrub22****. L'origine de la présence de chacune des réactions est mentionnée. Les noms des métabolites ont été remplacés par des* ***carrés noirs*** *pour signifier que ces composés sont présents dans le draft tout comme les* ***flèches*** *de la même couleur. Les* ***flèches et cercles orange*** *représentent les entités ajoutées par gap-filling. L'ergotamine appartient à l'ensemble de Targets 1, avec un poids de confiance théoriquement plus élevé mais l'utilisation erronée du lysergate en tant que seed explique l'incomplétude de cette voie. En revanche, trois réactions ont été ajoutées automatiquement dont une réaction spontanée afin de débloquer la production de l'élymoclavine. Ce métabolite appartient à la liste Targets 3 regroupant des composés du draft annotés « spécialisés » et qui sont soit seulement consommés, soit seulement produits.*

1 *Un gap-filling semi-automatique*

L'automatisation par les algorithmes de *gap-filling* facilite l'obtention de résultats, mais combler les lacunes d'un réseau métabolique reste un processus où l'intervention humaine est indispensable, d'où la dénomination de semi-automatique. En effet, la plausibilité biologique d'une réaction, ainsi que sa cohérence et son adéquation avec les connaissances existantes sur le métabolisme de l'organisme, ne peuvent encore être systématiquement et uniformément évaluées par un ordinateur. Toutefois, n'ayant ni le temps matériel, ni l'expertise nécessaire pour évaluer la totalité des résultats et décortiquer chaque ajout, nous avons décidé de déporter la problématique de la curation des résultats sur la qualité et le choix des paramètres d'entrée fournis à Meneco. Certes, cela ne permet pas de s'affranchir de l'examen minutieux des résultats, mais nous espérons ainsi obtenir des résultats plus pertinents de façon semi-automatique.

Ainsi, nous avons fait le choix de réaliser une complétion automatique par *gap-filling* pour des ensembles de métabolites de haute confiance mais également sur l'ensemble des *dead-ends* du draft qui étaient annotées « spécialisées » dans MetaCyc. Sur le plan biologique, cette approche, relativement brutale à première vue, peut soulever des questions sur la validité biologique des réactions ainsi complétées, notamment en ce qui concerne leur authenticité et leur pertinence fonctionnelle. Néanmoins, la présence de ces composés dans le draft garantit l'existence d'au moins une réaction menant à leur consommation ou production, étayée par une preuve génomique. Combiné avec l'utilisation de données d'entrée au poids de confiance modulables, nous nous appuyons donc sur ce fait pour justifier les ajouts de réactions de *gap-filling* à ***i*Prub22**.

En nous basant sur la classification de nos listes de cibles, le métabolite sur lequel nous concentrons notre raisonnement est un composé issu de la liste Targets 1. À ce titre, il correspond à un composé dont le poids de confiance est maximal. En conséquence, de prime abord, la question ici n'est pas de s'interroger sur la présence effective de la production de l'ergotamine chez *P. rubens* Wisconsin 54-1255, mais plutôt de comprendre les raisons de la présence ou de l'absence des réactions qui composent sa voie de biosynthèse au sein du réseau, et ce, avant et après l'exécution du *gap-filling*.

2 *Analyse de la voie de biosynthèse de l'ergotamine dans le draft puis de sa complétion dans* ***i*Prub22**

La **Figure 3-17** représente le pathway de biosynthèse de l'ergotamine (`PWY-6495`) tel qu'il était exposé dans la version 23.0 de MetaCyc (*i.e.* version utilisée pour le *gap-filling*). Trois des sept réactions composant cette voie métabolique sont présentes dans le draft et proviennent exclusivement du **GSMN** de 2018, Prubens *(Prigent et al. 2018)*. Les deux premières réactions (`RXN-13483` et `RXN-11394`) décrivent la conversion de l'élymoclavine en paspalate, avec la formation d'un produit intermédiaire, l'aldéhyde d'élymoclavine. Ces deux réactions sont catalysées par la même mono-oxygénase dans MetaCyc, ce qui explique la similarité des associations GPR. La troisième réaction mentionnée dans le draft (`RXN-11399`) correspond, quant à elle, à la production de l'ergotamine.

À ce stade, nous constatons que la voie de biosynthèse de l'ergotamine est morcelée puisque seule la partie finale du pathway semble être conservée. Il est également pertinent de noter que l'aldéhyde de chanoclavine-I, premier métabolite impliqué dans le pathway `PWY-6495`, est inclus dans le draft bien qu'aucune réaction de production ne lui soit associée. Cette présence est justifiée par le fait que ce composé est également impliqué dans une réaction issue de Prubens liée à la biosynthèse de la fumigaclavine, un composé de la même famille que l'ergotamine.

Parmi tous les composés impliqués dans la voie métabolique `PWY-6495`, cinq sont répertoriés dans nos différentes listes de cibles. L'ergotamine est incluse dans la liste des Targets 1, tandis que le paspalate, l'élymoclavine et l'aldéhyde de chanoclavine-I sont classés dans la liste des Targets 3 du fait de leur statut de métabolites spécialisés et de *dead-ends*. À noter que le lysergate, bien qu'il soit également une impasse métabolique, perd son caractère de cible à cause de sa présence dans la liste des *seeds*.

Soulignons désormais qu'en raison de l'utilisation erronée de lysergate en tant que *seed*, la complétion de `PWY-6495` provient de l'utilisation de l'élyclomavine en tant que cible. En effet, comme le lysergate, un métabolite uniquement consommé dans le draft, a été utilisé à tort comme un composé d'initiation, aucune réaction de production n'a pu être ajoutée par *gap-filling* à ce dernier. Vraisemblablement, si le lysergate avait été exclu de la liste des *seeds*, le *gap-filling* réalisé par Meneco aurait proposé un comblement de cette lacune par la réaction `RXN-11395`, un chaînon manquant de la voie de biosynthèse de l'ergotamine dans ***i*Prub22**.

Outre cette réaction, le draft est dépourvu des trois premières réactions de `PWY-6495` qui permettent la transformation de l'aldéhyde de chanoclavine-I en élymoclavine. Ces réactions ont été ajoutées par *gap-filling*, et nous les décrivons ci-après pour déterminer leur pertinence. La première réaction de la voie de biosynthèse de l'ergotamine est la déshydratation de la chanoclavine-I aldéhyde, une réaction spontanée (`RXN-13469`). Ensuite, le produit résultant est transformé en agroclavine par une déhydroxygénase (`RXN-11390`). Cette réaction a été ajoutée lors du *gap-filling* qui utilisait les cibles de l'ensemble Targets 3 contre la totalité des réactions de MetaCyc v23.0. Elle dispose donc du poids de confiance minimal dans les réactions ajoutées. Enfin, l'intervention d'une oxydoréductase permet la transformation de l'agroclavine en élymoclavine (`RXN-11414`). Tout comme la réaction `RXN-11395`, cette réaction appartient à l'ensemble des réactions orphelines.

En résumé, en supposant que la réaction d'isomérisation du paspalate en lysergate ait été ajoutée, la complétion de `PWY-6495` implique l'ajout d'une réaction spontanée, de deux réactions orphelines et d'une seule réaction biochimique caractérisée. Il s'agit principalement de réactions actuellement non étayées par des données génomiques et qui ne pouvaient donc pas être initialement contenues dans le draft. Nous pouvons toutefois nous interroger sur l'absence systématique de sources complémentaires pour les réactions apportées par Prubens, notamment d'annotation fonctionnelle, venant soutenir l'hypothèse de la présence de cette voie chez notre champignon.

3 *Actualisation et évolution du pathway PWY-6495*

Un facteur non négligeable à prendre également en compte est la temporalité des informations disponibles. En effet, contrairement au métabolisme de base, où nous pouvons supposer une plus grande stabilité des données, les connaissances liées au métabolisme spécialisé sont généralement plus sujettes à l'évolution.

La voie de biosynthèse de l'ergotamine est répertoriée dans MetaCyc depuis 2010, mais une mise à jour significative a été réalisée entre mars et mai 2023. Cette révision a entraîné l'ajout d'intermédiaires lors de la dernière étape de biosynthèse de la voie, ainsi qu'une modification conséquente de la réaction `RXN-11399`. Cette dernière a été subdivisée en cinq sous-réactions qui, vraisemblablement, sont toutes catalysées par la même NRPS. Attention toutefois car l'équation de cette réaction a été modifiée, puisque, désormais elle ne produit plus l'ergotamine mais l'"*ergotamam*". Parmi ces cinq réactions, quatre impliquent l'estérification du lysergate et des acides aminés phénylalanine, proline et alanine avec l'adénosine monophosphate (AMP), formant respectivement l'adénylate d'acide lysergique, le L-phenylalanyl-adénylate, le L-prolyl-adénylate et le L-alanyl-adénylate. Les réactions d'estérification de la phénylalanine (`RXN-23892`) et du lysergate (`RXN-24056`) ont été créées et intégrées dans MetaCyc en 2023, de même que la cinquième réaction (`RXN-24054`) qui correspond à la fusion des quatre adénylates qui forment l'"*ergotamam*". En revanche, les réactions de couplage de l'alanine (`RXN-17190`) et de la proline (`RXN-19079`) étaient déjà encodées et associées à des séquences génomiques dans la version 23.0. de MetaCyc, et auraient pu être intégrées au draft, si une correspondance génomique avait été détectée.

La réaction produisant de l'ergotamine à partir de l'"*ergotamam*" (*i.e.* nouveau composé ajouté en mai 2023 sans référencement vers d'autres bases de données) est catalysée par une dioxygénase isolée chez *Claviceps purpurea*. Cependant, une recherche rapide d'homologie effectuée *via* un blastp sur le serveur du NCBI, en ciblant exclusivement le protéome de *P. rubens* Wisconsin 54-1255 avec le reste des paramètres laissés par défaut, n'a donné aucun résultat.

Au regard de ces nouveaux éléments, et en admettant que *P. rubens* soit effectivement capable de produire de l'ergotamine, il semblerait que cette production résulte d'un autre mécanisme. Les voies présentes dans ***i*Prub22** seraient donc de probables faux positifs.

4 *L'ergotamine, une cible pertinente ?*

À la vue de ces diverses informations, il est désormais nécessaire de nous interroger sur la pertinence de la présence de l'ergotamine dans la liste des cibles à prioriser (Targets 1). En examinant son origine, nous constatons qu'elle provient de la liste des métabolites secondaires du réseau de 2018, qui, nous le rappelons, est un réseau issu d'un processus automatique de reconstruction effectué sur 24 espèces de *Penicillium*. Ainsi, afin d'affiner nos cibles, ajouter une couche de granularité dans la confiance en faisant la distinction entre les composés détectés *in vivo* ou *in vitro* et ceux détectés uniquement *in silico,* pourrait sembler approprié. Parmi les 243 métabolites présents dans l'ensemble Targets 1, 12 autres métabolites proviennent exclusivement de la reconstruction de 2018 et mériteraient à ce titre d'être vérifiés. En outre, même si selon

les ontologies MetaCyc, ils sont tous annotés comme étant des métabolites secondaires, à la lumière de la liste suivante, cette classification est largement discutable (cf. composés écrits en gras) : le bikisocoumarine, l'**époxy-squalène**, l'ergotamine, le **diphosphate de géranyle-géranyle**, le phomopsénoate de méthyle, le **myo-inositol**, le nivalénol, l'ophioboline F, l'acide stellatique, le stipitatate, la versicolorine A et le **(R)-mévalonate**. Néanmoins, une attention particulière sur les voies de biosynthèse des « authentiques » métabolites spécialisés devraient être envisagée prioritairement pour les curations futures d'***i*Prub22**.

5 *Biosynthèse des alcaloïdes dérivés de l'ergoline dans la littérature*

En l'état, le simple constat des informations présentes ou absentes du draft ne fait que générer des hypothèses. Pour approfondir notre compréhension, deux possibilités s'offrent à nous : **(1)** la recherche bibliographique et **(2)** l'expérimentation. Cette dernière peut-être réalisée soit *in silico*, par du *genome mining*, soit par des approches dites plus classiques, telles que des expériences de *knock-out* ou des études biochimiques d'enzymes d'intérêt. Nous nous concentrons ici sur une exploration plus précise des données disponibles dans la littérature.

L'ergotamine est une mycotoxine appartenant aux alcaloïdes indoliques, une famille de molécules connues pour leurs effets psychotropes, analgésiques, et parfois toxiques sur les organismes. L'ergotamine a été découverte historiquement chez *Claviceps purpurea*, un champignon vénéneux et parasite des graminées de l'ordre des Ascomycètes, connu communément sous le nom de l'ergot de seigle. Les diverses molécules apparentées qui ont été décrites par la suite ont héritées de la racine de son nom et sont classées comme alcaloïdes dérivés de l'ergoline *(Gerhards et al. 2014)*.

À ce jour, de nombreuses espèces de champignons Ascomycètes, réparties notamment dans les ordre des Hypocreales (*e.g. Claviceps*) et des Eurotiales (*e.g. Aspergillus*, *Penicillium*), sont connues pour produire ces types d'alcaloïdes *(Robinson et Panaccione 2015)*. Ces molécules, ayant eu historiquement un fort intérêt pharmacologique, ont longtemps été utilisées, parfois à mauvais escient, comme médicaments pharmaceutiques ou comme précurseurs de ceux-ci *(Haarmann et al. 2009)*. Elles sont toutes produites à partir du tryptophane et partagent une structure commune, le noyau ergoline. Ces alcaloïdes sont divisés en trois catégories : les clavines, les amides de l'acide lysergique et les ergopeptines *(Gerhards et al. 2014)*.

La chanoclavine est un élément clé dans la voie de biosynthèse de ces alcaloïdes. La biosynthèse de cet intermédiaire semble jusqu'alors être partagée par l'ensemble des espèces connues et sa forme aldéhyde marque un point de bifurcation à partir duquel les trois principales familles d'alcaloïdes se développent *(Coyle et Panaccione 2005)*. La revue sur les voies de biosynthèse des alcaloïdes de l'ergot *(Gerhards et al. 2014)* les classe ainsi : l'agroclavine chez *C. purpurea* qui donnera le lysergate, la pyroclavine chez *P. commune* et la festuclavine chez *A. fumigatus* qui sont toutes deux des intermédiaires à la production des fumigaclavines.

À notre connaissance, il n'existe pas de preuves formelles et probantes de la production des alcaloïdes mentionnés présentement chez la souche *P. rubens* Wisconsin 54-1255. En revanche, il semblerait que des espèces du genre *Penicillium* puissent avoir, ou eurent, la capacité de produire ces molécules. À titre

d'exemple, dans le cadre d'une fermentation par culture de surface, il a été observé chez *P. citrinum* une réponse de production d'alcaloïdes dérivés de l'ergoline en fonction de l'utilisation de divers substrats, sans toutefois préciser le type d'alcaloïdes produits *(Shahid et al. 2015)*. Au sein de ces mêmes travaux une production de ces composés a également été détectée chez *P. oxalicum*, *P. italicum* et *P. digitatum*.

La revue de Martín et *al. (2017)* sur les clavines, stipule quant à elle, que leurs producteurs ne possèdent pas, ou plus, l'enzyme mono-oxygénase responsable de la formation du lysergate. Spécifiquement, les espèces de *Penicillium* concernées par cette affirmation sont *P. roqueforti*, *P. commune*, *P. camemberti*, *P. expansum*, *P. steckii* et *P. griseofulvum*. Or, dans notre draft, les réactions `RXN-11383` et `RXN-11394` sont soutenues par les gènes *Pc13g14380* et *Pc16g00360*. Ces deux séquences sont annotées sous UniProt comme ayant une activité de mono-oxygénase et sont impliquées dans la biosynthèse de mycotoxines, avec, toutefois un score qualité d'annotation respectif de 2 et 1 sur 5. Néanmoins, le genre des *Penicillium* est vaste et regroupe un grand nombre d'espèces. Ainsi, pour trancher sur le maintien ou non de ces réactions dans ***i*Prub22**, des analyses supplémentaires seraient donc nécessaires. En outre, un BGC fonctionnel de la clavine a été identifié chez *P. commune*, *P. biforme* et *P. roqueforti* et des groupements de gènes putatifs apparentés ont également été détectés chez les espèces *P. expansum*, *P. steckii*, *P. griseofulvum* *(Martín et al. 2017)* et *P. camemberti* *(Fabian et al. 2018)*. Cependant, les gènes essentiels n'ont pas encore été identifiés et ces espèces ont vraisemblablement perdu la capacité de produire ces alcaloïdes.

6 *Biosynthèse des alcaloïdes dérivés de l'ergoline dans le réseau*

Les alcaloïdes dérivés de l'ergoline trouvent leur origine commune dans le tryptophane prénylé *(Gerhards et al. 2014)*. Dans le draft, deux métabolites correspondant à cette nomenclature sont identifiés : le 4-(3-méthylbut-2-ényle)-L-tryptophane (`CPD-460`) et le 7-(3-méthylbut-2-ényle)-L-tryptophane (`CPD-12156`). Alors que le premier est impliqué dans la voie de biosynthèse de l'aldéhyde de la chanoclavine-I, le second n'est associé sur MetaCyc qu'à sa seule réaction de production (`RXN-11191`). Dans le draft, ces composés sont tous deux des impasses métaboliques uniquement produites qui doivent leur présence à des réactions dont les séquences génomiques ont été détectées par orthologie chez les *templates* *A. nidulans* et le complexe d'espèces de *P. chrysogenum*.

Si nous nous focalisons sur la topologie des pathways de l'aldéhyde de chanoclavine-I (`PWY-6493`) et des fumigaclavines (`PWY-7059`), voici ce que nous observons. Hormis la prénylation du tryptophane (`TRYPTOPHAN-DIMETHYLALLYLTRANSFERASE-RXN`) qui est la première des quatre réactions de `PWY-6493`, les trois autres réactions sont absentes du draft. Sachant qu'elles n'appartiennent, ni aux réactions spontanées, ni aux réactions orphelines et qu'elles sont catalysées par trois enzymes différentes, leur absence est source de questionnement. En effet, cette voie est commune aux familles d'alcaloïdes dérivés de l'ergoline cités précédemment. Pour déterminer si cette voie est effectivement absente du métabolisme de notre organisme, si elle souffre d'une caractérisation insuffisante, si un pathway alternatif pour la production de l'aldéhyde de la chanoclavine-I existe, ou, si les éléments présents dans le draft sont des éléments ancestraux d'anciennes fonctionnalités désormais perdues, des recherches ciblées sont nécessaires.

Le pathway de biosynthèse des fumigaclavines comprend six réactions successives, dont seulement deux sont présentes dans le draft. Si nous nous en tenons à ce chiffre, seulement deux tiers de la voie métabolique est reconstruite automatiquement. Ces réactions correspondent à la cyclisation intra moléculaire de l'aldéhyde de chanoclavine-I (`RXN-13476`) par formation d'une imine, première réaction du pathway qui provient une fois de plus de Prubens et à la réaction d'acétylation de la fumigaclavine B en fumigaclavine A (`RXN-13366`), détectée par l'annotation fonctionnelle du génome et correspondant à la cinquième réaction du pathway. La deuxième réaction de `PWY-7059` est une réaction spontanée de cyclisation (`RXN-13455`) qui est suivie par la formation de la festuclavine (`RXN-13478`). Cette réaction est associée à une enzyme dans MetaCyc contrairement à la quatrième réaction (`RXN-13479`), qui est donc une réaction orpheline permettant la production de fumigaclavine B. Enfin, la dernière réaction de cette voie de biosynthèse fait intervenir une diméthylallyltransférase qui permet la formation de fumigaclavine C à partir de la fumigaclavine A (`RXN-11367`). La complétion par *gap-filling* a permis de combler les lacunes entre les deux réactions initialement présentes dans le draft en ajoutant les réactions nécessaires. Sur ces trois réactions, remarquons une fois de plus, qu'en raison des connaissances actuelles, deux d'entre-elles ne pouvaient être contenues dans le draft. Enfin, un argument, certes non suffisant, mais qui pourrait être en faveur du maintien de ce comblement de lacunes dans ***i*Prub22**, est la caractérisation du cluster de gènes responsable de la production de fumigaclavine A chez l'espèce *P. commune*. De plus, contrairement à *A. fumigatus* qui a la capacité de produire de la fumigaclavine C (*i.e.* produit final de la voie de biosynthèse présentée), cette espèce de *Penicillium* ne posséde pas l'enzyme diméthylallyltransférase responsable de cette transformation *(Robinson et Panaccione 2015)*.

7 *Conclusion*

La description détaillée de cet exemple visait à présenter les étapes et les questions pouvant être soulevées lors du *gap-filling*. Cet exemple met en lumière les limites de l'automatisation et souligne l'importance des réflexions et de l'attention requises pour chaque ajout effectué par le *gap-filling*. L'analyse de l'origine de chaque donnée permet néanmoins de comprendre les raisons derrière ces ajouts et ouvre des perspectives pour les futures améliorations d'***i*Prub22**.

2.2.1.3. Résultats du *gap-filling* et impact sur la productibilité des métabolites d'***i*Prub22**

En résumé, l'expression « *gap-filling* raisonné » se justifie et s'explique par le croisement des diverses listes de cibles et des réseaux de réparation que nous avons utilisés. Le comblement des lacunes s'effectue ainsi méthodiquement, en utilisant la granularité du taux de confiance accordé à la fois aux sous-ensembles de réparation et aux listes de cibles utilisées. De plus, la traçabilité de l'origine des réactions fournit les annotations nécessaires pour pondérer la pertinence des réactions ajoutées et, le cas échéant, effectuer les corrections et ajustements nécessaires. Une synthèse linéarisée des combinaisons de cibles et de réseaux de réparation a été présentée en **Figure 3-16** (page 241). La **Figure 3-18** quant à elle, expose l'origine des 510 réactions introduites par *gap-filling*, représentant ainsi 8,6 % de l'ensemble des réactions d'***i*Prub22**. Certes, bien que ce nombre soit relativement conséquent, il est néanmoins essentiel de souligner que 44,7 % d'entre elles ne peuvent être inférées automatiquement lors de la reconstruction en raison de l'absence de sources génomiques associées (*i.e.* réactions spontanées et orphelines).

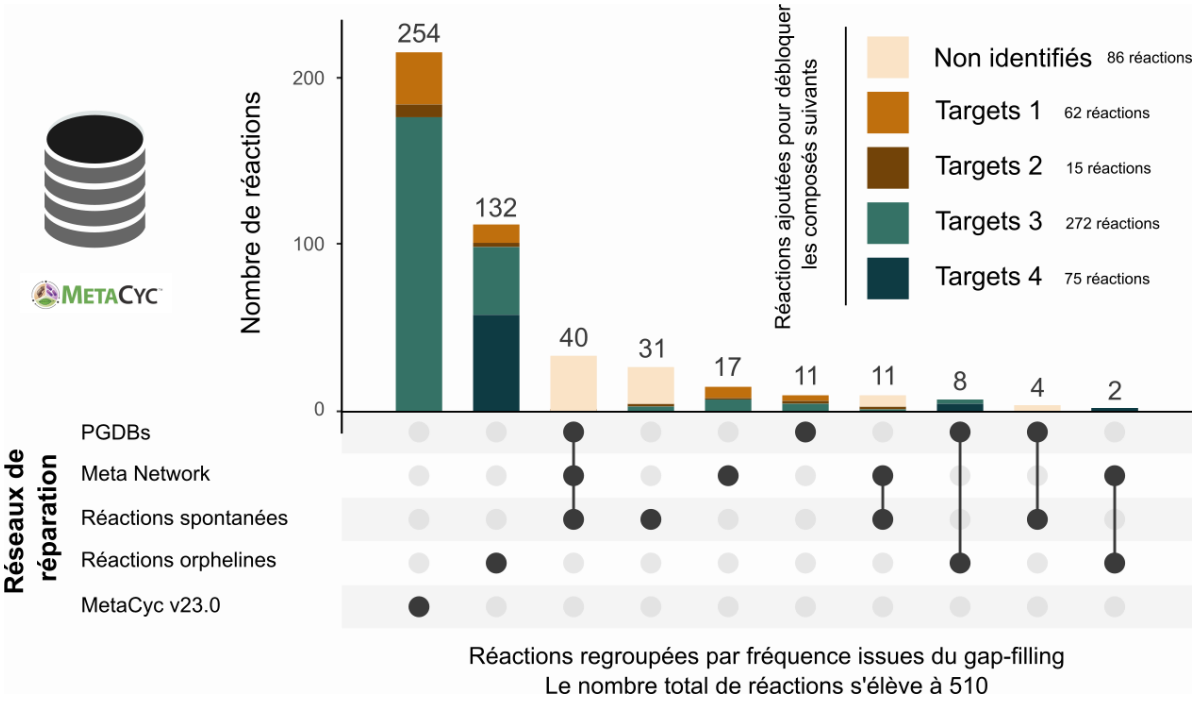


Figure 3-18 **: *Présentation des « réseaux de réparation » pour un « gap-filling raisonné ».*** *Distribution des 510 réactions ajoutées à la reconstruction au cours des étapes de gap-filling. À l'exception de la plupart des réactions répertoriées « spontanées » (i.e. de par leur nature, ces réactions ont été regroupées dans un seul document, et ce faisant, leur origine a été égarée), les réactions illustrées en Figure B sont regroupées en fonction des listes des cibles ayant permis leur détection. Parmi les réactions ajoutées, 256 d'entre elles proviennent de sous-ensembles préférentiels extraits de MetaCyc. Ce gap-filling « raisonné » renforce la confiance dans la présence de la moitié de ces réactions dans la reconstruction.*

Avant de poursuivre sur l'impact du *gap-filling* sur la topologie d'***i*Prub22**, nous souhaitons faire un aparté sur les associations GPR. Le *gap-filling* se décompose en trois étapes majeures : **(1)** l'identification des *dead-ends*, **(2)** le comblement des lacunes résultantes par l'ajout de réactions, et lorsque cela est possible, **(3)** le renforcement de ces ajouts avec des preuves par l'intégration d'associations GPR. Ces associations établissent des liens fonctionnels entre les gènes, les enzymes et les réactions métaboliques qu'elles catalysent. Ces triades génomiques sont définies dans un contexte systémique visant à appréhender le système biologique dans son ensemble, en examinant les interactions entre ses diverses parties afin d'obtenir une compréhension générale de son fonctionnement.

L'exactitude d'une association GPR ne revêt pas une importance fondamentale pour la conception de la reconstruction. En effet, la présence de faux positifs au sein d'une association GPR peut, par exemple, être représentative d'une faible spécificité au niveau de la réaction. En revanche, l'exactitude d'une association GPR devient cruciale pour son application utilisant le principe de manipulation génétique (*e.g. in silico knock-out*). Cet aspect, non limité aux réactions issues du *gap-filling*, demande néanmoins un intense travail de curation et une approche interdisciplinaire impliquant la vérification expérimentale et une expertise approfondie. Concernant les 350 réactions biochimiques issues du *gap-filling* et catalysées par une enzyme, la complétion génomique, étayée par des références issues de la littérature scientifique, s'est restreinte à 20 réactions (cf. section *3.2 Reconstruire les voies de biosynthèse des métabolites spécialisés*, page 318).

Le dernier aspect à aborder concerne l'incidence de l'intégration des 510 réactions issues du *gap-filling* sur la topologie de la reconstruction. Notons que nous préférons ici raisonner en termes de pourcentage, étant donné que le nombre total de métabolites peut être amené à varier entre chaque itération.

Afin d'effectuer les comparaisons entre le draft et ***i*Prub22**, les métabolites sont classés en trois catégories : **(1)** ceux uniquement consommés, **(2)** ceux uniquement produits et **(3)** ceux à la fois consommés et produits. Rappelons que cette annotation ne concerne que les voisins directs du métabolite étudié, et qu'un métabolite consommé et produit, n'est pas, *de facto*, automatiquement productible au sens topologique. Ensuite, en simulant un environnement nutritionnel ultra riche à l'aide de la liste de « *seeds* » précédemment décrite, nous pouvons quantifier les potentialités de production de la reconstruction.

Les résultats présentés en **Tableau 3-1** (page 141) font, certes, état d'une amélioration globale de la connectivité du réseau, avec une augmentation de 15 % de métabolites productibles entre le draft et ***i*Prub22**, mais ils témoignent également que des pistes d'amélioration subsistent. Toutefois, la productibilité des composés cibles a été améliorée, quant à elle, de façon majeure, entre 15 et 70 points de pourcentage.

Nous observons, de plus, une diminution de 8,3 points de pourcentage du nombre de *dead-end*, avec respectivement 5,0 % et 3,3 % de métabolites uniquement consommés et uniquement produits (**Figure 3-19.A**). En termes de chiffres, cela se traduit par la connexion de 331 métabolites initiaux et de l'ajout de 208 nouveaux composés à la fois consommés et produits. Cependant, en contrepartie, nous constatons l'ajout de 25 nouveaux métabolites *dead-ends* (**Figure 3-19.B**). Il est à remarquer que sur les 20 nouveaux métabolites uniquement produits, il s'agit en réalité d'un prolongement des voies métaboliques préexistantes.

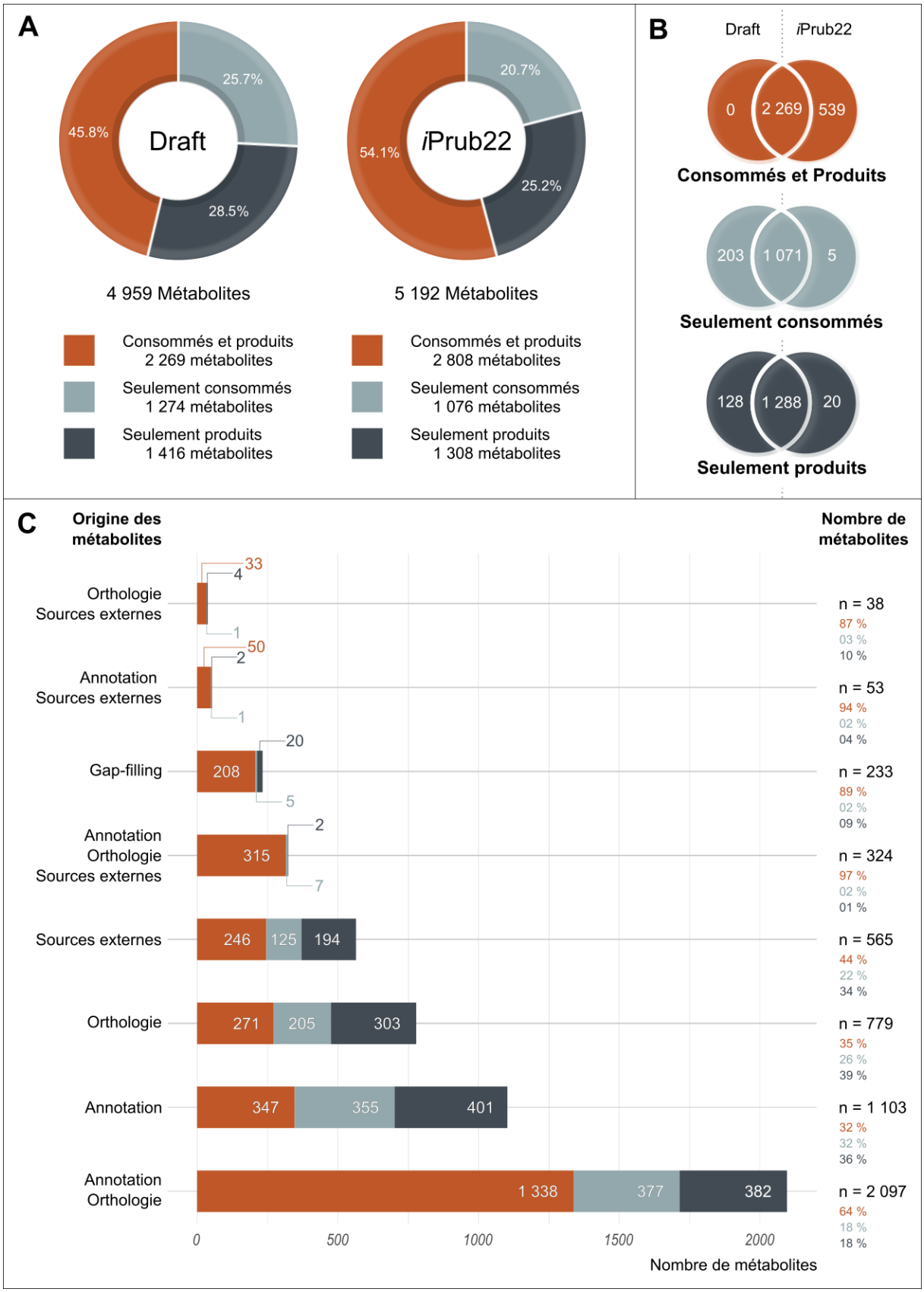


Figure 3-19 : Impact du gap-filling sur l'amélioration de la connectivité d'iPrub22. ***(A)*** *Évolution des ensembles de métabolites uniquement consommés, uniquement produits, ou consommés et produits avant et après gap-filling. Une augmentation de 8,3 points de pourcentage de connectivité entre le modèle initial et* ***iPrub22*** *est constatée.* ***(B)*** *Comparaison des ensembles de métabolites avant et après gap-filling.* ***(C)*** *Origine des métabolites présents dans* ***iPrub22*** *selon les sources de reconstruction des réactions. Le gap-filling contribue à l'addition de 233 métabolites, dont 89 % sont à la fois consommés et produits.*

2.2.2. MODÉLISATION DE DIFFÉRENTES CONDITIONS DE CULTURE

2.2.2.1. Définir les éléments constitutifs des milieux de culture

Expérimentalement, les champignons filamenteux sont cultivés en milieu liquide ou solide, sur des milieux contrôlés ou non. La culture sur gélose est privilégiée pour l'isolement des souches pures, la conservation à moyen terme et la production de sporulation. Cette méthode offre une surface plane qui favorise le développement distinct des colonies fongiques et facilite l'observation des caractéristiques morphologiques. En revanche, la culture en milieu liquide est préférée lors d'études en laboratoire ou à des fins industrielles, notamment pour la production de biomasse à grande échelle, la fermentation et la croissance en suspension. Elle permet un contrôle plus précis des conditions de croissance telles que la température, le pH et la dispersion homogène des nutriments.

Parmi les *media* non contrôlés couramment usités, nous pouvons citer le CYA (« *Czapek Yeast Autolysate* »), le PDA (« *Potato Dextrose Agar* ») ou bien le MEA (« *Malt Extract Agar* »). Le CYA est un milieu composé d'autolysat de levure qui favorise la croissance fongique en raison de sa teneur en éléments nutritifs variés. Le PDA, constitué de pomme de terre et de dextrose, est largement utilisé pour l'isolement et la culture des champignons, tandis que le MEA, à base d'extrait de malt, est réputé pour favoriser la formation de spores et la croissance mycélienne. Ces milieux de culture sont souvent enrichis pour fournir aux champignons les nutriments nécessaires à leur croissance et à leur développement. Cependant, il convient de noter que la composition exacte de ces milieux naturels n'est pas toujours connue de manière précise et peut varier d'un fournisseur à l'autre, voire en lots d'un même fournisseur. Ainsi, la variabilité des ingrédients et des proportions utilisées dans la préparation des *media* rendent leur transposition dans un contexte informatique complexe.

De plus, un contrôle du milieu de culture permet, dans certains cas, d'influencer favorablement la modulation de la morphologie des champignons filamenteux (*e.g.* une diminution de la viscosité favorise l'agrégation des mycéliums) et de moduler la production de composés d'intérêt (*i.e.* approche **OSMAC**). Les milieux chimiquement définis se décomposent essentiellement en éléments structuraux et catalytiques *(Raven et al. 2011)*.

Parmi les éléments structuraux, outre l'oxygène et l'hydrogène indispensables à leur métabolisme, les champignons ont besoin de sources de carbone et d'azote pour leur croissance et leur développement. Ces sources peuvent être indépendantes (*e.g.* glucose et ammonium) ou non (*e.g.* en raison de la présence d'un groupement azoté sur une molécule organique, de nombreux acides aminés peuvent servir à la fois de sources de carbone et d'azote). Le phosphore et le soufre sont, quant à eux, généralement introduits sous forme d'ions pour répondre aux besoins nutritionnels des champignons.

Les éléments catalytiques englobent les constituants des enzymes et les cofacteurs, comprenant des minéraux tels que le magnésium, le fer, le cuivre, le calcium, le manganèse ou le zinc. D'autres cations comme les ions sodium ou potassium sont souvent incorporés pour réguler le pH intracellulaire et maintenir l'osmolarité. L'ammoniaque, souvent utilisée comme source d'azote sous forme de sels d'ammonium (*e.g.* ammonium sulfate, ammonium phosphate, ammonium nitrate) permet également d'ajuster le pH du milieu. Optionnellement, les milieux de culture peuvent être enrichis en agents chélatants tel que l'acide citrique, en agents tampons comme le carbonate de calcium, ainsi qu'en vitamines *(Raven et al. 2011)*.

La combinaison de ces éléments est nécessaire à la croissance des organismes et, par conséquent, à la formation de biomasse. Une liste non exhaustive de ces composés est présentée en **Figure 3-20**. La question se pose alors de savoir comment intégrer ces informations dans un réseau métabolique pour une modélisation approfondie de la physiologie fongique.

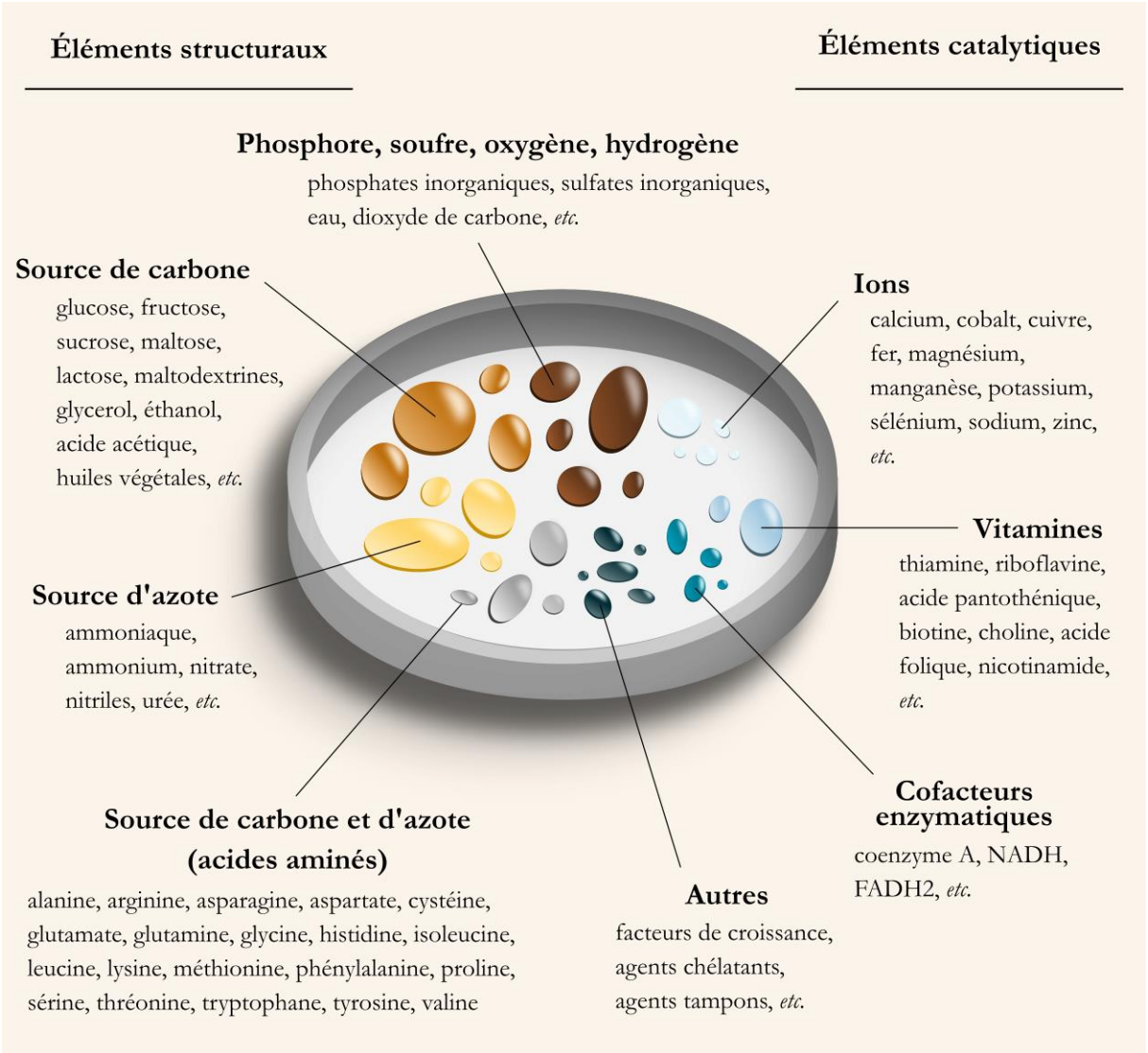


Figure 3-20 : Schématisation des éléments nécessaires à la conception d'un milieu de culture contrôlé. *L'élaboration d'un milieu de culture contrôlé crée un environnement propice à la croissance d'un organisme. Cette figure expose une liste non exhaustive de composés essentiels, classés en deux catégories principales : les éléments structuraux qui sont les composants de base des biomolécules constitutives des cellules et les éléments catalytiques qui contribuent à la régulation des processus cellulaires. Ensemble, ces éléments fournissent aux organismes les conditions nécessaires à leur survie et à leur développement en laboratoire, facilitant ainsi l'étude de leur physiologie, de leur biochimie et de leur écologie.*

2.2.2.2. Transposer ces composants dans la modélisation métabolique

L'objectif initial de la reconstruction d'***i*Prub22** était de créer une plateforme de ressources permettant de simuler la production de divers métabolites dans une gamme étendue de conditions expérimentales (*i.e.* simulation ***in silico*** de l'approche **OSMAC**).

À cette fin, nous nous sommes efforcés d'être le plus exhaustifs possible afin de fournir à l'utilisateur de notre reconstruction une large palette de ressources nutritives. Ainsi, nous nous sommes d'abord appuyés sur des descriptions de milieux expérimentaux existants. Nous avons ensuite complété ces informations, d'une part, avec les données des **GSMNs** antérieurs, et d'autre part, avec la liste des métabolites extracellulaires présents dans le draft. Cette approche vise à établir une connectivité topologique de tous les éléments nécessaires à la conception de divers milieux. Par la suite, leurs compositions pourront être affinées, par exemple, par des analyses PhPPs (*phenotype phase plan*) ou des échantillonnages aléatoires, permettant ainsi d'évaluer la sensibilité des modèles aux variations de composition de l'environnement et de quantifier les proportions d'éléments nécessaires.

Toutefois, une question subsiste : comment adapter l'ensemble de ces informations à la modélisation d'un réseau métabolique ? Concrètement, la réponse semble évidente : il convient de sélectionner et d'intégrer les métabolites largement documentés constituant les milieux nutritionnels potentiels de nos modèles, tout en garatissant leur cohérence topologique au sein de la reconstruction. Cette cohérence implique l'existence de réactions permettant leur acheminement de l'extérieur du système au milieu extracellulaire, puis au milieu intracellulaire. Néanmoins, du point de vue de la modélisation, la mise en œuvre de ces éléments est bien plus complexe.

Au sein de l'ontologie des composés de MetaCyc, les diverses entités sont classées, par définition, selon une spécificité croissante, où les feuilles correspondent généralement à des métabolites du type « *compound* », tandis que les termes parents sont référencés sous l'étiquette « *compound class* ». L'utilisation conjointe de termes ultra génériques avec des identifiants plus spécifiques entraîne une redondance et une complexification des modèles. Néanmoins, selon la profondeur des connaissances d'une part, et la qualité de l'annotation employée d'autre part, des réactions propres à chacun de ces termes coexistent. Ainsi, le nombre relativement élevé de réactions composant ***i*Prub22** par rapport aux autres reconstructions actuelles trouve une première explication dans l'organisation des métabolites dans MetaCyc.

En tenant compte de ces considérations, une fois que nous avons établi la liste des composés que nous souhaitons utiliser pour simuler différents milieux nutritionnels, la première étape consiste à déterminer quels sont les identifiants à leur attribuer. Pour répondre à cette question, nous illustrons nos propos en nous axant sur la source de carbone la plus représentative : le glucose (cf. encart : *Un « même » composé mais des identifiants différents, lequel choisir ? Illustration avec l'entité glucose*, page 259). Notons que les considérations exposées ci-dessous concernent exclusivement les sucres mais un raisonnement similaire est effectué pour tout autre type de molécules.

Un « même » composé mais des identifiants différents, lequel choisir ? Illustration avec l'entité glucose

Chaque monosaccharide aldéhydique existe sous forme d'un mélange d'au moins six composés linéaires ou cycliques : l'aldéhyde libre, l'aldéhyde hydraté (*i.e.* le gem-diol), deux isomères de pyranose et également deux isomères de furanose. Chacune de ces formes peut interagir de manière différente dans diverses réactions chimiques en raison de leur structure spatiale unique. En théorie, en raison des interconversions intramoléculaires naturelles, sur une échelle de temps suffisamment longue, toutes les formes du même sucre tendraient vers un équilibre, bien que cette convergence puisse être influencée par des facteurs tels que le pH, la température et la présence de catalyseurs.

Pour illustrer ce point, nous présentons ci-dessous l'ensemble des termes enfants référencés sous le terme parent **Glucose**. Les identifiants MetaCyc sont représentés entre parenthèses, les termes en gras sont ceux présents dans le draft avec le nombre de réactions dans lesquelles ils sont impliqués. Au regard de ces éléments, le choix du ou des identifiants qui modéliseront l'entrée du glucose dans le système est donc primordial.

Glucose (**Glucose**) — 2 réactions
- D-Glucose (*D-Glucose*)
  - D-glucofuranose *(D-Glucofuranose*)
    - *a*-D-glucofuranose (*CPD-24951*)
    - *β*-D-glucofuranose (*CPD-24952*)
  - D-glucopyranose *(**Glucopyranose***) — 62 réactions
    - *a*-D-glucopyranose (***ALPHA-GLUCOSE***) — 25 réactions
    - *β*-D-glucopyranose (***GLC***) — 8 réactions
  - Aldehydo-D-glucose (*CPD-15374*)
  - D-glucose gem-diol (*CPD*-2953)
- L-glucose (*L-glucose*)
  - L-glucofuranose (*L-glucofuranose*)
    - *a*-L-glucofuranose (*CPD-24996*)
    - *β*-L-glucofuranose (*CPD-24997*)
  - L-glucopyranose (*L-glucopyranose*)
    - *a*-L-glucopyranose (*CPD-3607*)
    - *β*-L-glucopyranose (*CPD-24994*)
  - Aldehydo-L-glucose (*CPD-24995)*

Les formes D (*i.e.* dextrogyre) et L (*i.e.* levogyre) définissent l'orientation spatiale des groupes fonctionnels. Cette configuration stéréochimique permet de distinguer les isomères optiques des sucres. Le préfixe « aldehydo » fait référence aux formes linéaires, tandis que les formes furanoses et pyranoses désignent les formes cycliques à 5 ou 6 carbones. Les anomères α et β caractérisent l'orientation du groupe hydroxyle sur le carbone impliqué dans la formation du cycle (*i.e.* soit opposé, soit sur le même plan).

En milieu naturel, l'occurrence de ces diverses formes est variable. De manière générale, en raison d'une plus grande stabilité en milieu aqueux, le D-glucose est plus abondant que le L-glucose. Les formes linéaires présentent une réactivité plus élevée que les formes cycliques, induisant alors une conversion rapide. Les formes furanoses et pyranoses peuvent coexister en équilibre dynamique. Toutefois, en raison d'une stabilité thermodynamique plus élevée, la forme pyranose est plus abondante dans le milieu nutritif et est majoritairement représentée dans les voies métaboliques. Enfin, les anomères α et β du D-glucopyranose peuvent être présents dans des proportions variables mais l'anomère β est généralement plus stable en solution. Notons également la possibilité de conversions chimiques spontanées entre les stéréoisomères du glucose avec une conversion de l'anomère β en α qui se produit généralement plus rapidement.

Si nous comparons ces informations à la nature des identifiants de glucose présents dans le draft, nous constatons que ces diverses généralités, à l'exception de la prédominance de l'anomère β, sont respectées.

Pour modéliser l'import du glucose dans ***i*Prub22** et être en adéquation avec les données présentes dans le draft, nous avons à notre disposition quatre identifiants distincts. En nous focalisant sur le terme de plus haut niveau, `Glucose`, nous constatons qu'il intervient dans seulement deux réactions appuyées par des séquences génomiques.

Réactions impliquant l'entité glucose (draft)

3.2.1.1-RXN	1 Starch + 1 WATER => 1 MALTOSE + 1 **Glucose**	(11 gènes)	Annotation
RXN-14352	1 **Glucose** [cytosol] => 1 **Glucose** [extracellulaire]	(9 gènes)	Prubens

Les équations de ces réactions nous indiquent, d'une part, que l'entité `Glucose` est déconnectée de l'ensemble de ses termes enfants et, d'autre part, que l'entité présente dans le milieu extracellulaire constitue une impasse métabolique uniquement produite. De plus, en raison de l'absence de connectivité entre les identifiants, la dissociation de l'amidon (*i.e.* `Starch`) en maltose et glucose, bien que présente dans la reconstruction, ne sera pas évaluée lors de la modélisation de flux, à moins d'effectuer une conciliation d'identifiants.

Ensuite, en examinant les équations des 62 réactions qui impliquent l'identifiant `Glucopyranose`, nous constatons à nouveau qu'il n'existe pas de réaction directe connectant cet identifiant à ses termes enfants. En revanche, il existe une réaction d'interconversion réversible des formes *a* et *β* du D-glucopyranose.

Réaction d'interconversion des anomères *α* et *β* du D-glucopyranose (draft)

ALDOSE-1-EPIMERASE-RXN	1 **ALPHA-GLUCOSE** <=> 1 **GLC**	(9 gènes)	Annotation Orthologie

Ainsi, les identifiants du glucose présents dans le draft correspondent à trois groupes de réactions distincts. Pour modéliser son entrée dans le système, plusieurs approches sont envisageables. Nous pouvons utiliser conjointement les trois identifiants, mais auquel cas, comment déterminer l'impact et les bornes de chacun des éléments ? Opter pour l'utilisation d'un seul identifiant offre l'avantage de simplifier la modélisation, mais implique de faire abstraction des groupes de réactions associées aux autres identifiants. Pour pallier ce problème, deux approches sont envisageables : ajouter une réaction d'interconversion artificielle entre tous les identifiants, ou effectuer une substitution d'identifiants en regroupant les composés sous le même terme parent ou enfant.

À ce sujet et avant de poursuivre sur la caractérisation et l'intégration des éléments constituant le milieu nutritif, nous faisons un aparté pour approfondir un aspect lié à la classification des métabolites dans MetaCyc. En effet, une perspective d'amélioration d'***i*Prub22** réside dans la conciliation et l'uniformisation des termes enfants (*i.e.* « *compound* ») et parents (*i.e.* « *compound class* »).

Intuitivement, il semble opportun de remplacer les identifiants de haut niveau par des termes plus spécifiques. Dans notre exemple, les identifiants `Glucose` et `Glucopyranose`, utilisés dans 64 réactions, pourraient être substitués par l'identifiant `GLC`, à quelques exceptions près. Généralement les enzymes ne différencient qu'un seul stéréoisomère ; en conséquence, l'*a*-D-glucopyranose et le *β*-D-glucopyranose auront des réactions spécifiques. En examinant en détail 13 types de sucres et dérivés, cette manipulation se résume à échanger l'identifiant des formes D ou L, selon le sucre considéré, par l'un de ses anomères *a* ou *β*. Néanmoins de plus amples considérations telle que l'évaluation de l'incidence de ces remplacements sur la modélisation sont nécessaires avant de généraliser cette observation.

Afin de préserver la précision et la granularité des termes enfants et parents au sein de la reconstruction, et de garantir leur traçabilité, nous proposons que les substitutions d'identifiants soient effectuées sur le modèle, c'est à dire *via* des instructions MATLAB et non directement encodées dans le fichier `*.sbml`. Cette piste, fastidieuse et réalisée entièrement manuellement, a été l'une des approches suivies lors de la phase de débogage des modèles. Nous avons ainsi effectué des vérifications et des ajustements, lorsque cela était nécessaire, pour différents types de sucres ainsi que pour l'ensemble des métabolites du draft appartenant à la classe des folates (*e.g.* ajout de réactions spontanées réversibles, création de métabolites, *etc.*). Notons que décortiquer l'organisation de seulement 13 sucres et dérivés (*e.g.* formes phosphorylées) conduit à la modification de 145 réactions.

Cependant, cette étape requiert une compréhension globale à la fois des éléments présents dans le draft et de ceux disponibles dans MetaCyc afin d'assurer une cohérence et une pertinence maximales. Nous avons alors observé une légère amélioration dans le nombre de réactions activables et de métabolites productibles, tant du point de vue topologique que dans l'analyse des flux. Citons par exemple le cas d'un des dérivés du sédoheptulose (*i.e.* un monosaccharide composé de sept carbones) dont le terme enfant correspond, sous MetaCyc, à une *dead-end* uniquement consommée, et qui, par substitution d'identifiant avec son terme parent direct, est résolue. Néanmoins, et en l'état, ce ne sont pas ces modifications qui ont permis la fonctionnalité du réseau et comme leur impact sur la modélisation n'a pas été effectuée avec exactitude, les substitutions de métabolites et les ajustements qui en découlent n'ont pas été retenus lors de la publication d'***i*Prub22**. Nous conservons toutefois ces informations pour d'éventuelles mises à jour et améliorations futures du réseau.

2.2.2.3. Assurer la connectivité : modélisation des réactions de transport

Une fois que nous avons sélectionné les identifiants des composés que nous souhaitons intégrer, le second point à aborder concerne la cohérence topologique des composés importés pour être en adéquation avec les contraintes de la modélisation. Le lien entre environnement et milieu extracellulaire est établi *via* des réactions qualifiées d'échange, tandis que la connexion entre les milieux intracellulaire et extracellulaire s'effectue à l'aide de réactions de transport. Ces différents types de réactions sont représentées en **Figure 3-21**.

Ainsi, lorsque des données attestent de l'utilisation d'un composé par l'organisme, que ce soit par sa détection dans le compartiment extracellulaire ou par l'identification d'une réaction de transport associée à une GPR, nous intégrons au besoin une réaction de transport et d'échange dans notre modèle.

Notons toutefois que la distinction entre environnement (*i.e.* limites du système) et milieu extracellulaire, bien que légère, revêt une importance pour la modélisation des champignons en raison de leur système de nutrition. (cf. encart : *Le phénomène d'absorption extracellulaire chez les champignons filamenteux*, page 263)

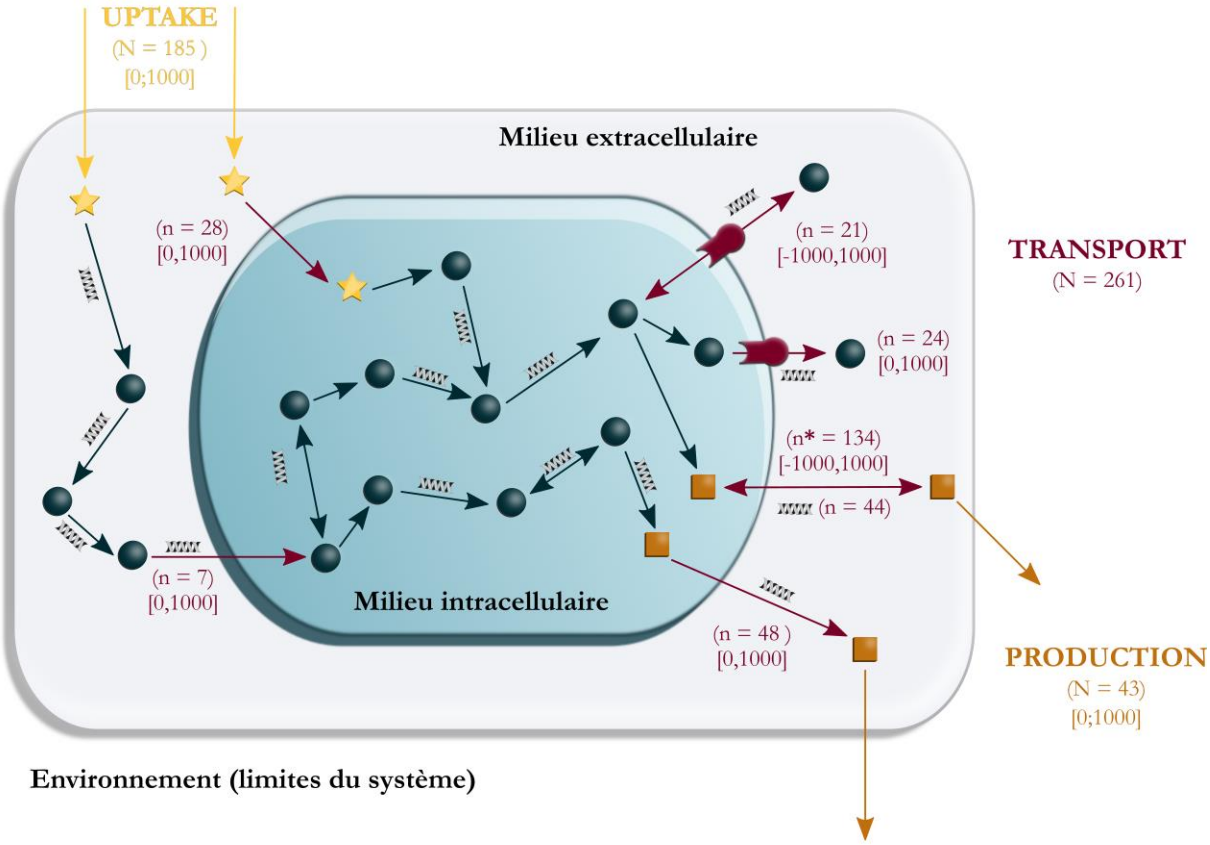


Figure 3-21 : Modélisation des réactions de transport et d'échange dans iPrub22. *Les étoiles jaunes (✶) représentent les composés du medium importés dans le milieu extracellulaire, tandis que les cercles bleus marine (●) symbolisent les métabolites généraux du système. Les carrés orange (■) représentent les métabolites d'intérêt dont la sécrétion peut être monitorée (e.g. pénicillines, roquefortines, etc.). Les deux formes rouge-brune (●) placées sur la membrane plasmique représentent le transport actif médié par une protéine (e.g. symporteurs, antiporteurs, etc.). Les réactions d'échange sont matérialisées par des flèches jaunes (→) représentant les réactions d'import et des flèches orange (→) représentant les réactions de sécrétion. Les flèches rouge-brune (→) indiquent les réactions de transport, tandis que les flèches bleues marine (→) représentent les autres réactions du système. Un brin d'ADN positionné sur le côté des flèches signifie que ces réactions sont soutenues par des informations génomiques. Le nombre de réactions par catégorie est indiqué entre parenthèses (i.e. N pour le nombre total de réactions de l'une des trois catégories, n pour chacune des sous-catégories de transport et n* signifie qu'au total, il existe 134 réactions de transport réversibles, dont 44 sont soutenues par au moins un gène). Les bornes des réactions sont indiquées entre crochets. Contrairement aux réactions d'échange qui servent exclusivement la modélisation, les réactions de transport représentent également des phénomènes biologiques. Ainsi, sur les 261 réactions de transport du réseau, 55 % d'entre elles sont soutenues par au moins une séquence génomique.*

Le phénomène d'absorption extracellulaire chez les champignons filamenteux

Les champignons sont des organismes hétérotrophes qui, par définition, ne peuvent synthétiser eux-mêmes leurs éléments constituants et dépendent de la matière organique existante pour se nourrir. Ainsi, contrairement à certaines bactéries et à la plupart des plantes, qui fixent le dioxyde de carbone atmosphérique, les champignons privilégient l'exploitation de composés organiques complexes comme source de carbone. En outre, les champignons n'ont pas la capacité de capturer l'azote atmosphérique, et, à l'instar des animaux, ils doivent le puiser dans leur environnement *(Raven et al. 2011)*.

En revanche, à l'opposé de la plupart des métazoaires qui ingèrent la nourriture puis la digèrent au sein d'organes spécialisés, les champignons possèdent un système d'enzymes extracellulaires leur permettant de réaliser ce processus dans l'ordre inverse. Les exoenzymes produites par ces champignons sont acheminées vers le milieu extracellulaire où elles dégradent les nutriments environnants au cours de la digestion externe. Cette transformation génère des petites molécules plus aisément absorbables qui sont captées à la surface du mycélium. Citons par exemple la dégradation de polysaccharides insolubles telle que la cellulose en glucose. Ce mécanisme de digestion externe libère dans l'environnement du carbone, de l'azote et divers éléments *(Raven et al. 2011)*.

*Processus matérialisé par la présence de métabolites et de réactions biochimiques dans le compartiment extracellulaire présenté en **Figure 3-21**.*

Au sein de notre reconstruction, une réaction de transport modélise le déplacement d'un métabolite entre les compartiments intracellulaire et extracellulaire. Même si leur nombre est peu élevé, de nombreux concepts de biologie cellulaire y sont associés. Ces réactions, qu'elles soient liées ou non à une association GPR, peuvent avoir un réel sens biologique ou au contraire être artificielles, c'est-à-dire nécessaires à la fonctionnalité et à la consistance du modèle. La **Figure 3-22** illustre les trois principaux mécanismes de transport de métabolites à travers les membranes :

- Le **transport actif**, où le métabolite traverse la membrane à l'encontre de son gradient électrochimique, est thermodynamiquement défavorable. Ce type de transport est médié systématiquement par un transporteur protéique et fait intervenir une molécule énergétique, généralement de l'ATP. Selon l'origine de l'énergie impliquée, nous distinguons le transport actif primaire (*e.g.* pompes), et secondaire (*e.g.* transport couplé).

- Le **transport passif**, quant à lui, permet des déplacements de molécules thermodynamiquement favorables. Il peut être facilité par une protéine (*e.g.* canaux ou transporteur protéique tels que les perméases - phénomène d'osmose avec les aquaporines) ou non (*e.g.* diffusion directe pour des composés liposolubles).

- Le **transport vésiculaire** autorise, en quantité plus conséquente, l'entrée par endocytose (*e.g.* phagocytose, pinocytose, et endocytose médiée par récepteurs) ou la sortie de métabolites par exocytose.

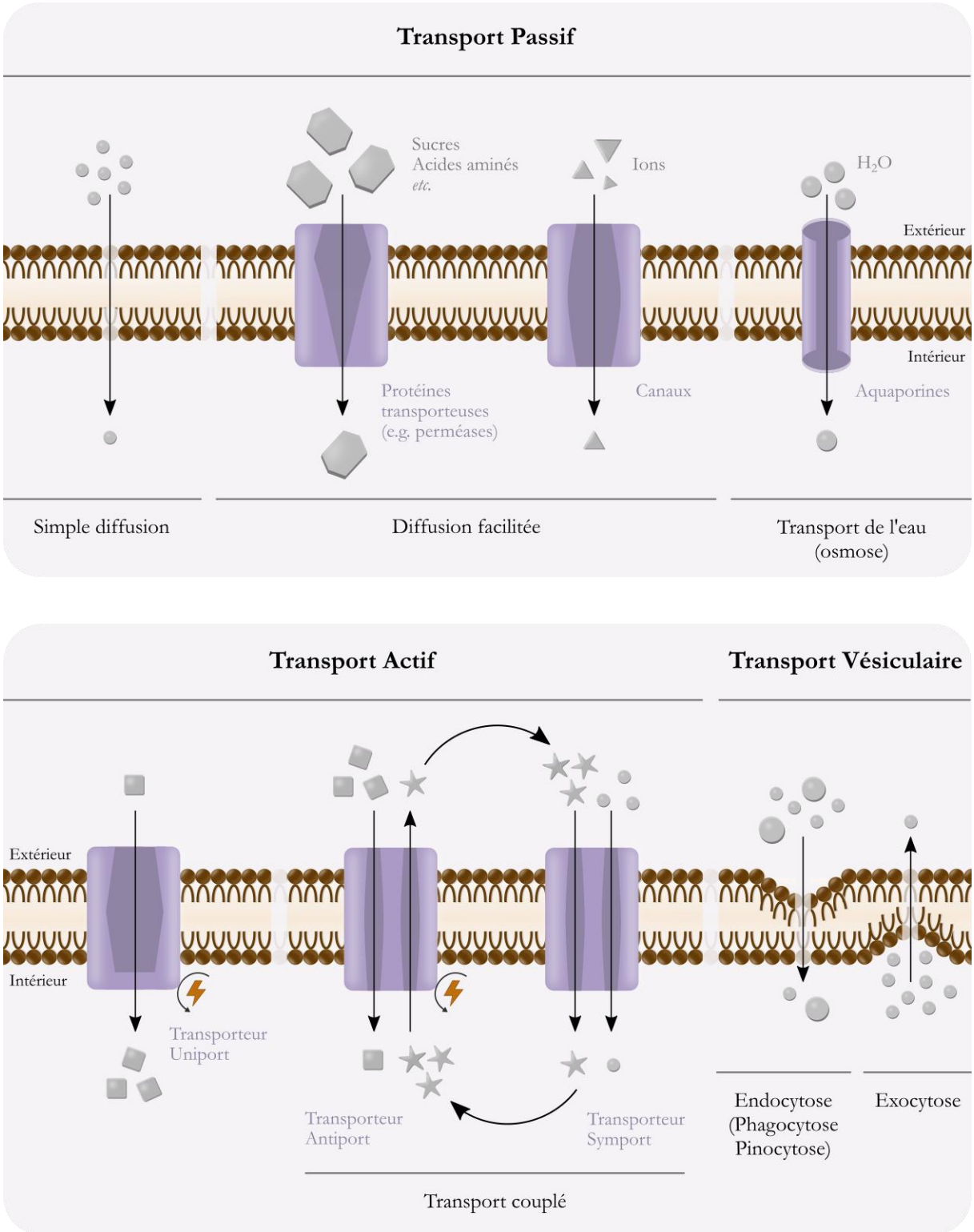


Figure 3-22 : Les mécanismes de transport à travers la membrane plasmique. *Le transfert de métabolites à travers les membranes peut s'effectuer avec ou sans consommation d'énergie. Le transport passif inclut la simple diffusion de métabolites de faible poids moléculaire, tandis que des protéines sélectives assurent une diffusion spécifique de molécules chargées ou hydrophiles de poids moléculaire plus élevé. Le transport actif utilise l'énergie pour transférer les métabolites à l'encontre de leur gradient de concentration. Les termes uniport, symport et antiport, non spécifiques au transport actif, représentent le nombre et le sens du déplacement des métabolites à travers la membrane. En conséquence, dans un réseau métabolique, le transfert d'un métabolite d'un compartiment à un autre est représenté par des réactions spontanées ou biochimiques. Ces réactions peuvent être réversibles ou non et être associées ou non à des séquences génomiques. Inspiré de (Raven et al. 2011).*

La première étape consiste à identifier les réactions de transport dans le draft afin de visualiser les manquements potentiels pré-existants. En explorant la topologie du draft et en se focalisant sur les réactions qui comportent le même métabolite en produit et en réactant, mais dans des compartiments différents (*i.e.* dans le cytosol et dans le milieu extracellulaire), 76 réactions de transport ont été détectées. Ces réactions sont associées à un total de 204 séquences génomiques, dont la localisation des produits, assignée par DeepLoc (cf. Chapitre 2 section *3.2 Annotation fonctionnelle et compartimentation subcellulaire*, page 99 et Chapitre 4 section *1.2.2 L'annotation fonctionnelle du protéome pour guider la filtration des données*, page 358), est la suivante : 107 affectés à la membrane cellulaire, 1 dans le milieu extracellulaire, 12 dans le cytosol et les 84 restant sont répartis entre divers organites.

Un fait non négligeable à prendre en compte pour la poursuite de la curation concerne la composition des fichiers `*.padmet` et `*.sbml`, utilisés pour la reconstruction et disponibles *via* le chargement d'AuReMe. Ces fichiers comprennent uniquement deux compartiments : le cytosol et le milieu extracellulaire. En conséquence, comment ont été intégrées les réactions modélisant le transport intracellulaire tel que, par exemple, un transfert entre le cytosol et les mitochondries ou tout autre organite ?

En toute logique, si l'un des substrats de la localisation possédait une annotation indiquant une localisation intracellulaire autre que le cytosol, elle a été substituée par l'annotation milieu extracellulaire, or nous nous somme appuyés, entre autres, sur la composition du milieu extracellulaire pour la composition des milieux nutritionnels. Une vigilance accrue doit donc être dirigée en ce sens, et, en tenant compte de ces informations, nous avons évalué au cas par cas les 22 réactions pour lesquelles aucun des produits géniques n'était associé à la membrane cellulaire. Onze d'entre elles ont été bloquées pour la modélisation (*i.e.* fermeture des bornes) puisque, toujours dans l'optique de conserver le maximum d'information génomique, elles n'ont pas été supprimées de la reconstruction. Ces informations demeurent primordiales pour la compartimentation intracellulaire et pourront ainsi être utilisée le cas échéant (cf. Chapitre 4 section *1 -Un GSMN d'eucaryote sans compartimentation intracellulaire*, page 353).

Ensuite, en nous appuyant sur les données des sources externes ainsi que sur la topologie de la reconstruction, nous avons ajouté 208 réactions de transport, dont 91 sont soutenues par au moins une séquence génomique. Ainsi, 60 % d'entre elles proviennent d'*i*AL1006, 24 % de PchCyc et les 16 % restantes ont été ajoutées manuellement pour surveiller, par exemple, la sécrétion des métabolites d'intérêt.

Les réactions provenant des sources externes ont été intégrées si, et seulement si, le métabolite concerné était présent dans la topologie du draft. Le cas échant, leur directionnalité a été conservée. Enfin, pour les réactions disposant d'une annotation GPR, la localisation du ou des produits géniques a été systématiquement vérifiée à l'aide de DeepLoc, et seules les séquences avec l'annotation « membrane cellulaire » ont été conservées. L'apport de chacune des sources externes est détaillé ci-dessous, tandis que les origines des réactions de transport et d'échange intégrées à ***i*Prub22** sont présentées en **Figure 3-23**.

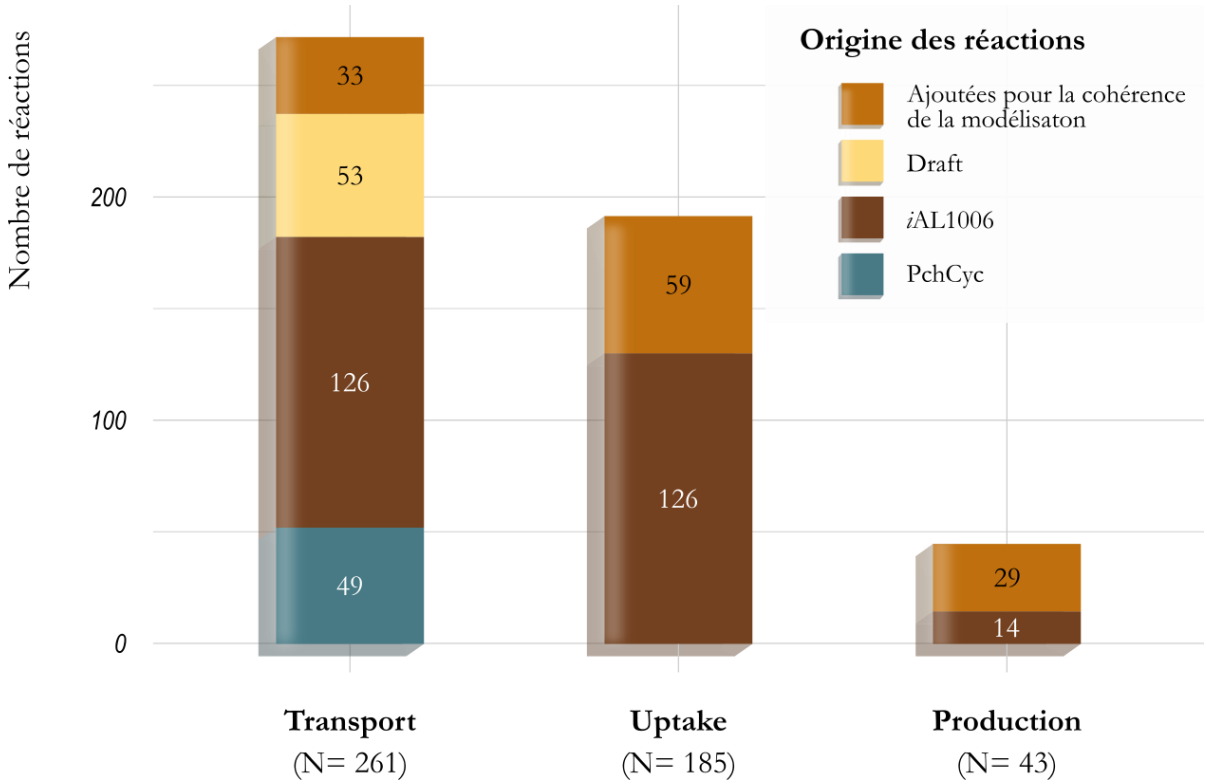


Figure 3-23 : Origine des réactions de transport et d'échange ajoutées à la reconstruction. *Ces réactions ont été sélectionnées en explorant les reconstructions antérieures, en examinant les métabolites présents dans le milieu extracellulaire et en définissant les conditions environnementales à étudier. En ce qui concerne les réactions de transport, nous avons ajouté respectivement 50 % et 84 % des réactions de transport contenues dans PchCyc et iAL1006. Nous avons inclus également 93 % des réactions d'uptake et 54 % des réactions de production de iAL1006. Le rejet d'une réaction s'explique soit par l'absence du métabolite ciblé dans la topologie du draft, soit par l'absence d'une identification du métabolite sur la base de données MetaCyc, soit, lorsqu'elle existe, par une annotation de la localisation de la protéine produite différente de la membrane cellulaire. À noter que l'ensemble des gènes associés à des réactions de transport ont été vérifiés et corrigés en fonction de l'attribution ou non de l'étiquette « membrane cellulaire » fournie par DeepLoc. À l'instar des réactions ajoutées par gap-filling, la justification et la traçabilité de chaque ajout sont mentionnées dans les informations supplémentaires de l'article (S4_file.xls).*

La reconstruction *i*AL1006 bénificie, entre autres, d'une intense curation manuelle axée sur l'étude bibliographique reposant sur 440 articles scientifiques qui soutiennent expérimentalement la majorité des réactions présentes dans le **GSMN**. Nous nous sommes donc fortement appuyés sur sa topologie pour enrichir notre recontruction. Le **GSMN** *i*AL1006 est composé de *1 632* réactions, filtrées dans un premier temps par l'étiquette SBO:0000185 (*i.e.* phénomène de translocation), puis en fonction des étiquettes « *exchange* », « *artificial* », et « *transport* ». Les 451 réactions résultantes, annotées par le terme SBO de référence, sont divisées en 26 réactions de production, 135 d'import, et 290 réactions de transport de métabolites au sens biologique du terme. Ces dernières sont à nouveau discrétisées en fonction du compartiment qui leur est assigné : 88 assurent un transport entre le cytosol et les mitochondries, 54 entre le cytosol et les peroxysomes, et 148 entre le cytosol et le milieu extracellulaire.

Selon les critères que nous avons établis, nous avons sélectionné et ajouté 126 réactions sur les 148 d'intérêt dans notre reconstruction. Toutes les réactions intégrées sont réversibles et un tiers d'entre elles sont soutenues par au moins une séquence génomique. Les réactions dépourvues de soutien génomique correspondent à 73 % à des perméases putatives et à 27 % à un phénomène de diffusion. À l'exception d'une seule réaction, toutes les réactions possédant une association GPR sont annotées comme étant des perméases.

Nous pouvons alors remarquer que l'accès à ces données est largement facilité par les formats de distribution de *i*AL1006 (*i.e.* **SBML** et classeur Excel permettant l'extraction de fichiers aux données structurées), ainsi que par la richesse des annotations associées aux diverses entités de la reconstruction.

En revanche, en l'absence d'annotations spécifiques, récupérer rapidement et exhaustivement les réactions de transport de Prubens s'avère délicat. Toutefois, en filtrant les identifiants MetaCyc commençant par « TRANS » (*i.e.* une portion, non quantifiable de réactions de transport partagent cette similarité au sein de MetaCyc) ou en recherchant les termes « *transport* », « *permease* », « *diffusion* » et « *export* » dans la description des réactions, nous avons identifié respectivement 35, 3, 60, 9 et 3 réactions qui font intervenir 87 gènes distincts. L'analyse des équations des réactions révèle, quant à elle, un total de 200 réactions. De plus, étant donné que Prubens compte 13 compartiments différents attribués automatiquement lors de sa reconstruction, et que notre intérêt se porte exclusivement sur le transfert entre le milieu intracellulaire et le milieu extracellulaire, divers autres niveaux de filtration sont nécessaires. De surcroît, comme la plupart des réactions sélectionnées étant similaires à celles détectées dans *i*AL1006, nous n'avons pas poursuivi l'investigation de cette source.

Le PGDB PchCyc présente une organisation de données structurées au sein de laquelle sont répertoriées 95 réactions de transport, dont 6 sont présentes dans le draft. Nous nous intéressons donc aux 89 réactions restantes, soutenues par 440 gènes et impliquant 87 identifiants de métabolites différents. De prime abord, 20 % de ces métabolites sont absents du draft. Cependant, une analyse détaillée révèle que la majorité de ces absences résulte de différences d'identifiants plutôt que de variations dans la nature des composés. Néanmoins, l'examen de la composition des substrats de chaque réaction ainsi que des associations GPR conduit au rejet de 39 réactions. Il est à noter que sur les 50 réactions restantes, l'une d'entre elles est commune à *i*AL1006. En conséquence, 49 nouvelles réactions ont été intégrées à ***i*Prub22**.

Concernant les réactions provenant de *i*AL1006, sur les 126 réactions sélectionnées, 40 d'entre elles possèdent une association GPR. Nous dénombrons alors 42 gènes distincts. La vérification de la localisation subcellulaire de leurs produits géniques révèle que 7 d'entre eux sont assignés au compartiment « lysosome/vacuole », tandis que les autres sont associés à la membrane cellulaire. Conformément aux règles que nous avons établies, les associations GPR ont alors été nettoyées pour ne garder que les gènes dont les produits géniques sont les plus cohérents. Toutefois, l'un des gènes, associé à trois réactions de transfert de bases azotées, a été conservé afin de permettre des scenari d'enrichissement du modèle. Enfin, la suppression de ces 6 gènes au sein des associations GPR a induit une modification de 34 d'entre elles.

Les 49 réactions sélectionnées dans PchCyc sont toutes associées à au moins un gène ; deux d'entre elles sont réversibles (*i.e.* transport des ions potassium et sodium) et les autres sont irréversibles. Avant le nettoyage des associations GPR, elles étaient associées à 311 séquences génomiques et lors du nettoyage, 96 séquences ont été retirées, induisant une modification de 28 associations GPR. Notons que sur les 215 séquences conservées, 205 sont effectivement associées à la membrane cellulaire selon DeepLoc mais que 10 entrées n'ont pas été annotées car elles étaient absentes des données associés à *P. rubens* Wisconsin 54-1255 téléchargées *via* EnsemblFungi (cf. section *2.1.1* *Description des données utilisées*, page 196).

En résumé, les réactions de transport provenant de *i*AL1006 et PchCyc sont associées à l'origine à un total de 316 séquences génomiques, dont 96 modélisent un probable transport intracellulaire. Parmi les 220 gènes conservés associés aux réactions de transport, 5 proviennent exclusivement de *i*AL1006 et 182 de PchCyc. Le croisement de ces sources permet d'intégrer un total de 175 réactions de transport et 62 associations GPR ont été nettoyées pour ne modéliser que le transport entre le milieu extracellulaire et le cytosol. Ainsi, 35 % des associations GPR ont été modifiées et ce pourcentage est un indicateur potentiel de la faible spécificité de ces réactions. Enfin, la localisation putative des produits géniques de l'ensemble de ces séquences est présentée en **Figure 3-24**.

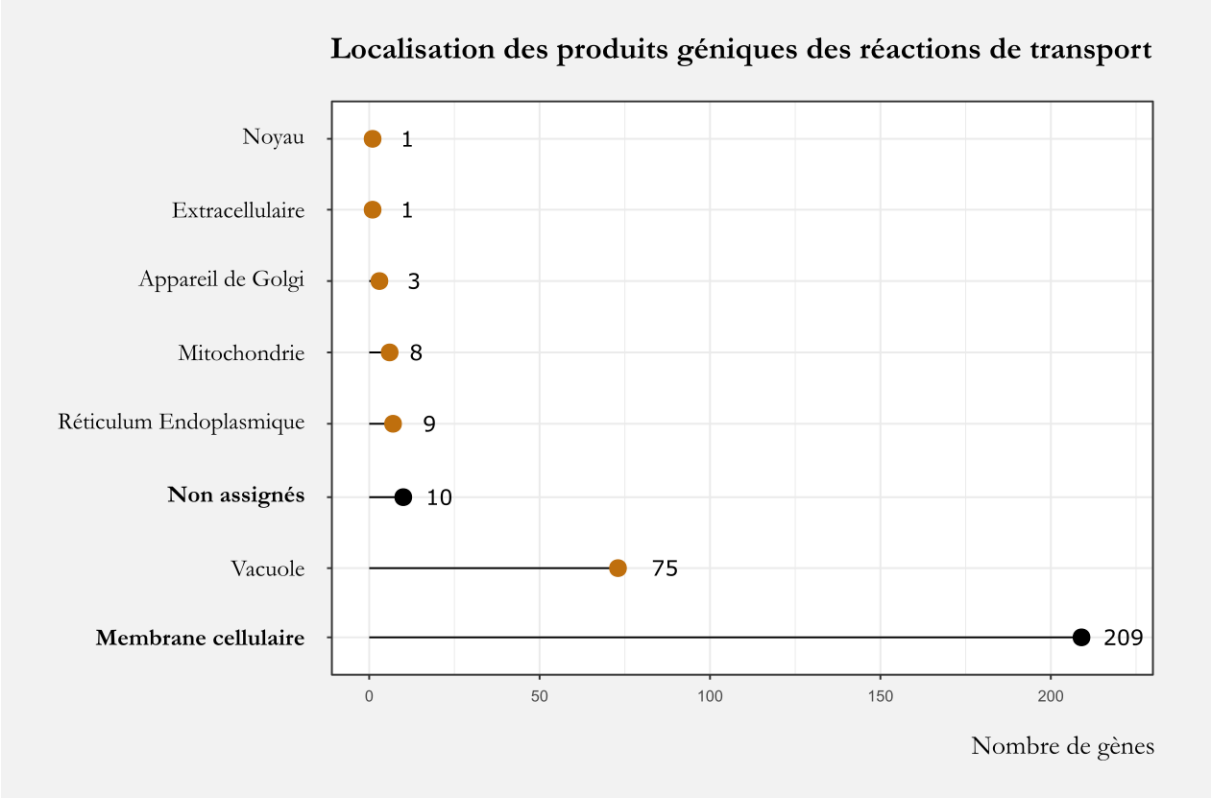


Figure 3-24 : Localisation subcellulaire putative des 316 séquences génomiques associées aux réactions de transport sélectionnées dans iAL1006 et PchCyc. *À l'exception des 10 gènes issus de PchCyc pour lesquels nous n'avons pas de localisation subcellulaire et de la séquence issue de Pc22g19850 dont le produit génique codant pour trois perméases de bases azotées (i.e. adénine : Transport_196, guanine : Transport_197 et cytosine : Transport_198) est associé à la vacuole, les 96 gènes qui ne sont pas assignés à la membrane cellulaire (■) ont été écartés des associations GPR.*

Enfin, pour garantir la discrétisation des réactions présentées dans cette section et intégrer leur granularité fonctionnelle, nous avons attribué à chacune de nos réactions le terme SBO le plus descriptif connu (**Tableau 3-7**). Cette approche simplifie l'accès aux réactions d'intérêt en facilitant la recherche sémantique et contribue ainsi à la compréhension et à l'appropriation d'***i*Prub22**.

Tableau 3-7 : Nature et nombre de termes SBO assignés aux réactions d'*i*Prub22. Parmi les 22 annotations SBO présentes dans ***i*Prub22**, 14 sont utilisées pour caractériser les réactions. Elles sont inscrites en gras. Pour situer ces termes dans l'ontologie, leurs termes parents sont référencés dans le tableau. Comme il n'existe pas de termes pour faire la différence entre réaction d'*uptake* et de production, elles sont regroupées sous le terme de réaction d'échange.

IDENTIFIANTS SBO ET NOMBRE DE RÉACTIONS ASSOCIÉES		
Systems biology representation	SBO:0000000	
Occurring entity representation	SBO:0000231	
Process	SBO:0000375	
Biochemical or Transport reaction	**SBO:0000167**	3
Biochemical reaction	**SBO:0000176**	5 280
Spontaneous reaction	**SBO:0000672**	86
Translocation reaction	**SBO:0000185**	60
Transport reaction	**SBO:0000655**	53
Co-transport reaction	SBO:0000654	
Symporter-mediated transport	**SBO:0000659**	21
Antiporter-mediated transport	**SBO:0000660**	2
Active transport	**SBO:0000657**	38
Passive transport	**SBO:0000658**	108
Encapsulating process	SBO:0000395	
Biomass Production	**SBO:0000629**	1
ATP Maintenance	**SBO:0000630**	1
Pseudo reaction	SBO:0000631	
Exchange reaction	**SBO:0000627**	228
Demand reaction	**SBO:0000627**	37
Sink reaction	**SBO:0000627**	1

2.2.2.4. Préparer le modèle à divers scénarios de nutrition : modélisation des réactions d'échange

Une fois que la communication entre les milieux intracellulaire et extracellulaire est assurée par le biais des réactions de transport, nous nous intéressons aux réactions d'échange. Ce type de réaction, entièrement artificiel, modélise le transfert du même métabolite entre le compartiment extracellulaire et l'environnement. À l'instar de ce qui a était effectué pour *i*AL1006 (*i.e.* 161 réactions d'échanges : 135 d'import et 26 de production) et dans le but de faciliter leur identification, les réactions d'échange ajoutées au modèle sont irréversibles et sont nommées en fonction de leur type. Les réactions d'import (*i.e.* `Uptake_\d{3}`) sont utilisées pour simuler l'environnement nutritionnel, tandis que les réactions de sécrétion (*i.e.* `Production_\d{3}`) permettent de surveiller l'excrétion des métabolites d'intérêt. La reconstruction comprend 43 réactions de production, dont 33 % proviennent de *i*AL1006 et 67 % ont été ajoutés pour monitorer des métabolites spéciaclisés d'intéret (cf. section *3 - Focus sur le métabolisme spécialisé : descriptions de diverses voies de biosynthèse de métabolites spécialisés*, page 308). Concernant les réactions d'import, 68 % proviennent de *i*AL1006 et 32 % ont été intégrées pour répondre aux besoins de la modélisation (**Figure 3-23**). Les 185 réactions d'*uptake* contenues dans ***i*Prub22** sont présentées en **Tableau 3-8**. En somme, il existe une réaction d'*uptake* si l'une des quatre conditions suivantes est vérifiée :

- Le composé est couramment utilisé dans de la culture de champignons filamenteux.
- La réaction d'*uptake* est répertoriée dans *i*AL1006 et le métabolite associé est présent dans le draft dans le compartiment extracellulaire et/ou intracellulaire.
- Il existe une réaction de transport dans le draft, acheminant le métabolite du compartiment extracellulaire vers le compartiment intracellulaire et, dans le milieu extracellulaire, ce métabolite est une impasse métabolique uniquement consommée.
- Le métabolite est considéré comme un « artefact », tels qu'un cofacteur (*i.e.* composé chimique non protéique nécessaire à une enzyme pour catalyser une réaction biochimique qui retrouve son état initial à la fin du cycle catalytique), un métabolite dont l'identifiant est de haut degré dans l'ontologie des composés MetaCyc (*i.e. « compound class »*), ou une *dead-end* uniquement consommée qui présente un intérêt dans le contexte de l'étude.

Enfin et en dernier recours, nous utilisons deux types de réactions fictives pour finaliser la connectivité : les réactions « *demand* » et « *sink* ». L'objectif de ces ajouts est d'accroître la proportion de réactions actives au sein du modèle en débloquant des métabolites spécifiques. Les réactions « *demand* » sont des réactions irréversibles qui représentent les besoins spécifiques de la cellule qui n'ont pu être satisfaits. Elles modélisent ainsi des apports en nutriments, en cofacteurs ou en produits intermédiaires nécessaires à des processus cellulaires ciblés (*e.g.* voies de biosynthèse de métabolites spécialisés). Les réactions « *sink* », quant à elles, sont réversibles et représentent la consommation ou la production de métabolites dont l'origine, ou la destination, est inconnue. Ces deux types de réactions sont principalement utilisées lors de l'évaluation de la reconstruction (*e.g. gap-filling*) et des modèles (*e.g.* analyses de flux) et sont vouées à être remplacées par des réactions fonctionnelles.

Tableau 3-8 : Liste des composés avec réactions d'*uptake* pour la simulation de milieux.
Ces composés sont classés de manière arbitraire en fonction de leur structure ou de leur rôle biologique. Les termes en gras suivis d'une astérisque* proviennent de méthodologies de publication telles que ce brevet sur la production fermentaire de composés d'intérêt industriel en utilisant des milieux chimiquement définis *(Laat et al. 2002)*. Outre cette source, la liste des métabolites a été enrichie par les données du premier **GSMN** de *P. rubens*, *i*AL1006, et par une liste de métabolites extracellulaires présents dans le draft (M_{ext}). Les numéros d'*uptake* servent uniquement à référencer les réactions, leur numérotation a été effectuée sans règle précise et n'a pas de signification particulière. Enfin, le terme artefact dans la catégorie des sources représente trois types de composés : des cofacteurs, des entités au nom générique (*i.e. compounds class*) ou des impasses métaboliques seulement consommées. Bien que non essentiels à la fonctionnalité du modèle, ces éléments permettent d'envisager divers scénarios d'enrichissement du milieu. À noter que ces éléments, en raison de leur caractère artificiel, ne sont pas associés à des réactions de transport mais à des *demand* réactions.

Éléments structuraux

Nom	*Formule brute*	*Identifiants* MetaCyc	*Uptake*	*Source*
● Sources de Carbone				
Glucides (carbohydrates classés par nombre de motifs et complexité structurale croissante)				
Monosaccharides				
Anhydromannitol	$C_6H_{12}O_5$	CPD-9446	004	Artefact
Arabinose ■	$C_5H_{10}O_5$	CPD-15699	016	*i*AL1006
	$C_5H_{10}O_5$	L-ARABINOSE	017	*i*AL1006 - M_{ext}
D-arabinitol	$C_5H_{12}O_5$	CPD-355	043	*i*AL1006
Fructose* ■	$C_6H_{12}O_6$	BETA-D-FRUCTOSE	045	*i*AL1006
	$C_6H_{12}O_6$	CPD-15382	046	*i*AL1006 - M_{ext}
Galactose	$C_6H_{12}O_6$	ALPHA-D-GALACTOSE	048	*i*AL1006
Glucose* ■	$C_6H_{12}O_6$	GLC	026	*i*AL1006 - M_{ext}
	$C_6H_{12}O_6$	Glucose	070	*i*AL1006 - M_{ext}
	$C_6H_{12}O_6$	ALPHA-GLUCOSE	015	*i*AL1006 - M_{ext}
L-arabinitol	$C_5H_{12}O_5$	L-ARABITOL	088	*i*AL1006
L-iditol	$C_6H_{14}O_6$	CPD-369	099	*i*AL1006
Mannitol	$C_6H_{14}O_6$	MANNITOL	054	*i*AL1006
Mannose	$C_6H_{12}O_6$	MANNOSE	055	*i*AL1006
Ribose	$C_5H_{10}O_5$	RIBOSE	057	*i*AL1006
Ribulose	$C_5H_{10}O_5$	L-RIBULOSE	108	*i*AL1006
Sorbose	$C_6H_{12}O_6$	CPD-9570	110	*i*AL1006
Xylitol	$C_5H_{12}O_5$	XYLITOL	182	*i*AL1006
Xylose	$C_5H_{10}O_5$	D-Xylose	058	*i*AL1006 - M_{ext}
Disaccharides				
Cellobiose	$C_{12}H_{22}O_{11}$	CELLOBIOSE	033	*i*AL1006 - M_{ext}
Lactose* ■	$C_{12}H_{22}O_{11}$	CPD-15972	086	*i*AL1006
	$C_{12}H_{22}O_{11}$	Alpha-lactose	085	*i*AL1006 - M_{ext}
Maltose	$C_{12}H_{22}O_{11}$	MALTOSE	118	*i*AL1006 - M_{ext}
Mélibiose	$C_{12}H_{22}O_{11}$	MELIBIOSE	122	*i*AL1006 - M_{ext}
Sucrose*	$C_{12}H_{22}O_{11}$	SUCROSE	165	*i*AL1006 - M_{ext}
Tréhalose	$C_{12}H_{22}O_{11}$	TREHALOSE	014	*i*AL1006 - M_{ext}
Trisaccharides				
Maltotriose	$C_{18}H_{32}O_{16}$	MALTOTRIOSE	119	*i*AL1006
Raffinose	$C_{18}H_{32}O_{16}$	CPD-1099	154	*i*AL1006 - M_{ext}
Polysaccharides				
Amidon*	$(C_6H_{10}O_5)n$	Starch	160	*i*AL1006
Cellulose	$(C_6H_{10}O_5)n$	CELLULOSE	034	*i*AL1006 - M_{ext}
Glucane	$(C_6H_{10}O_5)n$	1-3-beta-D-Glucans	003	*i*AL1006 - M_{ext}
Glycogène	$(C_6H_{10}O_5)n$	Glycogens	074	*i*AL1006 - M_{ext}
Inuline*	$(C_6H_{10}O_5)n$	Inulin	082	-
Maltodextrines*	$(C_6H_{10}O_5)n.H_2O$	Maltodextrins	117	-
Mannanes	$(C_6H_{10}O_5)n$	Mannans	121	*i*AL1006
Xylanes	$(C_5H_8O_4)n$	Xylans	181	*i*AL1006 - M_{ext}

■ Cf. encart : *Un « même » composé mais des identifiants différents, lequel choisir ? Illustration avec l'entité glucose*, page 259

Tableau 3-8 (suite)

Alcools*

Éthanol*	C_2H_6O	ETOH	060	*i*AL1006
Glycérol*	$C_3H_8O_3$	GLYCEROL	072	*i*AL1006
Méthanol*	CH_4O	METOH	123	*i*AL1006
Myo-inositol	$C_6H_{12}O_6$	MYO-INOSITOL	124	*i*AL1006

Acides gras

Saturés à chaîne courte ($C_{3:4}$)

Butyrate	$C_4H_7O_2$	BUTYRIC_ACID	028	*i*AL1006
Propionate*	$C_3H_5O_2$	PROPIONATE	149	*i*AL1006

Saturés à chaîne moyenne ($C_{5:12}$)

Décanoate	$C_{10}H_{19}O_2$	CPD-3617	044	*i*AL1006
Heptanoate	$C_7H_{13}O_2$	CPD-7619	079	*i*AL1006
Laurate	$C_{12}H_{23}O_2$	DODECANOATE	092	*i*AL1006
Nonanoate	$C_9H_{17}O_2$	CPD-8505	135	*i*AL1006
Octanoate	$C_8H_{15}O_2$	CPD-195	137	*i*AL1006
Valérate	$C_5H_9O_2$	VALERATE	178	*i*AL1006

Saturés à chaîne longue ($C_{14:18}$)

Heptadecanoate	$C_{17}H_{33}O_2$	CPD-7830	078	*i*AL1006
Myristate	$C_{14}H_{27}O_2$	CPD-7836	125	*i*AL1006
Palmitate* ▲	$C_{16}H_{31}O_2$	PALMITATE	141	*i*AL1006
Pentadecanoate	$C_{15}H_{29}O_2$	CPD-8462	143	*i*AL1006
Stéarate* ▲	$C_{18}H_{35}O_2$	STEARIC_ACID	161	*i*AL1006

Saturés à chaîne très longue ($C_{>20}$)

Icosanoate	$C_{20}H_{39}O_2$	ARACHIDIC_ACID	081	*i*AL1006

Monoinsaturés à chaîne longue ($C_{14:18}$)

Oléate* ▲	$C_{18}H_{33}O_2$	OLEATE-CPD	138	M_{ext}

Polyinsaturés à chaîne longue ($C_{14:18}$)

Linolénate* ▲	$C_{18}H_{29}O_2$	LINOLENIC_ACID	018	

▲ Ces quatre acides gras sont des composés couramment retrouvés dans l'huile de soja, une huile végétale utilisée pour enrichir les milieux de culture des champignons filamenteux. Deux acides gras insaturés, l'**acide érucastique** (CPD-12189) et l'**acide γ-linoléique** (CPD-8117) complètent sa composition. Étant donné que ces éléments ne figuraient pas dans la topologie initiale du draft, ils n'ont pas été intégrés à la modélisation. En revanche, l'huile de soja contient également des **vitamines E** (Vitamin-E) et des **vitamines K** (CPD-11501). Ces identifiants correspondent à des « *compounds class* » et sont absents du draft initial. Néanmoins, les identifiants Vitamin-E et CPD-11501 englobent respectivement 11 et 25 entités, dont 4 de chaque sont retrouvées dans le draft. Pour la vitamine E nous observons les formes *α*, *β*, *δ* et *γ* du tocophérol (ALPHA-TOCOPHEROL, BETA-TOCOPHEROL, DELTA-TOCOPHEROL, GAMA-TOCOPHEROL), tandis que pour la vitamine K, le phylloquinone (2-METHYL-3-PHYTYL-14-NAPHTHOQUINONE), le phylloquinol (CPD-12831), le ménadione et la classe des ménaquinones sont présents.

Acides Carboxyliques

Monocarboxylates

Acétate*	$C_2H_3O_2$	ACET	011	*i*AL1006 - M_{ext}
Glycolate	$C_2H_3O_3$	GLYCOLLATE	075	*i*AL1006
Lactate (acide lactique)	$C_3H_5O_3$	L-LACTATE	001	*i*AL1006
Phénoxyacétate*	$C_8H_7O_3$	CPD-10902	144	*i*AL1006
Phénylacétate*	$C_8H_7O_2$	PHENYLACETATE	145	*i*AL1006
Pyruvate	$C_3H_3O_3$	PYRUVATE	151	*i*AL1006
Quinate	$C_7H_{11}O_6$	QUINATE	152	Artefact

Dicarboxylates

α-cétoglutarate	$C_5H_4O_5$	2-KETOGLUTARATE	005	*i*AL1006
Fumarate	$C_4H_2O_4$	FUM	068	*i*AL1006
Malate	$C_4H_4O_5$	MAL	002	*i*AL1006

Tableau 3-8 (suite)

Oxalate	C_2O_4	OXALATE	139	*i*AL1006
Oxaloacétate	$C_4H_2O_5$	OXALACETIC_ACID	140	*i*AL1006
Pimélate	$C_7H_{10}O_4$	CPD-205	147	*i*AL1006
Succinate	$C_4H_4O_4$	SUC	164	*i*AL1006
Tartrate	$C_4H_4O_6$	TARTRATE	170	Artefact
Tricarboxylates				
Isocitrate	$C_6H_5O_7$	THREO-DS-ISO-CITRATE	083	*i*AL1006
Lactones (esters cycliques)				
Cellobiono-1,5-lactone	$C_{12}H_{20}O_{11}$	CPD-379	032	*i*AL1006 - M_{ext}
Galactono-1,4-lactone	$C_6H_{10}O_6$	D-GALACTONO-1-4-LACTONE	047	Artefact
Glucono-1,5-lactone	$C_6H_{10}O_6$	GLC-D-LACTONE	049	*i*AL1006 - M_{ext}
Glycoside Phénolique				
Abrutine	$C_{12}H_{16}O_7$	HYDROQUINONE-O-BETA-D-GLUCOPYRANOSIDE	024	M_{ext}
● Sources d'Azote				
Composés Inorganiques				
Ammoniac*	H_3N	AMMONIA	130	*i*AL1006
Ammonium*	H_4N	AMMONIUM	019	M_{ext}
Cyanate	CNO	CPD-69	042	*i*AL1006
Nitrate	NO_3	NITRATE	133	*i*AL1006
Nitrite	NO_2	NITRITE	134	*i*AL1006 - M_{ext}
Composés Organiques				
Choline*	$C_5H_{14}NO$	CHOLINE	039	*i*AL1006
Éthyl-nitronate	C_2H_5NO	CPD-320	061	*i*AL1006
Urée*	CH_4N_2O	UREA	176	*i*AL1006 - M_{ext}
Composants d'Acides Nucléiques				
Bases puriques				
Adénine*	$C_5H_5N_5$	ADENINE	012	*i*AL1006 - M_{ext}
Guanine*	$C_5H_5N_5O$	GUANINE	076	*i*AL1006 - M_{ext}
Xanthine	$C_5H_4N_4O_2$	XANTHINE	180	*i*AL1006 - M_{ext}
Hypoxanthine*	$C_5H_4N_4O$	HYPOXANTHINE	080	*i*AL1006
Bases pyrimidines				
Uracile*	$C_4H_4N_2O_2$	URACIL	175	*i*AL1006
Ribonucléoside purique				
Adénosine*	$C_{10}H_{13}N_5O_4$	ADENOSINE	013	M_{ext}
Ribonucléoside pyrimidique				
Uridine*	$C_9H_{12}N_2O_6$	URIDINE	177	*i*AL1006 - M_{ext}
● Sources potentielles de Carbone et d'Azote				
Acides Aminés				
Protéinogènes				
Alanine	$C_3H_7NO_2$	L-ALPHA-ALANINE	087	*i*AL1006
Arginine	$C_6H_{15}N_4O_2$	ARG	089	*i*AL1006
Asparagine	$C_4H_8N_2O_3$	ASN	090	*i*AL1006
Aspartate	$C_4H_6NO_4$	L-ASPARTATE	091	*i*AL1006 - M_{ext}
Cystéine	$C_3H_7NO_2S$	CYS	094	*i*AL1006
Glutamate*	$C_5H_8NO_4$	GLT	095	*i*AL1006

Tableau 3-8 (suite)

Glutamine	$C_5H_{10}N_2O_3$	GLN	096	*i*AL1006
Glycine	$C_2H_5NO_2$	GLY	073	*i*AL1006
Histidine	$C_6H_9N_3O_2$	HIS	097	*i*AL1006
Isoleucine	$C_6H_{13}NO_2$	ILE	101	*i*AL1006
Leucine	$C_6H_{13}NO_2$	LEU	102	*i*AL1006
Lysine*	$C_6H_{15}N_2O_2$	LYS	103	*i*AL1006 - M_{ext}
Méthionine	$C_5H_{11}NO_2S$	MET	104	*i*AL1006
Phénylalanine	$C_9H_{11}NO_2$	PHE	106	*i*AL1006
Proline	$C_5H_9NO_2$	PRO	107	*i*AL1006
Sérine	$C_3H_7NO_3$	SER	109	*i*AL1006
Thréonine	$C_4H_9NO_3$	THR	111	*i*AL1006
Tryptophane	$C_{11}H_{12}N_2O_2$	TRP	112	*i*AL1006
Tyrosine	$C_9H_{11}NO_3$	TYR	113	*i*AL1006 - M_{ext}
Valine	$C_5H_{11}NO_2$	VAL	114	*i*AL1006

Non-protéinogènes

β-alanine	$C_3H_7NO_2$	B-ALANINE	025	*i*AL1006
Citrulline	$C_6H_{13}N_3O_3$	L-CITRULLINE	093	*i*AL1006 - M_{ext}
γ-aminobutyrate	$C_4H_9NO_2$	4-AMINO-BUTYRATE	069	*i*AL1006
Homocystéine	$C_4H_9NO_2S$	HOMO-CYS	098	*i*AL1006
L-2-aminoadipate	$C_6H_{10}NO_4$	CPD-468	084	*i*AL1006
Ornithine	$C_5H_{13}N_2O_2$	L-ORNITHINE	105	*i*AL1006 - M_{ext}

Sucres Aminés

Chitobiose	$C_{12}H_{26}N_2O_9$	CPD-13545	036	*i*AL1006 - M_{ext}
Glucosamine	$C_6H_{14}NO_5$	GLUCOSAMINE	050	*i*AL1006 - M_{ext}
N-acétylglucosamine	$C_8H_{15}NO_6$	N-acetyl-D-glucosamine	126	*i*AL1006 - M_{ext}

Aminoglycanes

Chitine	$H_2O(C_8H_{13}NO_5)_n$	CHITIN	035	*i*AL1006 - M_{ext}
Chitosane	$(C_8H_{13}NO_5)_n$	Chitosan	037	*i*AL1006 - M_{ext}

Aminobenzoates

Anthranilate	$C_7H_6NO_2$	ANTHRANILATE	023	*i*AL1006
P-aminobenzoate	$C_7H_6NO_2$	P-AMINO-BENZOATE	006	*i*AL1006

Pyridinedicarboxylate

Quinolinate	$C_7H_3NO_4$	QUINOLINATE	153	*i*AL1006

- Sources d'Oxygène, d'Hydrogène, de Phosphore et de Soufre

Éléments structuraux

Dioxyde de carbone*	CO_2	CARBON-DIOXIDE	040	M_{ext}
Eau*	H_2O	WATER	179	M_{ext}
Oxygène*	O_2	OXYGEN-MOLECULE	136	*i*AL1006 - M_{ext}
Soufre*	S	Elemental-Sulfur	169	*i*AL1006

Anions

Formate	CHO_2^-	FORMATE	067	*i*AL1006 - M_{ext}
Ion chlorure	Cl^-	CL-	038	M_{ext}
Phosphate*	HPO_4^{2-}	Pi	146	*i*AL1006 - M_{ext}
Sulfate*	SO_4^{2-}	SULFATE	166	*i*AL1006
Sulfite	SO_3^{2-}	SO3	168	*i*AL1006

Donneur d'Hydrogène

Donneur d'hydrogène	$A\text{-}H_2$	Donor-H2	056	Artefact
Sulfide	H_2S	HS	167	*i*AL1006

Molécule réactive

Peroxyde d'hydrogène*	H_2O_2	HYDROGEN-PEROXIDE	077	*i*AL1006 - M_{ext}

Tableau 3-8 (suite)

Éléments catalytiques (constituants des cofacteurs inorganiques et organiques)

Ions métalliques

Cations inorganiques

Cuivre*	Cu^{+}	CU+	041	M_{ext}
Fer (II)*	Fe^{2+}	FE+2	062	-
Ion calcium*	Ca^{2+}	CA+2	029	M_{ext}
Ion ferrique*	Fe^{3+}	FE+3	063	M_{ext}
Ion potassium*	K^{+}	K+	148	M_{ext}
Magnésium*	Mg^{2+}	MG+2	116	-
Manganèse*	Mn^{2+}	MN+2	120	M_{ext}
Sodium*	Na_{+}	NA+	158	M_{ext}
Zinc*	Zn^{2+}	ZN+2	183	M_{ext}

Les éléments suivants sont également cités dans la littérature : le **bore***, le **cobalt*** (CO+1, CO+2 et CO+3), le **molybdène*** (MO+2 et CPD0-2009) et le **sélénium*.** Le bore et le sélénium ne possèdent pas de référencement sous MetaCyc et ne peuvent donc être inclus automatiquement à la reconstruction. Les identifiants du cobalt et du molybdène sont quant à eux absents des composés présents initialement dans le draft, et *de facto* n'ont pas été ajoutés à la modélisation. Pour la même raison, les identifiants CU+2 et FE+4 des ions métalliques cuivre et fer n'ont pas été intégrés. Concernant le calcium, l'identifiant CA2+ est présent dans le draft mais il n'est associé qu'a une seule réaction annotée ATPase transporteuse d'ion calcium (TRANS-RXN-194). Au moment de la reconstruction, outre cette réaction, aucune autre réaction biochimique pertinente pour notre oganisme n'était liée à cet ion (*i.e.* CA2+ est uniquement impliqué dans quatre réactions liées au phénomène de bioluminescence des méduses et coraux). Néanmoins, nous avons choisi d'associer des réactions de modélisation d'import de calicum en prévision de futures mises à jour de MetaCyc.

Vitamines et associés (groupe de composés organiques sans lien structurel)

Vitamines

Acide pantothénique* *(Vitamine B5)*	$C_9H_{16}NO_5$	PANTOTHENATE	142	
Biotine* *(Vitamine B7)*	$C_{10}H_{15}N_2O_3S$	BIOTIN	027	
Carnitine *(Vitamine BT)*	$C_7H_{15}NO_3$	CARNITINE	031	Artefact - M_{ext}
Folates* *(Vitamine B9)*	$C_{19}H_{17}N_7O_5R$	Folates	066	Artefact
Nicotinamide* *(Vitamine B3)*	$C_6H_6N_2O$	NIACINAMIDE	131	*i*AL1006
Nicotinate* *(Vitamine B3)*	$C_6H_4NO_2$	NIACINE	132	*i*AL1006
Pyridoxal* *(Vitamine B6)*	$C_8H_9NO_3$	PYRIDOXAL	150	
Riboflavine* *(Vitamine B2)*	$C_{17}H_{19}N_4O_6$	RIBOFLAVIN	157	
Thiamine* *(Vitamine B1)*	$C_{12}H_{17}N_4OS$	THIAMINE	171	*i*AL1006

Composants de Vitamines

5-methyltetrahydropteroyl tri-L-glutamate* *(Composant des folates)*	$C_{30}H_{35}N_9O_{12}$	CPD-1302	007	
FMN - flavine mononucléotide *(Composant de la riboflavine)*	$C_{17}H_{18}N_4O_9P$	FMN	065	*i*AL1006 - Artefact

Cofacteurs contenant du soufre

Acide lipoïque*	$C_8H_{13}O_2S_2$	LIPOIC-ACID	100	-
Glutathione	$C_{10}H_{16}N_3O_6S$	GLUTATHIONE	071	*i*AL1006

Tableau 3-8 (suite)

Cofacteurs d'oxydoréduction

Liés au transport d'électrons à travers les membranes

Ferrocytochrome-B5	$C_{34}H_{30}FeN_4O_4^{2-*}$	FERROCYTOCHROME-B5	064	Artefact
Quinol de transfert d'électrons	$C_6H_2O_2R_4$	ETR-Quinols	020	Artefact
Ubiquinone-6	$C_{39}H_{58}O_4$	UBIQUINONE-6	173	Artefact
Ubiquinone-8	$C_{49}H_{74}O_4$	UBIQUINONE-8	174	Artefact

Liés au transfert d'électrons (Accepteur/Donneur)

Dinucléotides

NADH-P-OR-NOP	$C_{21}H_{25}N_7O_{17}P_3$ ou $C_{21}H_{26}N_7O_{17}P_3$	NADH-P-OR-NOP	127	Artefact
NADP	$C_{21}H_{25}N_7O_{17}P_3$	NADP	128	Artefact - M_{ext}
NADPH	$C_{21}H_{26}N_7O_{17}P_3$	NADPH	129	Artefact - M_{ext}

Protéines contenant des clusters fer-soufre

Ferrédoxine oxydée	$Fe_2S_2^{2+*}$	Oxidized-ferredoxins	022	Artefact - M_{ext}
Ferrédoxines réduites	$Fe_2S_2^{+*}$	Reduced-ferredoxins	156	Artefact - M_{ext}

Protéines de régulation du potentiel redox

Flavoprotéine de transfert d'électrons oxydée	$C_{13}H_{10}N_4O_2$*	ETF-Oxidized	021	Artefact
Red-NADPH-hémoprotéine-réductase	$C_{12}H_{11}N_4O_2$*	Red-NADPH-Hemoprotein-Reductases	155	Artefact
Thioredoxine réduite	$C_6H_9NO_2S_2$*	Red-Thioredoxin	009	Artefact

Des éléments artéfactuels de modélisation

★ Les éléments présentés dans cette section sont des termes de haut niveau dans la hiérarchisation des métabolites. Ils représentent des classes de composés et ils ont été principalement utilisés comme composés d'initiation pour la réalisation du *gap-filling*. De plus, en raison de l'absence d'identifiants adéquats et cohérents, l'intégralité de la biomasse du **GSMN** *i*AL1006 n'a pu être intégrée dans ***i*Prub22** (cf. section *2.2.3* *Modélisation de la croissance de l'organisme, la réaction de biomasse*, page 277). Les éléments écartés de notre réaction de biomasse sont des ensembles de composés au nom extrêmement générique. Les éléments présentés ci-dessous ont néanmoins été conservés afin d'enrichir, au besoin, la modélisation. Afin de mettre en avant leur caractère artificiel, ils sont liés au compartiment intracellulaire par des *demand* réactions.

Diverses classes de composées

Acides aminés	$C_2H_4NO_2$*	Amino-Acids-20	184	Artefact
Carboxylates	$C_nH2_{n-1}O_2$	Carboxylates	030	Artefact
Dicarboxylate	$C_{n-1}H_{2n-3}O_4$	dicarboxylate	052	Artefact
Hexoses	$C_6H_{12}O_6$	D-Hexoses	051	Artefact
Nucléosides	$C_5H_8O_3R_2$	Nucleosides	185	Artefact
Stérols	$C_{19}H_{31}OR$	Sterols	162	Artefact

Le Sous-Système de Biosynthèse des Acides Gras

Protéine porteuse d'acyle	HSR	ACP	010	Artefact

Le Sous-Système de Biosynthèse des Esters de Stérols

Esters stéaryliques	$C_{20}H_{30}O_2$*	Steryl-Esters	163	Artefact
Oléate d'ergostéryl	$C_{46}H_{76}O_2$	CPD-16017	059	Artefact

Le Sous-Système de Biosynthèse des Glycérides

Triacylglycérides	$C_6H_5O_6R_3$	Triacylglycerides	172	Artefact

Un élément du Sous-Système de Biosynthèse des Lipides

Sphingomyélines	$C_{24}H_{48}N_2O_6P$*	Sphingomyelins	159	Artefact

Tableau 3-8 (suite et fin)

Quelques intrus…

★ Cf. encart : *Les risques de la curation manuelle - Des intrus se sont glissés dans la liste des composés d'initiation*, page 245

⚠	6-Déméthylstérigmatocystine	$C_{17}H_{10}O_6$	6-DEMETHYLSTERIGMATOCYSTIN	008
	Dioscine	$C_{45}H_{72}O_{16}$	CPD-15438	053
	Lysergate	$C_{16}H_{16}N_2O_2$	CPD-12364	115

2.2.3. MODÉLISATION DE LA CROISSANCE DE L'ORGANISME, LA RÉACTION DE BIOMASSE

2.2.3.1. Définir les fondements de la réaction de biomasse : biomolécules et énergie

La simulation du phénomène de croissance d'un organisme s'effectue par la modélisation de la biomasse, un composé artefactuel résultant de la somme des macromolécules du système et de l'énergie requise pour leur production. Cette énergie est matérialisée par une réaction dénommée « *growth-associated maintenance* » (GAM) qui simule les besoins énergétiques lors de la polymérisation des macromolécules (*i.e.* réactions d'hydrolyse de l'ATP au sein des divers systèmes). Par ailleurs, les coûts de maintenance représentant la quantité d'énergie nécessaire à l'organisme indépendamment de la croissance, sont, quant à eux, représentés par la réaction « *non growth-associated maintenance* » (NGAM) *(Thiele et Palsson 2010a)*. Un résumé graphique de ces concepts est présenté en **Figure 3-25**.

Dans un système vivant, la charpente des biomolécules est essentiellement composée de carbone et les fonctions et propriétés de ces molécules organiques sont déterminées par leurs groupements fonctionnels. Nous dénombrons alors quatre grands types de macromolécules biologiques constituant les matériaux de construction : les glucides, les lipides, les protéines et les acides nucléiques *(Raven et al. 2011)*. Une description succincte des éléments constitutifs de ces classes est présentée dans l'encart : *Les quatre types de macromolécules*, page 278. Au sein des cellules, la principale source d'énergie provient de la dégradation de l'ATP en ADP et phosphate inorganique. Cette énergie est transférée aux réactions biochimiques qui dépendent des changements d'énergie (*i.e.* une réaction se définissant par la formation ou la rupture de liaisons chimiques). Le métabolisme est composé de réactions organisées en voies biochimiques divisées en deux grands types : les réactions cataboliques qui libèrent de l'énergie en dégradant des molécules, et les réactions anaboliques qui utilisent de l'énergie pour la biosynthèse de molécules. En conséquence, la réaction de biomasse résulte de l'équilibre dynamique entre les processus cataboliques qui fournissent l'énergie et les précurseurs, et les processus anaboliques, qui utilisent cette énergie et ces éléments pour construire et maintenir la structure cellulaire.

Les quatre types de macromolécules d'après *(Raven et al. 2011)*

La compréhension des macromolécules biologiques est fondamentale pour la modélisation des systèmes biologiques, notamment dans le cadre de la croissance des organismes. Ces macromolécules essentielles pour diverses fonctions cellulaires sont constituées de glucides, de lipides, de protéines et d'acides nucléiques.

À l'exception de la classe des lipides, construits principalement à partir de glycérol et d'acides gras, les macromolécules sont des polymères résultant de l'assemblage de monomères (*i.e.* les polysaccharides sont formés de monosaccharides ; les acides nucléiques, ADN et ARN, sont formés par les nucléotides ; les protéines ou polypeptides sont formés d'acides aminés). La dégradation de ces polymères s'effectue généralement par des réactions d'hydrolyse (*i.e.* ajout d'une molécule d'eau conduisant à la rupture de liaisons entre sous-unités) et, à l'inverse, leur biosynthèse s'effectue par la libération d'une molécule d'eau lors de l'ajout d'un monomère.

Les **glucides**, ou carbohydrates, sont représentés par des sucres simples tels que le glucose, et leur agglomération forme des glucides complexes employés pour le stockage de l'énergie (*e.g.* amidon et glycogène) ou pour le maintien de la structure cellulaire (*e.g.* chitine - polymère de *N*-acétylglucosamine, une version modifiée de glucose - dans les parois des champignons filamenteux). Notons que les plus petits glucides répertoriés sont des monosaccharides composés de trois carbones et, à travers l'organisme, les glucides sont communément transportés sous forme de disaccharides (*e.g.* saccharose, ou maltose), moins facilement métabolisés lors de leur transport.

Les **protéines** sont des polymères d'acides aminés aux structures et aux fonctions variées. Dans le cadre de la modélisation du métabolisme, nous nous intéressons préférentiellement aux enzymes en raison de leur rôle catalytique. Cependant, il existe une large gamme de classes de protéines telles que : les protéines de transport qui facilitent le mouvement sélectif de molécules à travers les membranes cellulaires (*e.g.* transporteurs membranes) ; les protéines de défense qui inhibent la croissance d'autres micro-organismes compétiteurs (*e.g.* toxines) ; les protéines de mouvement qui sont impliquées dans les processus de motilité cytoplasmique (*e.g.* actine, myosine, tubuline) ; les protéines de soutien et de structure qui confèrent aux parois une résistance aux stress environnementaux (*e.g.* chitine) ; les protéines de régulation qui contrôlent l'expression des gènes (*e.g.* facteurs de transcription, protéines régulatrices de la phosphorylation comme les kinases) ; les protéines de stockage et de réserve qui régulent la concentration ionique et d'éléments essentiels (*e.g.* complexants d'ions comme la ferritine, protéines de réserve de carbone et d'azote).

Les classes principales de **lipides** sont les triglycérides, les phospholipides et les stéroïdes. Les lipides sont un ensemble de molécules caractérisées par leur propriété hydrophobe, leur conférant une force cohésive à l'origine de la base de la structure des membranes cellulaires. Le squelette des phospholipides et des triglycérides est composé d'un glycérol associé à des acides gras alors que les stéroïdes se caractérisent par une succession de cycles carbonés. Les triglycérides sont des molécules énergétiques au potentiel bien plus élevé que celui des glucides, tandis que les phospholipides et les stéroïdes, notamment l'ergostérol, sont impliqués dans la composition et dans la fonctionnalité des membranes cellulaires.

Les **acides nucléiques**, quant à eux, sont porteurs de l'information génétique et contiennent des bases azotées, du phosphate et du ribose. L'ADN est nécessaire à l'encodage des gènes alors que l'ARN permet leur expression. Ces molécules de l'information sont des polymères de nucléotides, composés chacun d'un sucre à cinq carbones (*i.e.* désoxyribose pour l'ADN et ribose pour l'ARN), d'un groupement phosphate et d'une base organique azotée purine (*i.e.* adénine et guanine) ou pyrimidine (*i.e.* cytosine, thymine et uracile). Le rôle des nucléotides ne se limite pas à l'encodage et à l'expression de l'information génétique puisqu'ils sont également essentiels à l'activation de réactions chimiques (*i.e.* jouent le rôle cofacteurs). De surcroît, l'ATP, considéré comme la monnaie d'échange énergétique des cellules, ou le NAD^+ et le FAD^+ sont des molécules essentielles liées au transport de substances comme des électrons ou des molécules plus complexes.

En résumé, la complexité et la diversité des macromolécules biologiques, de leur composition à leurs fonctions variées, sont essentielles à la modélisation précise du métabolisme et de la croissance des organismes. Une compréhension approfondie de ces éléments permet de mieux appréhender les interactions biochimiques et les mécanismes cellulaires.

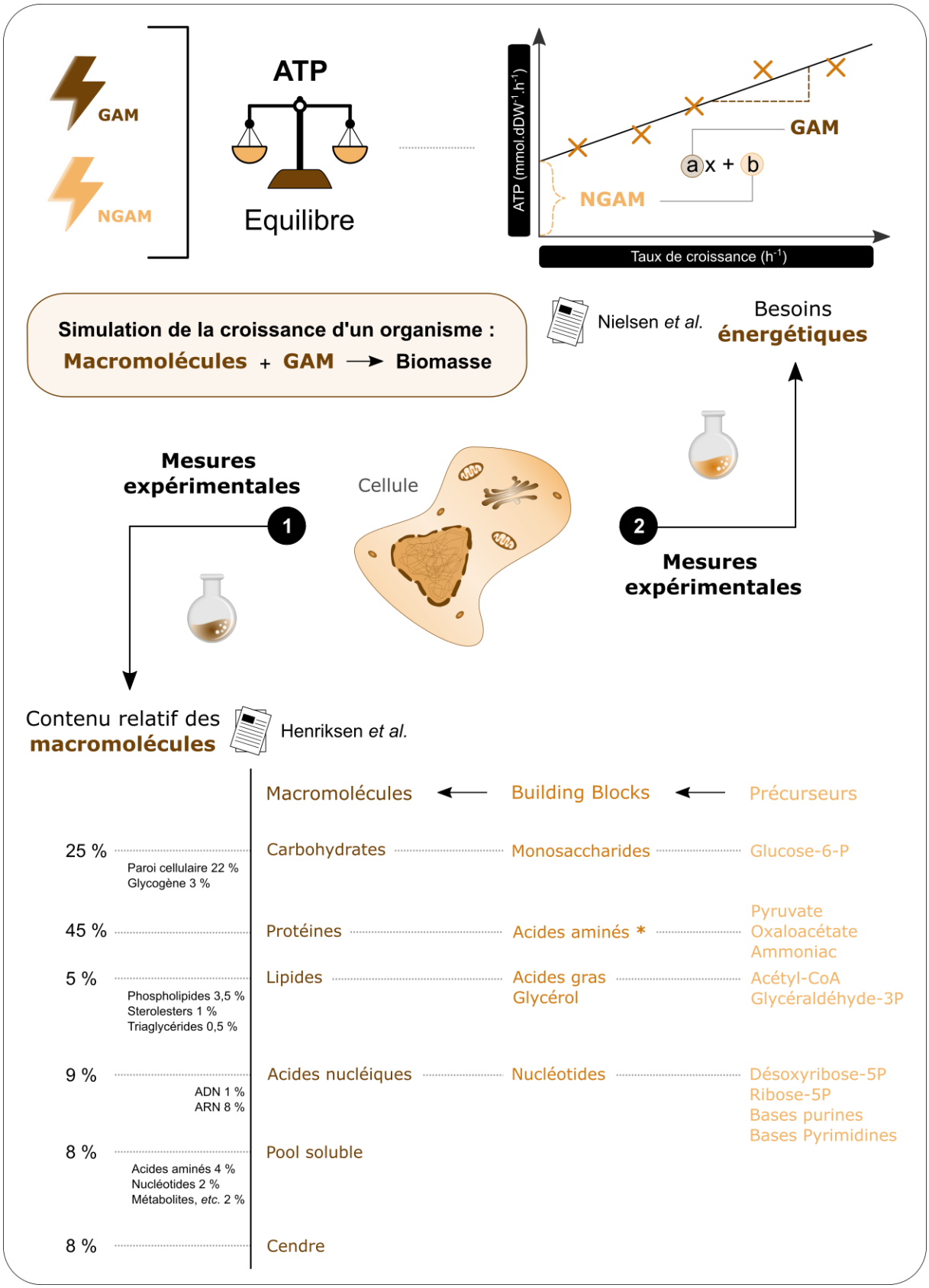


Figure 3-25 : Résumé synthétique des éléments constitutifs de la réaction de biomasse. (1) *Les données numériques concernant le contenu relatif des macromolécules proviennent de la publication Growth energetics and metabolic fluxes in continuous cultures of Penicillium chrysogenum (Henriksen et al. 1996).* ***(2)*** *Les coefficients des réactions GAM et NGAM ont été déterminés par régression linéaire lors d'expériences en chémostat limité par le glucose en présence d'acide phénoxyacétique. Les données expérimentales sont tirées du livre Physiological Engineering Aspects of Penicillium chrysogenum (Nielsen 1997), et leurs valeurs ont été estimées à 104 mmol·gDW⁻¹·h⁻¹ pour la réaction GAM et à 4.14 mmol·gDW⁻¹·h⁻¹ pour la réaction NGAM lors de la conception d'iAL1006.*

L'anabolisme représente la composante du métabolisme qui englobe, d'une part, la biosynthèse des éléments constitutifs (*i.e. building blocks*) à partir de précurseurs métaboliques et, d'autre part, la polymérisation de ces éléments en macromolécules. Il est important de noter qu'il existe une centaine d'éléments constitutifs, de cofacteurs et de groupements prosthétiques (*i.e.* composés organiques non-protéiques tels que l'hème, la thiamine, la vitamine B12, *etc.*) dérivés d'une dizaine de précurseurs métaboliques (*i.e.* intermédiaires de la glycolyse tels que le glucose-6-phosphate ou le pyruvate, de la voie des pentoses phosphates comme le ribose-5-phosphate, ou du cycle de Krebs tels que l'acétyl-CoA ou l'oxaloacétate). Le catabolisme, quant à lui, regroupe des réactions de dégradation permettant la formation de métabolites précurseurs, la libération d'énergie libre de Gibbs principalement sous forme d'ATP utilisée pour alimenter d'autres réactions cellulaires et la génération de pouvoir réducteur nécessaire aux réactions de biosynthèse. Enfin, les réactions d'absorption de substrats provenant du milieu environnant viennent compléter cet ensemble de réactions chimiques permettant de modéliser le métabolisme d'un organisme *(Nielsen 1997)*.

2.2.3.2. Formuler la réaction de biomasse : un enjeu sous-estimé ?

La réaction de biomasse constitue un inventaire des composés essentiels à la croissance de l'organisme. Une fois la nature des biomolécules déterminées, leur quantité, c'est-à-dire leur fraction molaire, portée par le coefficient stœchiométrique des éléments, doit être caractérisée. Pour ce faire, l'idéal est d'avoir recours à des données mesurées expérimentalement lors de phase de culture (cf. encart : *Obtenir des données expérimentales pour la conception de la réaction de biomasse : croissance continue en chémostat*, page 282). Les coefficients de ces éléments sont alors pondérés afin de refléter leur quantité dans un gramme de biomasse en poids sec. Notons que le taux de croissance s'exprime en h^{-1} puisque toutes les fractions des précurseurs de la biomasse sont converties en $mmol.gDW^{-1}$ *(Thiele et Palsson 2010a)*. Le détail du contenu molaire des constituants de la biomasse permet alors de calculer les rendements de la biomasse sur la base de la stœchiométrie.

L'aspect qualitatif de la composition de la biomasse, à savoir la présence ou l'absence de précurseurs dans la réaction de biomasse, est critique dans les simulations *in silico* de suppression génique (*i.e. knock-out*) et, par conséquent, pour la caractérisation de l'essentialité des réactions et des gènes. Déterminer spécifiquement les réactions de biomasse est donc un enjeu majeur, et des avancées significatives ont été réalisées en ce sens ces dernières années, à l'instar de BOFdat. BOFdat (*Biomass Obective Function from experimental data*) est un algorithme conçu pour générer des fonctions objectives de biomasse pour les **GSMNs** en se basant sur des données expérimentales *(Lachance et al. 2019)*. Avec une implémentation modulaire, le processus de définition de la réaction de biomasse est divisé en trois phases : **(1)** calcul des coefficients des principales macromolécules, **(2)** identification des coenzymes et des ions inorganiques puis estimation de leurs coefficients, et **(3)** extraction à partir de données expérimentales des précurseurs de la biomasse métabolique spécifiques à l'espèce. Ainsi, cet algorithme améliore la précision des **GSMNs** en ajustant qualitativement et quantitativement la composition de la biomasse pour qu'elle reflète fidèlement les conditions expérimentales observées. D'une part, il ajuste les coefficients stœchiométriques des éléments, et d'autre part, il identifie les éléments nécessaires à la croissance de l'organisme à partir d'analyses omiques et de mesures biochimiques.

Selon cette définition, et en toute logique, la réaction de biomasse est hautement spécifique à l'organisme étudié. Pourtant, sa formulation est souvent héritée d'organismes similaires en raison des défis expérimentaux associés à sa détermination correcte *(Beck et al. 2018)*. Cet aspect constitue d'ailleurs l'une des principales sources d'incertitudes dans la formulation de la composition de la biomasse *(Bernstein et al. 2021)*. En effet, des réactions de biomasse d'organismes plus ou moins similaires sont régulièrement utilisées pour caractériser l'organisme d'intérêt, sans tenir compte toutefois des variations susceptibles d'exister entre les différentes espèces. Pour illustrer ce point, nous citerons les travaux de Xavier *et al. (2017)*, qui ont démontré, par un clustering hiérarchique basé sur une comparaison des compositions de réactions de biomasse issues de **GSMNs** de 71 espèces procaryotes, que les regroupements effectués étaient relativement indépendants de leur parenté phylogénétique et qu'ils étaient plus largement liés à la nature des réactions de biomasse comparées.

Outre une différence de composition de biomasse entre organismes, nous pourrions également nous attendre à des variabilités de la composition de biomasse au sein du même organisme. En effet, en réponse à des modifications du taux de croissance, de la disponibilité des nutriments ou de paramètres physiques tels que la température *(Bernstein et al. 2021)*, la croissance d'un organisme est impactée. Par exemple, chez les champignons filamenteux, selon leur substrat de croissance, des formations filamenteuses ou des agrégats de pelotes sont observés *(Nielsen 1997)*. Il est alors plausible que les cellules réagissent par des réarrangements métaboliques, influençant la biosynthèse des précurseurs métaboliques nécessaires à la construction des constituants moléculaires d'une cellule. Il est donc essentiel de prendre en considération la dépendance de la composition de la biomasse par rapport aux conditions environnementales.

À ce titre, les travaux de Schulz *et al. (2021)* proposent deux approches simples et rapides à appliquer pour générer des compositions de biomasse pour un **GSMN** en fonction des variations de l'environnement nutritif : la *Biomass Tradeoff Weighting* (BTW) et la *Higher-dimensional-plane InterPolation* (HIP). Ces méthodes reposent sur les hypothèses que la composition de la biomasse dépend de l'environnement et que des environnements similaires produisent des compositions de biomasse similaires. Lorsqu'elles ont été testées avec le modèle *E. coli i*ML1515, elles ont révélé des profils phénotypiques variés en termes de taux de croissance, de sécrétion d'acétate, d'essentialité des gènes et de quotient respiratoire. Ainsi, appliquer ces méthodes à ***i*Prub22** permettrait d'approfondir les connaissances sur les relations entre l'environnement, le métabolisme et la composition de la biomasse de notre organisme.

Obtenir des données expérimentales pour la conception de la réaction de biomasse : croissance continue en chémostat

Un chémostat, ou bioréacteur, est un dispositif de culture continue où les paramètres environnementaux tels que la température, le pH, la concentration en nutriments et le taux de dilution sont soumis à des conditions strictement contrôlées. Le milieu de culture est continuellement alimenté en nutriments, tandis que le volume de culture est maintenu constant par l'élimination simultanée du même volume excédentaire (**Figure 3-26**). Cette configuration permet de maintenir une densité cellulaire stable, permettant ainsi d'étudier la croissance cellulaire à un taux de dilution spécifique. Ainsi, les chémostats sont particulièrement utiles pour étudier la cinétique de croissance des organismes, la régulation de l'expression génique en réponse à des conditions environnementales spécifiques et pour déterminer des paramètres métaboliques, tels que le rendement cellulaire ou le taux de croissance spécifique. La capacité à contrôler précisément les conditions de croissance permet d'obtenir des données précises sur le métabolisme et la physiologie des organismes, améliorant ainsi la compréhension des processus biologiques fondamentaux.

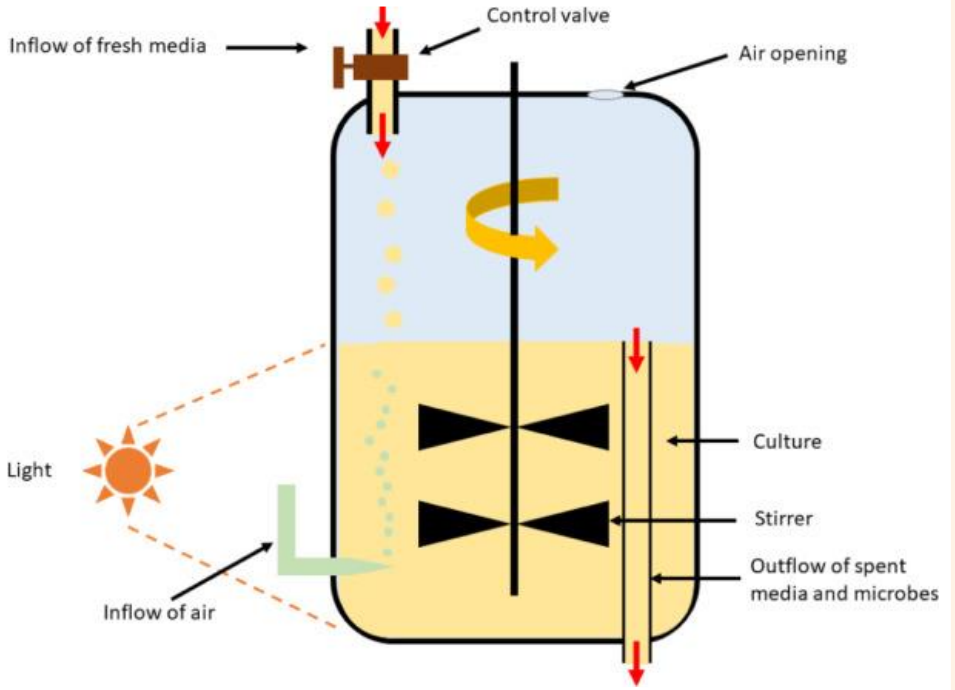


Figure 3-26 : Schéma d'un chémostat. *Le réacteur à cuve agitée montre un apport continu de milieu frais contrôlé pour équilibrer l'évacuation du milieu usé, permettant de maintenir un volume constant. L'agitation homogénéise la répartition des nutriments et des micro-organismes. Figure extraite de (Alzahrani et al. 2020)*

Pour les champignons filamenteux, les mesures intracellulaires sont généralement effectuées *in vitro* et se divisent en trois catégories principales : **(1)** la mesure des pools macromoléculaires (*i.e.* données cruciales pour les études énergétiques et la quantification des flux métaboliques), **(2)** la mesure des activités enzymatiques et **(3)** la mesure des métabolites. Diverses méthodes analytiques peuvent être utilisées pour quantifier la croissance des champignons filamenteux et évaluer les paramètres de leur environnement de culture, telles que les méthodes enzymatiques, les chromatographies en phase gazeuse (GC) ou liquide (HPLC), la spectrométrie de masse (MS), *etc.*

■ Biomasse

- Poids sec — Mesure de la biomasse après séchage et pesée. Indication directe de la quantité de biomasse produite.
- Densité optique — Mesure de la turbidité de la culture à une longueur d'onde spécifique par spectrophotomètrie. Reflète indirectement la concentration de biomasse.

■ Composition chimique de la biomasse (évaluation de la proportion des éléments)

- Protéines — Dosage ou autres techniques spectrophotométriques.
- Polysaccharides — Mesure par des techniques colorimétriques ou chromatographiques.
- Lipides — Quantification par chromatographie et extraction par solvants suivie de pesée ou analyse par GC.
- ADN — Quantification de l'ADN total par des méthodes fluorimétriques ou spectrophotométriques.
- ARN — Quantification par des méthodes similaires à celles utilisées pour l'ADN.

■ Substrat (mesures de concentration)

- Sucres — Mesure par des méthodes enzymatiques ou chromatographiques.
- Acides aminés — Analyse par chromatographie avec détection par fluorescence ou spectrophotométrie.
- Nutriments — Mesure par colorimétrie ou chromatographie pour l'azote, le phosphore, *etc.*

■ Produits métaboliques

- Acides organiques — Mesure par HPLC (*e.g.* acide acétique).
- Alcools — Mesure par GC (*e.g.* éthanol).
- Gaz produits — Mesure des gaz produits (*e.g.* O_2,CO_2) par analyseurs de gaz en ligne ou GC.
- Métabolites spécialisés — Mesure de composés spécifiques produits par les champignons, tels que des antibiotiques ou des mycotoxines, par HPLC, LC-MS, ou GC-MS.

■ Paramètres de croissance

- Taux de croissance spécifique — Calcul basé sur la variation de la biomasse dans le temps. Reflète la vitesse de croissance des champignons filamenteux dans le chémostat.
- Rendement en biomasse — Rapport entre la biomasse produite et le substrat consommé. Calculé en divisant la quantité de biomasse sèche par la quantité de substrat utilisé.
- Productivité volumétrique — Quantité de produit (*e.g.* biomasse, métabolites) formée par unité de volume par unité de temps.

■ Paramètres environnementaux

- pH — Mesure et contrôle avec une sonde pH pour maintenir des conditions optimales. Le pH est souvent ajusté automatiquement avec des solutions acides ou basiques.
- Température — Mesure et contrôle avec un thermomètre ou une sonde de température pour assurer des conditions de culture stables et optimales.
- Oxygène dissous — Mesure avec une sonde d'oxygène pour surveiller la concentration en oxygène dissous dans le milieu de culture. Variable essentielle pour la respiration cellulaire et la croissance des champignons filamenteux.

En combinant ces mesures, il est possible d'obtenir une compréhension détaillée de la cinétique de croissance des champignons filamenteux dans un chémostat, ainsi que des conditions environnementales optimales pour leur culture. Les expériences de croissance en chémostat représentent également une méthode optimale pour déterminer les besoins énergétiques de l'organisme étudié (*i.e.* détermination des coefficients des réactions GAM et NGAM).

2.2.3.3. Une réaction de biomasse bien établie pour *Penicillium rubens*

La problématique relative à la spécificité des données pour la conception de la réaction de biomasse de notre organisme est écartée en raison de l'abondance d'études effectuées sur *P. chrysogenum/rubens* depuis les années 1950. En effet, l'équation de la réaction de biomasse du modèle *i*AL1006 est basée sur les travaux de Nielsen *et al.* *(1997)*, qui répertorient les stœchiométries globales nécessaires pour la polymérisation de 1 g de protéines, ainsi que pour la synthèse de 1 g d'ARN, 1 g d'ADN, 1 g de phospholipides, 1 g d'esters, 1 g de triacylglycérols et 1 g de glucides. La réaction de biomasse du réseau de 2013 modélise ainsi la biosynthèse des macromolécules à travers dix sous-systèmes, présentés dans le **Tableau 3-9** : production des acides aminés, des composants de la paroi cellulaire, des cofacteurs, de l'ADN, de l'ARN, des acides gras, des phospholipides, des glycérides, des stérols esters et d'autres lipides.

Dans la version de 2018, les sous-systèmes correspondant à la synthèse des autres lipides, des stérols esters et des glycérides sont absents de la réaction de biomasse. De plus, le sous-système des phospholipides ne couvre pas les composés de la famille des céramides mannose-(P-inositol). Il a été tenté, en vain, de récupérer ces informations pour la conception de la réaction de biomasse d'***i*Prub22**. L'absence de correspondance claire entre les identifiants, l'utilisation de termes génériques et le contenu des bases de données rendent cette tâche complexe, voire impossible dans le contexte actuel.

En conséquence, la réaction de biomasse présentée dans ***i*Prub22** se rapproche, par la nature des sous-systèmes modélisés, de celle présentée dans Prubens, bien que nous ayons conservé les coefficients d'*i*AL1006 et le phénomène d'hydrolyse de l'ATP (*i.e.* énergie requise pour la croissance cellulaire, équivalente à la réaction GAM) pour la réaction de biomasse et pour les sous-systèmes de biosynthèse des protéines (*i.e.* `r1456`), de l'ADN (*i.e.* `r1458`) et de l'ARN (*i.e.* `r1457`).

Tableau 3-9 : La réaction de biomasse de *Penicillium rubens* Wisconsin 54-1255, de *i*AL1006 à *i*Prub22. Ce tableau présente les coefficients stœchiométriques des éléments constitutifs de la réaction de biomasse et des dix sous-systèmes associés, tels qu'encodés dans le premier **GSMN** de *P. rubens* Wisconsin 54-1255, *i*AL1006 *(Agren et al. 2013)*. Les composants surlignés en orange et précédés d'un triangle (▲) sont absents de la réaction de Prubens *(Prigent et al. 2018)* et d'***i*Prub22**, tandis que ceux précédés d'un carré (■) ont été modifiés entre les versions. Pour une meilleure lisibilité, les substrats sont présentés par leur abréviation ou leur nom commun. Les identifiants ayant servi dans ***i*Prub22** sont extraits de Prubens, et nous avons conservé pour ***i*Prub22** les coefficients d'*i*AL1006. Les coefficients des composés précédés d'un cercle (●) sont différents dans Prubens, et ceux précédés d'une étoile (★) sont absents (*i.e.* hydrolyse de l'ATP).

r1463 : Biomasse			
	Reactant	AAPOOL	0.04
★	Reactant	ATP	104
	Reactant	CELLWALL	0.25
	Reactant	COF	0.0001
	Reactant	DNA	0.01
▲	Reactant	GLYFFA	0.005
	Reactant	H2O	104
▲	Reactant	OL	0.00001
	Reactant	PROTEIN	0.45
	Reactant	RNA	0.08
▲	Reactant	STERE	0.01
	Reactant	PLIPIDS	0.035
★	Product	ADP	104
★	Product	Pi	104
	Product	**BIOMASS**	**1**
r1455 : formation de la paroi cellulaire et stockage			
	Reactant	Trehalose	0.1652
	Reactant	GDP-Mannose	0.3304
	Reactant	UDP-Galactose	0.8261
	Reactant	UDP-Glucose	2.8088
	Reactant	UDP-GlcNAc	1.5421
	Product	GDP	0.3304
	Product	UDP	5.177
	Product	**CELLWALL**	**1**
▲ r1462 : synthèse des esters de stérols			
	Reactant	Ester d'ergostérol	0.4992
	Reactant	Ergostérol	1.694
	Product	**STERE**	**1**
r1459 : pool d'acides aminés			
	Reactant	GLY	0.25719
	Reactant	AMA	0.04822
	Reactant	ALA	1.0449
	Reactant	ARG	0.20897
	Reactant	ASN	0.27327
	Reactant	ASP	0.46616
	Reactant	CYS	0.04019
	Reactant	GLU	2.0415
	Reactant	GLN	1.3342
	Reactant	HIS	0.22504
	Reactant	ILE	0.10448
	Reactant	LEU	0.1286
	Reactant	LYS	0.17682
	Reactant	MET	0.04019
	Reactant	PHE	0.04019
	Reactant	PRO	0.41392
	Reactant	SER	0.37775
	Reactant	THR	0.24916
	Reactant	TRP	0.01607
	Reactant	TYR	0.04822
	Reactant	VAL	0.1286
	Product	**AAPOOL**	**1**
r1457 : synthèse de l'ARN			
	Reactant	AMP	0.79
	Reactant	ATP	7.44
	Reactant	CMP	0.61
	Reactant	GMP	0.89
	Reactant	H2O	7.44
	Reactant	UMP	0.81
★	Product	ADP	7.44
★	Product	Pi	7.44
	Product	**RNA**	**1**
r1456 : formation des protéines			
★	Reactant	ATP	34.99
●	Reactant	GLY	0.62095
●	Reactant	ALA	0.58394
●	Reactant	ARG	0.55516
●	Reactant	ASN	0.14085
●	Reactant	ASP	0.42254
●	Reactant	CYS	0.14804
●	Reactant	GLU	0.72479
●	Reactant	GLN	0.2416
●	Reactant	HIS	0.20561
●	Reactant	ILE	0.34543
●	Reactant	LEU	0.69498
●	Reactant	LYS	0.62095
●	Reactant	MET	0.44413
●	Reactant	PHE	0.27552
●	Reactant	PRO	0.42356
●	Reactant	SER	0.42356
●	Reactant	THR	0.44001
●	Reactant	TRP	0.05757
●	Reactant	TYR	0.19328
●	Reactant	VAL	0.57572
★	Product	ADP	34.99
★	Product	PI	34.99
	Product	**PROTEIN**	**1**
r1458 : synthèse de l'ADN			
★	Reactant	ATP	11.22
●	Reactant	DAMP	0.79
●	Reactant	DCMP	0.86
●	Reactant	DGMP	0.86
●	Reactant	DTMP	0.79
●	Reactant	H2O	11.22
★	Product	ADP	11.22
★	Product	Pi	11.22
	Product	**DNA**	**1**

Tableau 3-9 (suite et fin)

r1465 : pool des cofacteurs		
Reactant	Biotine	1
Reactant	CoA	1
Reactant	NADH	1
Reactant	NADPH	1
Reactant	Siroheme	1
Reactant	Tetrahydrofolate glutamate	1
Reactant	Thiamine diphosphate	1
Reactant	FAD	1
Reactant	Folates	1
■ Reactant	Heme a	1
■ Reactant	Ubiquinone	1
Product	**COF**	**1**

NB. Les substrats Heme A et Ubiquinone ont été remplacés par les éléments Ferroheme A et Ubiquinol

r1460 : formation des phospholipides		
Reactant	Cardiolipine	0.04314
Reactant	Phosphatidylcholines	0.3783
Reactant	Phosphatidyl-Éthanolamines	0.85994
Reactant	Phosphatidylsérine	0.00887
▲ Reactant	Trigalactosyldimannosylinositol-Pcéramide	0.00012
Product	**PLIPIDS**	**1**

▲ r1461 : synthèse des glycérides		
Reactant	Diglycerides	0.07179
Reactant	Acides gras libres	0.37947
Reactant	Monoglycerides	0.63411
Reactant	Triglycerides	0.71743
Product	**GLYFFA**	**1**

▲ r1464 : formation d'autres lipides		
Reactant	Cérébrine 1	1
Reactant	Cérébrine 2	1
Reactant	Digalactosyl-diacylglycérol	1
Reactant	Galactosylcéramides	1
Reactant	Glucocérébroside 1	1
Reactant	Glucocérébroside 2	1
Reactant	Hexadécadiénoate	1
Reactant	Octadécatriénoate	1
Product	**OL**	**1**

2.2.3.4. Vérification de la robustesse de la réaction de biomasse d'*i*Prub22 par analyse de sensibilité

Lorsque la réaction de biomasse employée dans un modèle dérive d'une autre reconstruction, du même organisme ou d'une espèce similaire, une méthode de vérification de la qualité de cette réaction peut être effectuée par des analyses de sensibilité. L'objectif principal de telles analyses est de comprendre la robustesse du modèle et d'identifier les paramètres critiques pour la production de biomasse. Ces analyses consistent à examiner comment un état optimal de sortie d'un modèle mathématique réagit, qualitativement ou quantitativement, à diverses variations dans les entrées du modèle. Ainsi, une analyse de sensibilité de la fonction objective de la biomasse aux variations des coefficients des précurseurs de la biomasse *(Nanda et al. 2020)* permet de vérifier si la croissance du modèle *in silico* est sensible à tous les précurseurs de la biomasse, tandis qu'une analyse de sensibilité aux variations des bornes des réactions d'échange *(Lachance et al. 2021)* permet, par exemple, de vérifier les réponses du modèle aux variations environnementales et, le cas échéant, de déterminer les plages optimales pour la conception de différents milieux de culture.

Nous venons d'expliquer que la réaction de biomasse d'***i*Prub22** provient initialement d'*i*AL1006 et que cette dernière a été établie à partir de données expérimentales. Nous avons également observé des différences sur les coefficients stœchiométriques des réactants du sous-système `PROTEIN` dans la réaction de biomasse de Prubens. Nous savons que la contribution individuelle de chaque précurseur, portée par la

distribution fractionnelle de chaque composé, a un impact limité dans l'analyse de l'essentialité des gènes et *de facto*, des réactions. Néanmoins, elle est cruciale pour prédire avec exactitude le taux de croissance optimal du système. C'est pourquoi, nous présentons en **Figure 3-27** une analyse de sensibilité sur les coefficients des précurseurs de la biomasse et de ceux relatifs à ses sous-systèmes afin de visualiser l'impact potentiel des variations de ces coefficients sur la production de biomasse.

Pour cela, nous avons augmenté ou diminué de 50 % la valeur de chaque coefficient de réactant dans l'équation de la biomasse ou de l'un de ses sous-systèmes, puis exécuté une FBA pour estimer le taux de croissance à chaque changement de coefficient et déterminer comment cette modification affecte ou non le flux optimal de biomasse. Les éléments dont les variations de coefficient entraînent des changements significatifs dans la biomasse sont considérés comme critiques et doivent faire l'objet d'une attention particulière. À l'inverse, une faible variation de la biomasse en réponse aux changements des coefficients suggère que le modèle est robuste vis-à-vis de ces variations.

De manière simpliste, si une augmentation de production de biomasse est observée avec l'augmentation d'un coefficient de précurseur, cela signifie que ce dernier est limitant au sein du système et que des quantités supplémentaires améliorent la croissance. Réciproquement, si la diminution du coefficient réduit la production de biomasse, cela confirme que le précurseur est essentiel et que sa disponibilité limite la croissance. Nous pourrions également supposer que, dans le cadre de relations linéaires, une augmentation du coefficient s'effectue au détriment de la croissance en raison d'une demande accrue de ressources. Or, il est établi que le métabolisme n'est pas régi exclusivement par des relations linéaires, et il est alors nécessaire de prendre en considération les potentielles interactions et réallocations des distributions des flux (*i.e.* changements d'équilibre) du système.

Concernant la réaction de biomasse, *stricto sensu*, nous constatons une variation maximale de six centièmes (**Figure 3-27.A**), indiquant une certaine robustesse du modèle puisque les plus grandes variations sont portées par, respectivement, les réactants `PROTEIN` et `CELLWALL`. De surcroît, des variations négligeables, de l'ordre de 10^{-3} à 10^{-6} h^{-1} sont observées pour les précurseurs des réactions artificielles à l'origine de la biosynthèse des systèmes : `COF` (**Figure 3-27.B**), `DNA` (**Figure 3-27.C**), `RNA` (**Figure 3-27.D**), `PLIPIDS` (**Figure 3-27.E**), `AAPOOL` (**Figure 3-27.G**) et `PROTEIN` (**Figure 3-27.H**). En revanche, il existe un point de vigilance concernant trois précurseurs du sous-système `CELLWALL` (**Figure 3-27.F**) : l'UDP (uridine diphosphate), l'UDP-α-glucose et l'UDP-*N*-acetyl-D-glucosamine.

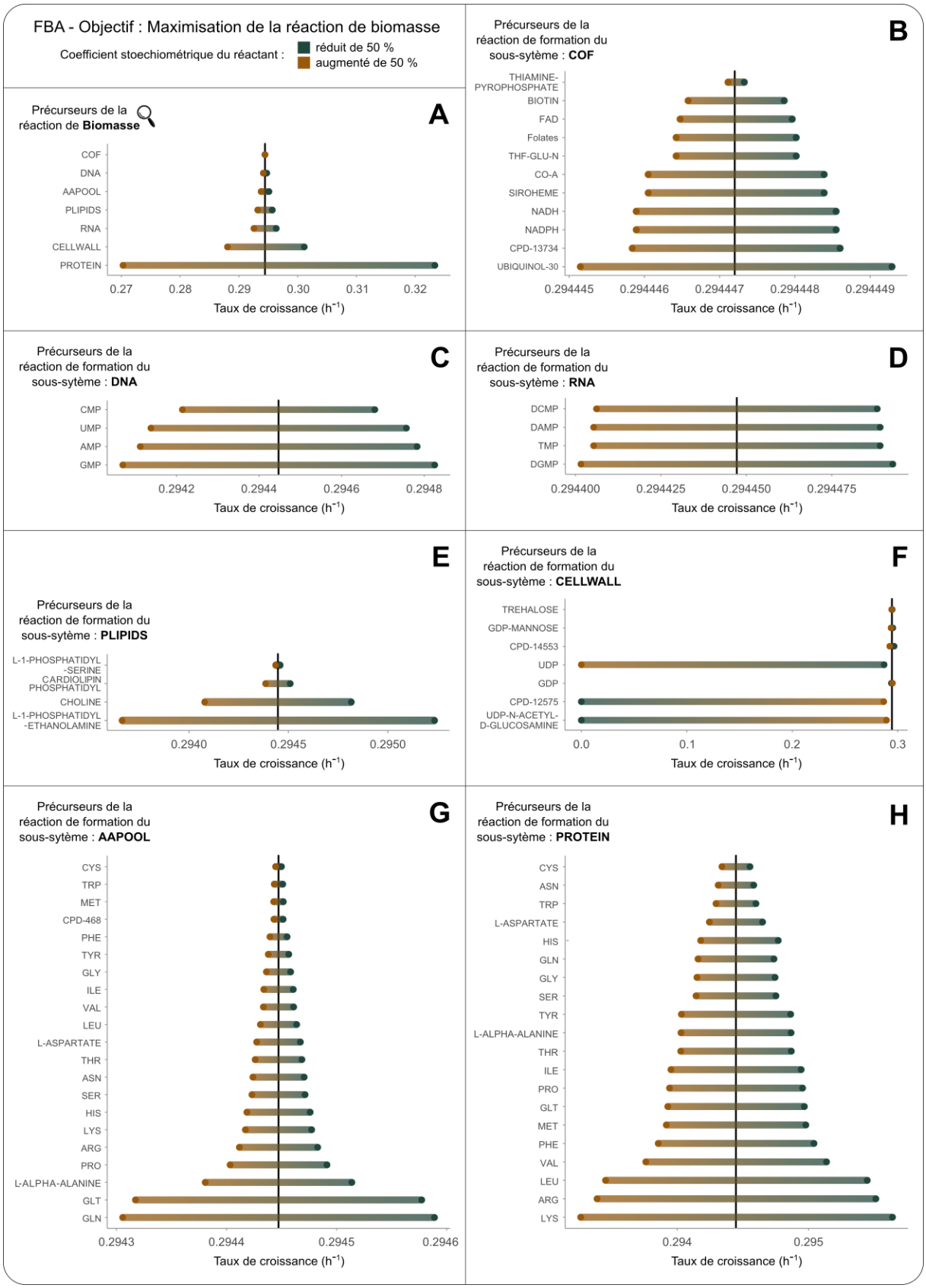


Figure 3-27 : Analyse de sensibilité de la fonction objective de la biomasse aux variations des coefficients des précurseurs de la réaction de biomasse et de ses sous-systèmes. *Une FBA maximisant la réaction de biomasse à 100 % est réalisée pour chaque modification du coefficient de l'élément considéré. Les contraintes du modèle sont celles définies par défaut, rappelées en* ***Tableau 3-10****. Les points marron (●) et vert (●) représentent respectivement le résultat de la FBA avec une augmentation et une réduction de 50 % de la valeur du coefficient. La ligne verticale noire représente la valeur du flux de biomasse dans les conditions par défaut, soit 0.2944472 h-1. Les divers précurseurs testés sont caractérisés par leur identifiant MetaCyc.*

L'UDP agit comme transporteur d'énergie et de substrats dans les cellules et il est essentiellement impliqué dans le métabolisme des glucides et des lipides. De plus, il sert de précurseur pour la synthèse de divers nucléotides et acides nucléiques. Nous observons que lors de la diminution de son coefficient, la valeur optimale de croissance est inférieure à celle obtenue en condition standard, ce composé semble donc être un facteur limitant pour la croissance. De surcroît, nous constatons que lors de l'augmentation de son coefficient, le système n'est plus en mesure de produire de la biomasse. L'effet négatif de cet excès peut entraîner une perturbation des voies métaboliques suite à l'accumulation de produits intermédiaires toxiques ou provoquer un stress métabolique sévère. Par exemple, le système est forcé de rediriger des ressources métaboliques pour gérer l'excès, épuisant les ressources disponibles pour d'autres processus vitaux. En conséquence, l'UDP en quantité excessive, a des effets profondément perturbateurs sur le métabolisme cellulaire puisque le système exprime une grande sensibilité aux variations de son coefficient. En outre, la valeur définie dans le modèle, issue de l'expérience et utilisé pour *i*AL1006 semble être une valeur spécifique permettant une croissance optimale du modèle.

Concernant l'UDP-α-glucose (*i.e.* `CPD-12575`) et l'UDP-N-acetyl-D-glucosamine (*i.e.* `UDP-N-ACETYL-D-GLUCOSAMINE`), leurs résultats expriment un profil similaire. Une augmentation de leur coefficient induit une diminution de la production de biomasse, tandis qu'une diminution de leur coefficient empêche la production de biomasse. Le premier joue un rôle clé dans la glycogénèse alors que le second participe à la biosynthèse des glycoprotéines et des glycolipides. La diminution de la biomasse avec l'augmentation des coefficients peut signifier un coût métabolique élevé avec une plus grande demande pour ces précurseurs. La cellule doit allouer plus de ressources et d'énergie pour synthétiser ou importer ces métabolites, ce qui peut réduire la disponibilité des ressources pour d'autres processus essentiels dont la production de biomasse. En revanche, la diminution du coefficient occasionne une production de biomasse nulle et signifie que ces précurseurs sont absolument essentiels pour la croissance. Une quantité insuffisante de ces métabolites ne permet pas de maintenir les processus cellulaires vitaux et ces éléments sont des composants critiques de la biomasse. Leur réduction en dessous d'un certain seuil compromet la capacité de la cellule à produire de la biomasse entraînant dès lors une défaillance complète de la croissance cellulaire. De manière similaire à l'UDP, il est intéressant de noter que la valeur issue de l'expérience semble correspondre à un seuil spécifique optimal pour la croissance du modèle.

2.2.3.5. Simulation de croissance : gage de fonctionnalité pour un modèle métabolique ?

La fonctionnalité d'un modèle, et en conséquence sa qualité, est généralement évaluée par l'utilisation d'une réaction de biomasse en tant que fonction objective afin de déterminer la distribution des flux dans un état optimal du réseau lors d'analyses FBA ou FVA. Nous rappelons à cet effet que la distribution des flux à travers les différentes voies cellulaires est calculée à partir des mesures des flux entrant et sortant des systèmes évalués.

Cependant, résumer la fonctionnalité d'un modèle à sa simple capacité de production de biomasse est réducteur ; encore faut-il que cette production soit cohérente. Dans les premières versions d'***i*Prub22**, nous avons pu observer la production de divers sous-systèmes de la réaction de biomasse alors que l'ensemble des réactions d'*uptake* était clos (*i.e.* modèle fermé). C'est pourquoi nous avons mentionné précédemment que, pour les besoins de la modélisation, entre autres, un certain nombre de réactions non équilibrées en charge et/ou en masse ont été bloquées. Cette solution, bien qu'efficace, n'est pourtant pas optimale. Certes, procéder ainsi nous a permis de briser un ensemble de cycles énergétiques futiles qui entraînaient un dysfonctionnement des modèles, sans toutefois identifier la, ou plus vraisemblablement les ensembles de réactions responsables de ces cycles. De surcroît, ce critère drastique, 1 402 réactions sont concernées, ne doit pas être employé de manière systématique puisque certaines réactions sont essentielles à la production de biomasse et à la production des métabolites spécialisés étudiés ultérieurement. Pour illustrer ce point, nous indiquons ci-dessous la répartition des réactions considérées comme non équilibrées par la fonction `checkMassChargeBalance()` de COBRA Toolbox :

- 228 sont des réactions d'échange.
- 170 ont été fermées lors de curations précédentes.
- 10 sont liées à la réaction de biomasse et assimilées.
- 25 sont nécessaires au maintien de la production de biomasse.
- 23 sont nécessaires à la production de métabolites spécialisés.
- 946 sont bloquées et annotées avec l'étiquette `imbalanced_reaction_bound` pour assurer leur traçabilité.

Avant d'aboutir à la solution proposée et détaillée dans le Livescript d'***i*Prub22**, diverses approches ont été testées. Certaines d'entre-elles nous ont permis d'apporter des améliorations graduelles au système, mais dans la majorité des cas, les résultats des analyses effectuées aboutissaient au simple constat de l'existence d'incohérences dans le modèle, sans mettre en évidence les causes, ni apporter de réelles solutions. Notons également qu'il n'existe pas de « recette magique » permettant de résoudre les anomalies d'un modèle. Ici encore, la curation manuelle et les choix de l'opérateur sont prépondérants (*e.g.* ordre des modifications effectuées sur le modèle, choix des critères utilisés ou des contraintes employées, *etc.*). De plus, la recherche

de solutions (*i.e.* rappelons à ce titre la pluralité de ces dernières) peut s'effectuer sous de nombreux angles classés en deux familles historiques : les analyses topologiques et les analyses de modélisation par contraintes, l'enjeu principal étant de trouver une ou des solutions à la justification cohérente. Nous établissons ci-dessous une liste des pistes, dont certaines ont été évoquées précédemment, que nous avons suivies pour tenter de résoudre les incohérences de notre modèle :

- identification et comparaisons des réactions actives et bloquées sous diverses conditions (*i.e.* modèle ouvert, fermé, minimal, *etc.*),
- recherche des réactions redondantes et dupliquées,
- correction et mise à jour de données concernant la réversibilité des réactions,
- explorations topologiques manuelles couplées à des simulations de *knock-out in silico* (*i.e.* analyses conduite essentiellement avec Fluxer *(Hari et Lobo 2020)*),
- comparaisons de topologie avec les réseaux *i*AL1006 *(Agren et al. 2013)* et Prubens *(Prigent et al. 2018)* et vérifications du poids des réactions intégrées (*e.g.* recherche d'un problème qui serait potentiellement lié à une réaction issue du *gap-filling*),
- réconciliation des identifiants pour des entités « similaires » (cf. encart *Un « même » composé mais des identifiants différents, lequel choisir ? Illustration avec l'entité glucose*, page 259),
- identification des impasses métaboliques,
- recherche de cycles qui génèrent de l'énergie et création de réaction de dissipation d'énergie,
- recherche de réactions essentielles et réduction de réseau,
- recherche de *leak* (*i.e.* production sans consommation) ou de *siphon* (*i.e.* consommation sans production),
- recherche de réactions aux flux couplés.

En dépit de ces divers tests, les causes des problèmes rencontrés lors des simulations initiales n'ont pas été précisément identifiées. Une analyse détaillée des 946 réactions bloquées sera essentielle pour identifier les réactions problématiques, proposer des solutions adaptées et améliorer la précision des prédictions métaboliques, renforçant ainsi la compréhension, la robustesse et la fiabilité du modèle ***i*Prub22**.

2.2.3.6. Besoins nutritionnels minimaux pour la croissance de *Penicillium rubens*, comparaison de modèles

Malgré les difficultés rencontrées lors des phases de réconciliation de données, l'un des avantages majeurs d'avoir accès à des reconstructions antérieures est de pouvoir les utiliser comme canevas de comparaison. Dans un premier temps, nous utilisons donc les simulations des réseaux pré-existants comme repères. Outre les similitudes de comportements que nous pourrions observer et qui viendraient renforcer le poids de la capacité prédictive d'***i*Prub22**, ce sont surtout les différences que nous allons préférentiellement rechercher.

Qualitativement et à l'exception des imports de pimelate (*i.e.* indispensable pour *i*AL1006 et Prubens) et de riboflavine (*i.e.* indispensable pour ***i*Prub22**), les besoins de nutriments minimaux nécessaires à la fonctionnalité des trois **GSMNs** sont identiques. Ces éléments, présentés en **Tableau 3-10** comprennent une source de carbone avec le glucose, une source d'azote avec l'ammoniac, de l'oxygène, du fer ferreux, du soufre, du phosphate inorganique, de la thiamine et, pour ***i*Prub22**, de la riboflavine. Ces deux derniers éléments artefactuels permettent de débloquer le sous-système `COF` en initiant la synthèse des cofacteurs thiamine diphosphate et FAD.

Tableau 3-10 : Conditions minimales assurant la fonctionnalité du modèle *i*Prub22. Les bornes ont été fixées arbitrairement, en respectant toutefois une proportion supérieure de carbone par rapport à l'azote. Notons que seuls les imports de glucose et d'ammoniac sont mentionnés dans ce tableau, mais nous avons vu précédemment que de nombreux éléments pouvaient servir de sources de carbone et/ou d'azote.

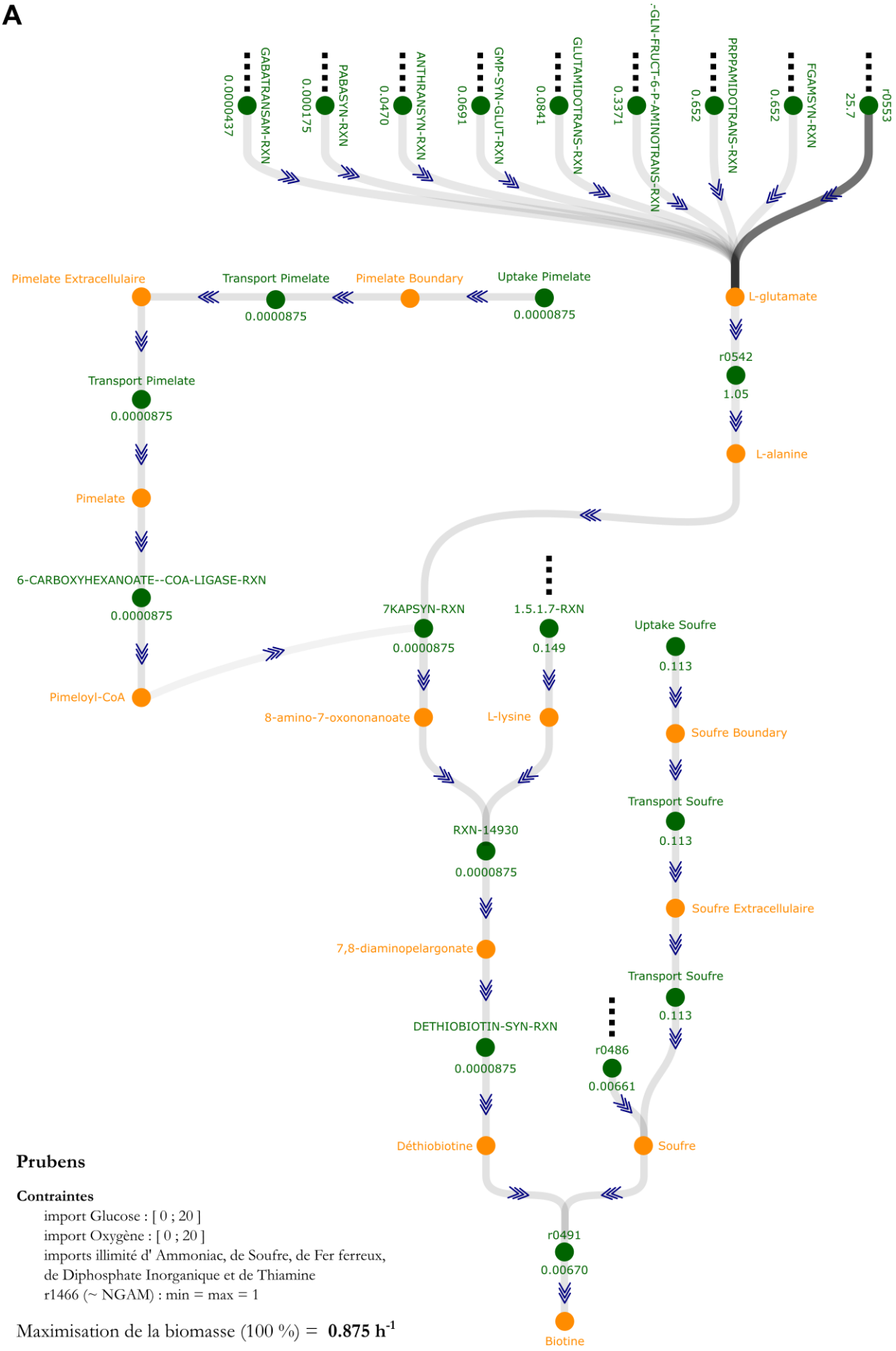
Source de Carbone		
● *Uptake*_015	ub=15 mmol.gDW^{-1}.h^{-1}	Glucose
Source d'Azote		
● *Uptake*_130	ub=5 mmol.gDW^{-1}.h^{-1}	Ammoniac
Éléments Non Substituables		
(i.e. les éléments suivants ne peuvent être remplacés par aucun import)		
● *Uptake*_062	ub=10 mmol.gDW^{-1}.h^{-1}	Fer ferreux
● *Uptake*_169	ub=10 mmol.gDW^{-1}.h^{-1}	Soufre
● *Uptake*_146	ub=10 mmol.gDW^{-1}.h^{-1}	Phosphate
● *Uptake*_171	ub=10 mmol.gDW^{-1}.h^{-1}	Thiamine
Éléments Substituables		
(i.e. les éléments suivants peuvent être remplacé par un autre import, écrit en gris)		
● *Uptake*_157	ub=10 mmol.gDW^{-1}.h^{-1}	Riboflavine
● *Uptake*_065	ub=10 mmol.gDW^{-1}.h^{-1}	FMN
● *Uptake*_136	ub=1000 mmol.gDW^{-1}.h^{-1}	Oxygène
● *Uptake*_077	ub=1000 mmol.gDW^{-1}.h^{-1}	Peroxyde d'hydrogène

Le premier élément de divergence entre ***i*Prub22** et les versions antérieures *i*AL1006 et Prubens, est le besoin d'un import de riboflavine. En explorant la topologie de Prubens, couplée à la distribution des flux résultant d'une FBA avec optimisation de la croissance, nous avons identifié une réaction essentielle qui est absente d'***i*Prub22** : la réaction `RIBOPHOSPHAT-RXN`. Cette réaction, bien que dépourvue de séquence et résultant probablement des propres étapes de *gap-filling* de Prubens, permet de consommer et de débloquer une *dead-end* uniquement produite, éliminant ainsi le besoin artificiel de riboflavine pour la fonctionnalité d'***i*Prub22**. Ainsi, l'ajout de cette réaction à ***i*Prub22** devra probablement être considéré lors de futures révisions de la reconstruction.

La seconde différence entre les trois modèles concerne l'import de pimelate, un élément essentiel pour la croissance dans les réseaux *i*AL1006 et Prubens, mais pas dans celui d'***i*Prub22**. Notons d'abord que le pimelate est un intermédiaire clé dans la voie de biosynthèse de la biotine (*i.e.* vitamine B8), un cofacteur essentiel dans le métabolisme des glucides et des acides gras chez les champignons filamenteux. La disponibilité en biotine influence directement leur croissance et leur morphogenèse en régulant les activités enzymatiques nécessaires à la construction et à la maintenance des structures cellulaires. La voie du pimelate n'étant pas, à ce jour, caractérisée avec précision, son import semble donc indispensable.

En examinant la distribution des flux menant à la biosynthèse de la biotine, exposée en **Figure 3-28** pour les modèles Prubens et ***i*Prub22**, nous constatons l'existence d'un chemin alternatif permettant la production de 8-amino-7-oxononanoate, un composé intermédiaire. Il semble donc que le pimelate puisse être substitué dans ***i*Prub22** par le pimeloyl-ACP, dont l'un des précurseurs initial est le malonyl-CoA. Les onze réactions impliquées dans cette voie métabolique, regroupées sous l'identifiant `PWY-6519`, sont toutes soutenues par au moins une séquence génomique, des séquences qui ont été détectées majoritairement par annotation et orthologie.

A



Prubens

Contraintes

import Glucose : [0 ; 20]
import Oxygène : [0 ; 20]
imports illimité d' Ammoniac, de Soufre, de Fer ferreux, de Diphosphate Inorganique et de Thiamine
r1466 (~ NGAM) : min = max = 1

Maximisation de la biomasse (100 %) = **0.875 h^{-1}**

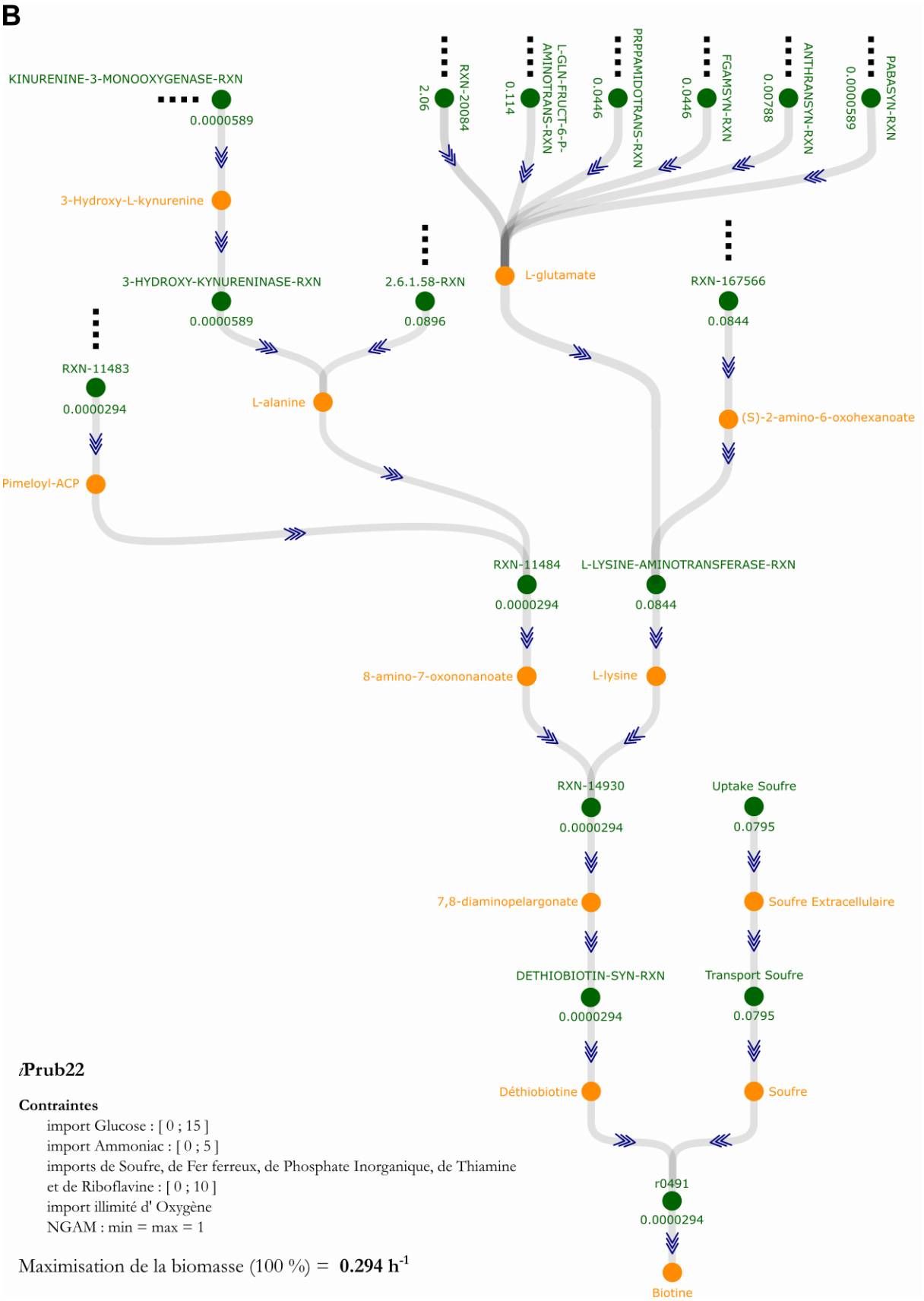


Figure 3-28 : Comparaison de la distribution des flux pour la biosynthèse de biotine. *Les voies métaboliques sont extraites de Fluxer (Hari et Lobo 2020) après le chargement par défaut des modèles Prubens* ***(A)*** *et* ***iPrub22 (B)****. Les métabolites, désignés par leur nom commun, sont représentés par des nœuds orange (●), tandis que les réactions, caractérisées par leur identifiant, sont matérialisées en vert (●), incluant la valeur de leur flux. Les contraintes ainsi que la valeur maximale de biomasse (i.e. fonction objective) obtenue par FBA sont rappelées pour chaque modèle.*

Outre ces informations, la **Figure 3-28** nous permet également d'établir trois constats différents relatifs à la biosynthèse des acides aminés L-lysine, L-alanine et L-glutamate. De manière générale, étudier les flux de distribution impliqués dans la biosynthèse des acides aminés constitue un indicateur de qualité de la modélisation du métabolisme constitutif et révèle ainsi la validité du modèle. Nous présentons à cet effet, en **Figure 3-29** un récapitulatif synthétique de ces diverses voies qu'il sera nécessaire de contrôler à l'avenir. Soulignons également que les détails que nous présentons ci-dessous, sont les conséquences d'une seule solution de simulation, obtenue par FBA sur deux réseaux métaboliques différents présentant des contraintes similaires. Des analyses plus approfondies seraient indispensables pour toute généralisation.

■ **Distribution des flux pour la biosynthèse de la L-lysine**

La biosynthèse de la L-lysine s'effectue en huit réactions à partir de l'acide *a*-cétoglutarique. Cette voie, représentée sous MetaCyc par le pathway `LYSINE-AMINOAD-PWY`, se conclut par la réaction `1.5.1.7-RXN` qui transforme la saccharopine en L-lysine. La distribution des flux de Prubens indique que la production de L-lysine emprunte effectivement ce seul chemin. En revanche, pour ***i*Prub22**, bien que ce pathway soit présent dans la reconstruction, la production de L-lysine résulte de l'association entre l'*a*-aminoadipate et le glutamate *via* la réaction `L-LYSINE-AMINOTRANSFERASE-RXN`. Notons que cette réaction est réversible et, bien qu'elle contribue à la régulation des niveaux de L-lysine et de glutamate dans la cellule, elle semble être préférentiellement utilisée dans le sens de la dégradation de la L-lysine. Comprendre pourquoi la biosynthèse de L-lysine s'effectue uniquement par cette voie dans ***i*Prub22** sera donc un élément à étudier pour les développements futurs du modèle.

■ **Distribution des flux pour la biosynthèse de la L-alanine**

La biosynthèse de la L-alanine se déroule principalement par la transamination du pyruvate. Dans Prubens, le groupe aminé transféré au pyruvate provient de la glutamine (*i.e.* réaction `r0542` : `L-alanine + 2-Ketoglutarate ↔ Glutamine + Pyruvate`), tandis que dans ***i*Prub22**, il provient de la phénylalanine (réaction `2.6.1.58-RXN` : `Phénylalanine + Pyruvate ↔ Phénylpyruvate + L-alanine`). Remarquons qu'au sein d'***i*Prub22**, les réactions `r0542`, `RXN-13698` et `ALANINE-AMINOTRANSFERASE-RXN` partagent la même équation (*i.e.* réaction tripliquée), ce qui nous a conduit à bloquer les deux premières réactions dans le modèle ***i*Prub22** (cf. *Exemples de réactions dupliquées*, page 208). Enfin, dans ***i*Prub22**, une autre voie menant à la production de L-alanine est observée *via* la réaction `3-HYDROXY-KYNURENINASE-RXN` (`3-hydroxy-L-kynurenine + H2O → H+ + 3-hydroxyanthranilate + L-alanine`), suggérant ainsi qu'un autre chemin issu de la dégradation du tryptophane pourrait être emprunté pour la production de L-alanine.

■ **Distribution des flux pour la biosynthèse du L-glutamate**

Les distributions des flux menant à la biosynthèse du L-glutamate sont relativement similaires dans les modèles Prubens et ***i*Prub22**. Cependant, dans le contexte de la production de biotine, le rôle du L-glutamate diffère : dans Prubens, il contribue à la biosynthèse de la L-alanine, tandis que dans ***i*Prub22**, il est impliqué dans la production de L-lysine. Principalement, le L-glutamate est formé à partir de l'*a*-cétoglutarate, un intermédiaire du cycle de l'acide citrique, par amination réductrice avec de l'ammonium ou de la glutamine

comme source d'azote. Cette réaction est représentée par les identifiants `r0553` dans Prubens et `RXN-20084` dans ***i*Prub22**. Notons que ces deux réactions modélise le même processus chimique (*i.e.* `NH3`$^{+}$`+ 2-Ketoglutarate + NADPH + 2H`$^{+}$ `→ Glutamate + NADP`$^{+}$ `+ H`$_2$`O`) mais que la réaction issues de *i*AL1006 et non équilibrée (cf. *Exemples de réactions redondantes*, page 209), en conséquence, elle a été bloquée au sein d'***i*Prub22**. Parmi les autres réactions générant un flux vers la production du L-glutamate, nous comptons une réaction de transamination impliquant le *γ*-aminobutyrate et l'*α*-cétoglutarate (*i.e.* `GABATRANSAM-RXN`), ainsi que sept réactions d'amination, où un groupe amino de la glutamine est transféré à un autre substrat (*i.e.* `PABASYN-RXN`, `ANTHRANSYN-RXN`, `L-GLN-FRUCT-6-P-AMINOTRANS-RXN`, `PRPPAMIDOTRANS-RXN`, `FGAMSYN-RXN`, et pour Prubens, s'ajoute également les réactions `GMP-SYN-GLUT-RXN` et `GLUTAMIDOTRANS-RXN`).

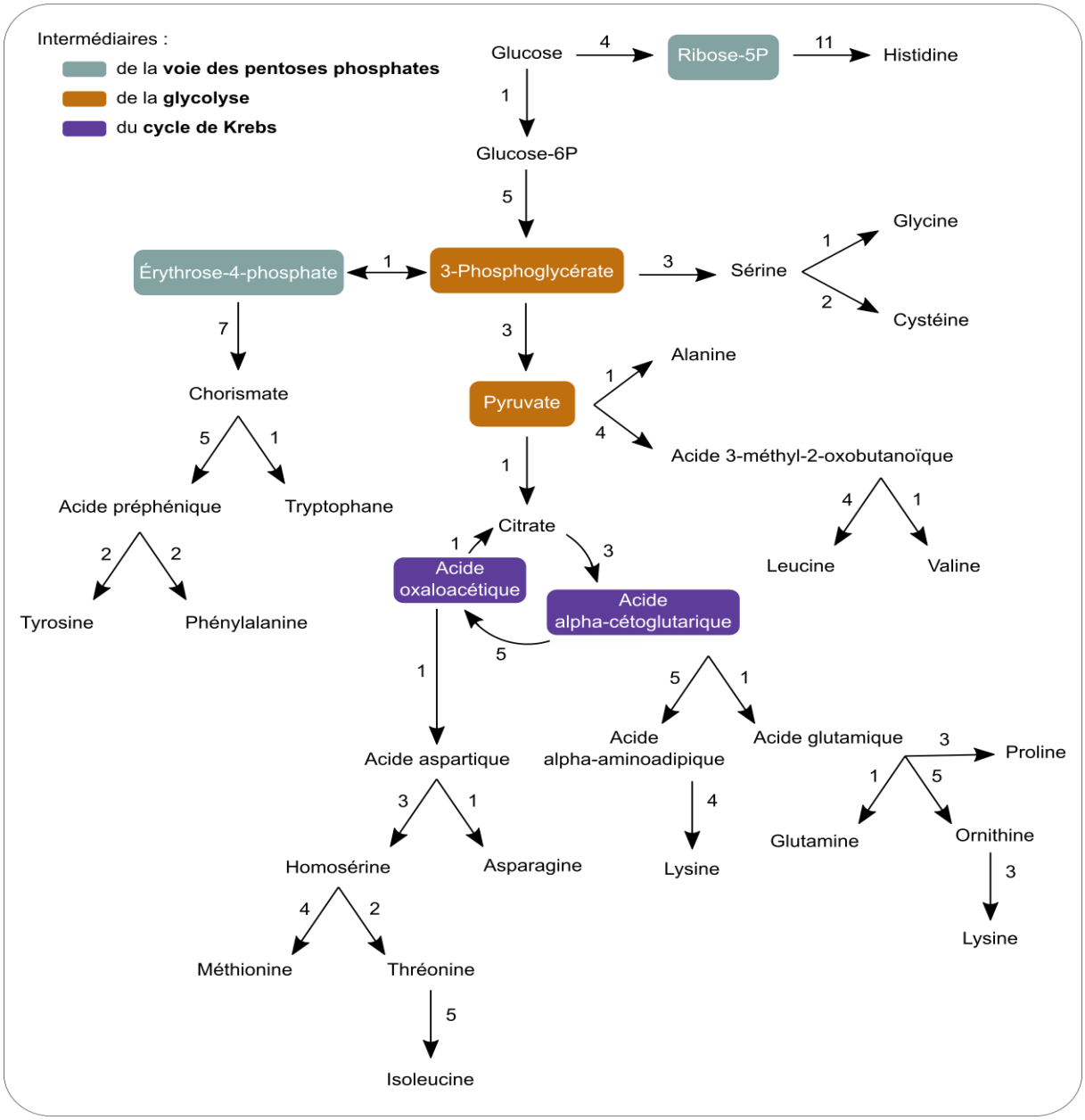


Figure 3-29 : Résumé des voies de biosynthèse des acides aminés de Penicillium rubens. *Les acides aminés sont classés en trois groupes selon la voie métabolique d'origine du précurseur initial de leur synthèse : la glycolyse, la voie des pentoses phosphates et le cycle de Krebs. Les données sont extraites de Nielsen et al. (1997) et les numéros indiquent le nombre de réactions intermédiaires entre chaque métabolite.*

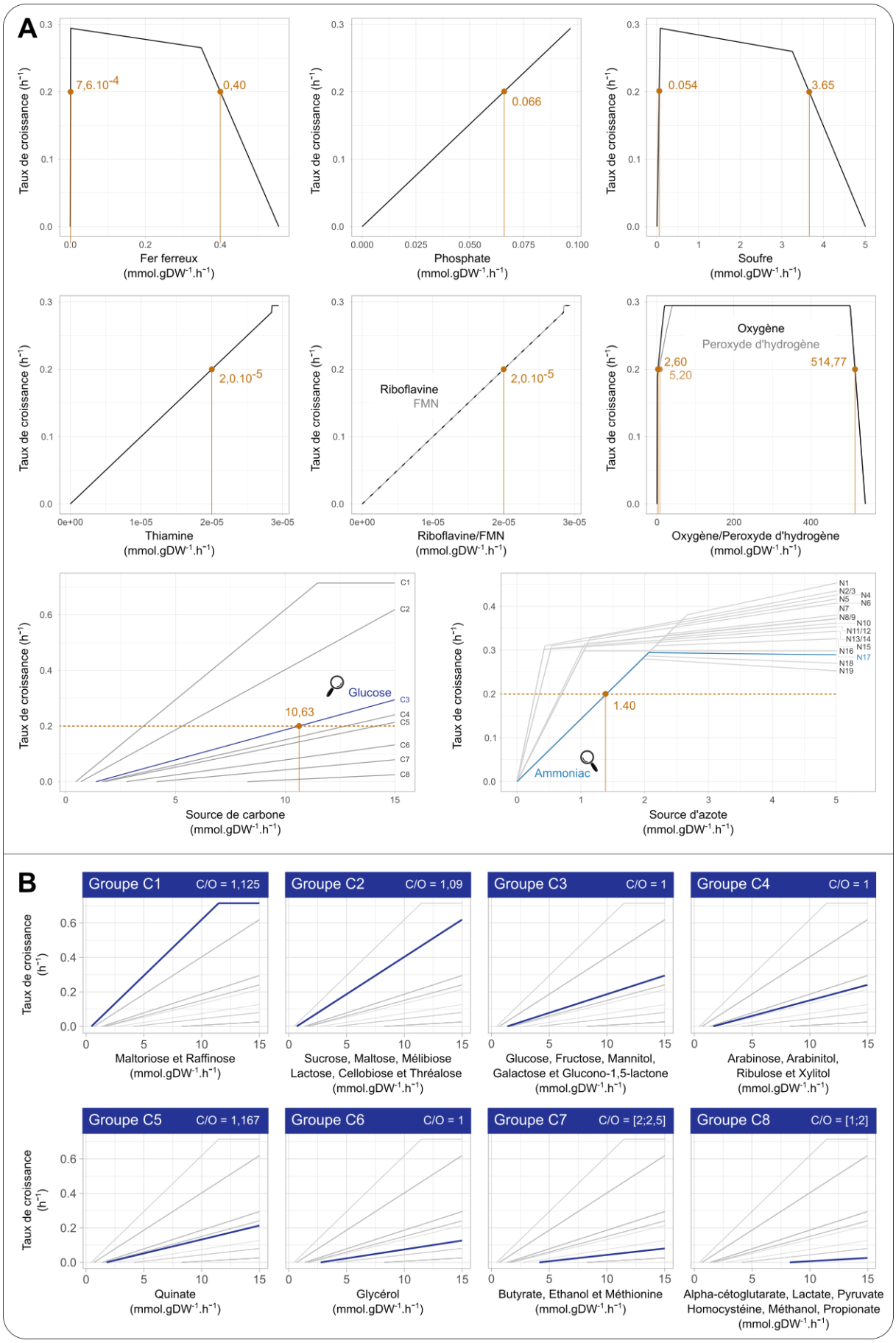
Nous nous sommes axés volontairement sur des comparaisons qualitatives puisque quantitativement, ces dernières se complexifient. Nous rappelons, que le taux de croissance maximal de *P. rubens* est évalué à environ 0,2 h^{-1} dans des conditions standards où la source de carbone est le glucose *(Nielsen 1997)*.

Initialement, étant donné que le modèle *i*AL1006 est non paramétré, l'utilisateur doit lui-même définir la fonction objective et fixer les contraintes. En accord avec les données de l'article, une première FBA a été réalisée en fermant toutes les réactions d'*uptakes* à l'exception des imports de glucose, d'oxygène, de phosphate, de soufre, de pimelate, de thiamine et d'ammoniac, laissés à l'infini, et en fermant les réactions `r0181`, `r0019`, `r0020` et `r0081` (*i.e.* réactions associées à des cycles futiles). Avec ces paramètres, le taux de croissance est de 21,8656 h^{-1}, comparé à 28,7070 h^{-1} sans contraintes. En fixant les bornes inférieure et supérieure de la réaction `r1466` (*i.e.* NGAM) à 1 $mmol.gDW^{-1}.h^{-1}$, et en bloquant les réactions artificielles utilisées pour la recherche des fuites énergétiques (*i.e.* freeATP, freeNADH et freeNADPH), le taux de croissance diminue à 10,7033 h^{-1}. Ces contraintes, issues des données supplémentaires de l'article, n'ont cependant pas été suffisante d'une part pour aboutir à un taux de croissance cohérent et d'autres part pour reproduire les résultats présenté dans ces données supplémentaires.

Pour Prubens, par défaut, les imports d'oxygène et de glucose sont fixés dans le modèle à un maximum de 20 $mmol.gDW^{-1}.h^{-1}$, les bornes de la fonction NGAM sont fixées à 1 $mmol.gDW^{-1}.h^{-1}$, et les imports de soufre, thiamine, diphosphate inorganique, pimelate et ammoniac sont laissés à une valeur infinie. Ceci aboutit à un flux de biomasse de 0,8745 h^{-1}. En ajustant l'import d'oxygène à 10 $mmol.gDW^{-1}.h^{-1}$ et celui de glucose à 5 $mmol.gDW^{-1}.h^{-1}$, conformément aux indications de l'article, un flux de 0,7709 $mmol.gDW^{-1}.h^{-1}$ est alors obtenu. Or, l'article mentionne que les taux de croissance maximaux des espèces de *Penicillum* se situent dans l'intervalle 0,305-0,323 h^{-1}. Par conséquent, pour ce modèle également, les résultats n'ont pas pu être reproduits, expliquant dès lors l'arrêt des comparaisons.

2.2.3.7. Robustesse d'*i*Prub22 aux variations des flux d'importation de nutriments

Dans le **Tableau 3-10** (page 292), nous avons rappelé quels étaient les imports minimaux nécessaires pour assurer la production de biomasse. Pour aller plus loin, nous présentons en **Figure 3-30**, une analyse de robustesse réalisée avec la fonction `robustnessAnalysis()` de COBRA Toolbox en examinant comment les variations dans les niveaux des flux d'import affectent la faisabilité ou les performances du flux de biomasse. Cette analyse permet de comprendre la tolérance du réseau métabolique ***i*Prub22** aux perturbations, et, le cas échant, d'identifier les points critiques du réseau. Il est cependant important de noter que l'analyse présentée se concentre uniquement sur la réponse du réseau à la variation d'un seul flux. Pour explorer les interactions entre deux ou plusieurs flux, cette analyse devra être complétée, par exemple, par une analyse PhPP. La complémentarité de ces approches permettra ainsi d'obtenir une compréhension plus complète de ce réseau métabolique.


A
Taux de croissance (h-1)
7,6.10-4
0,40
Fer ferreux
(mmol.gDW-1.h-1)
0.066
Phosphate
(mmol.gDW-1.h-1)
0.054
3.65
Soufre
(mmol.gDW-1.h-1)
2,0.10-5
Thiamine
(mmol.gDW-1.h-1)
Riboflavine
FMN
2,0.10-5
Riboflavine/FMN
(mmol.gDW-1.h-1)
Oxygène
Peroxyde d'hydrogène
2,60
5,20
514,77
Oxygène/Peroxyde d'hydrogène
(mmol.gDW-1.h-1)
C1
C2
Glucose
C3
C4
C5
C6
C7
C8
10,63
Source de carbone
(mmol.gDW-1.h-1)
N1
N2/3
N4
N5
N6
N7
N8/9
N10
N11/12
N13/14
N15
N16
N17
N18
N19
1.40
Ammoniac
Source d'azote
(mmol.gDW-1.h-1)
B
Groupe C1 C/O = 1,125
Maltoriose et Raffinose
(mmol.gDW-1.h-1)
Groupe C2 C/O = 1,09
Sucrose, Maltose, Mélibiose
Lactose, Cellobiose et Thréalose
(mmol.gDW-1.h-1)
Groupe C3 C/O = 1
Glucose, Fructose, Mannitol,
Galactose et Glucono-1,5-lactone
(mmol.gDW-1.h-1)
Groupe C4 C/O = 1
Arabinose, Arabinitol,
Ribulose et Xylitol
(mmol.gDW-1.h-1)
Groupe C5 C/O = 1,167
Quinate
(mmol.gDW-1.h-1)
Groupe C6 C/O = 1
Glycérol
(mmol.gDW-1.h-1)
Groupe C7 C/O = [2;2,5]
Butyrate, Ethanol et Méthionine
(mmol.gDW-1.h-1)
Groupe C8 C/O = [1;2]
Alpha-cétoglutarate, Lactate, Pyruvate
Homocystéine, Méthanol, Propionate
(mmol.gDW-1.h-1)

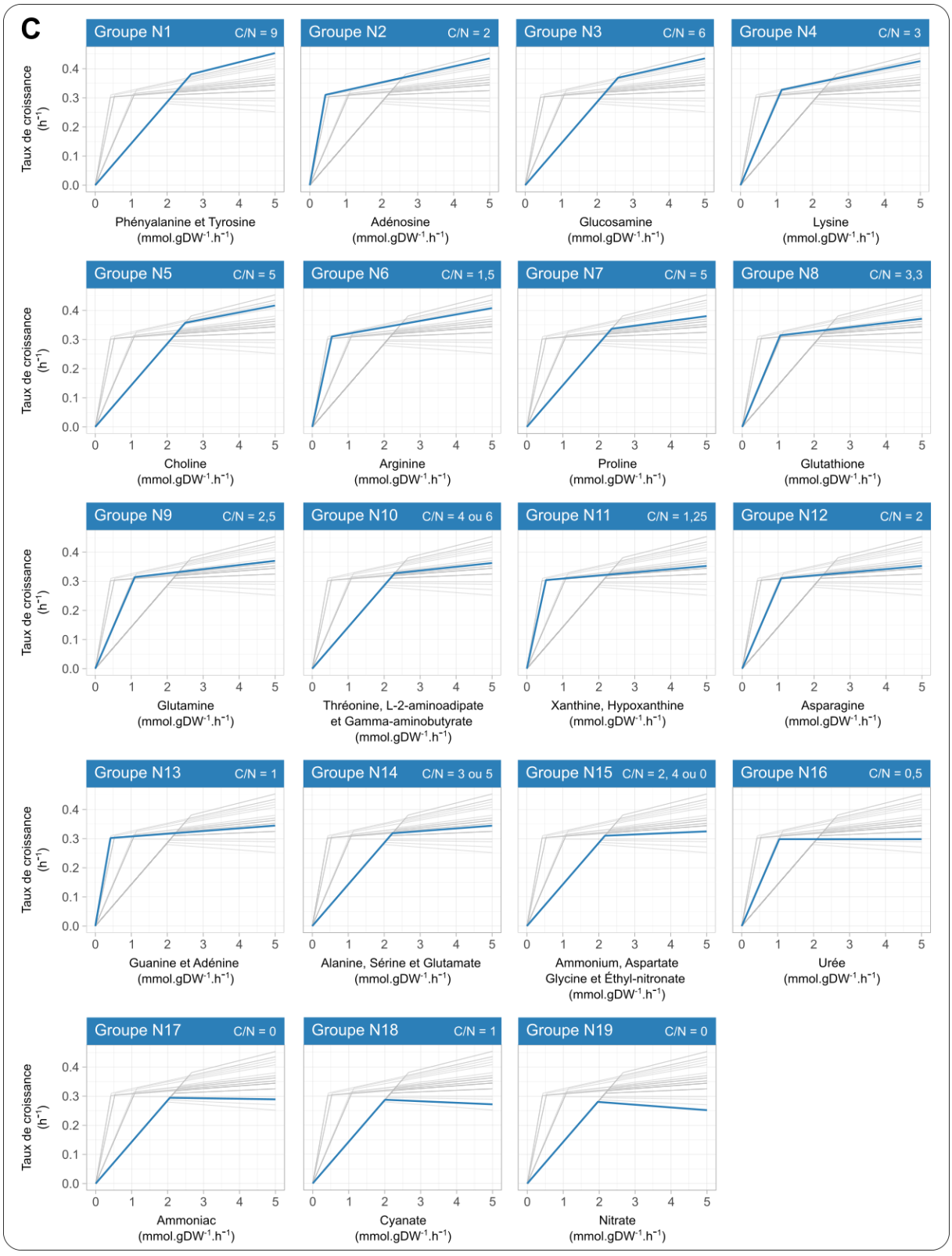


Figure 3-30 : Analyses de robustesse sur les uptakes minimaux nécessaires pour assurer la fonctionnalité d'iPrub22.
Les divers profils de robustesse ont été obtenus à l'aide de la fonction `robustessAnalysis` *de la Cobra Toolbox.* ***(A)*** *Profils de robustesse réalisés pour chacun des imports minimaux déterminés précédemment (formules brutes indiquées ci-contre). Les points orange correspondent aux valeurs de flux des imports lorsque la valeur de 0,20 h^{-1} de biomasse est atteinte, valeur déterminée expérimentalement selon Nielsen et al. (1997).* ***(B)*** *Détails des analyses de sensibilité pour les diverses sources de carbone, regroupées en huit catégories (C1 à C8). À l'exception des groupes C5, C7 et C8, les tendances montrent de manière assez logique que plus le ratio carbone/oxygène (i.e. disponibilité en carbone) diminue, et plus le nombre de carbones diminue, plus l'apport doit être élevé pour atteindre la valeur optimale de l'objectif. Notons toutefois que la borne arbitraire de 15 $mmol.gDW^{-1}.h^{-1}$ n'est pas suffisante pour atteindre les plateaux de saturation.* ***(C)*** *Détails des analyses de sensibilité pour les diverses sources d'azote, regroupées en seize catégories (N1 à N19). Pour la source d'azote, le ratio carbone/azote semble moins influent pour la production de biomasse comparé au ratio carbone/oxygène des sources de carbone.*

Liste des formules brutes des sources de carbone présentées en Figure 3-30

Groupe C1 : *Maltoriose et Raffinose $C_{18}H_{22}O_{16}$*

Groupe C2 : *Sucrose, Maltose, Mélibiose, Lactose, Cellobiose et Thréalose $C_{12}H_{22}O_{11}$*

Groupe C3 : *Glucose, Fructose, Galactose $C_6H_{12}O_6$, Mannitol $C_6H_{14}O_6$ et Glucono-1,5-lactone $C_6H_{10}O_6$*

Groupe C4 : *Arabinose, Ribulose $C_5H_{10}O_5$, Arabinitol et Xylitol $C_5H_{12}O_5$*

Groupe C5 : *Quinate $C_7H_{11}O_6$*

Groupe C6 : *Glycérol $C_3H_8O_3$*

Groupe C7 : *Butyrate $C_4H_7O_2$, Éthanol C_2H_6O et Méthionine $C_5H_{11}NO_2S$*

Groupe C8 : *Lactate C3H6O3, α-cétoglutarate C5H6O5, Pyruvate C3H4O3, Homocystéine C4H9NO2S, Méthanol CH4O et Propionate $C_3H_5O_2$.*

Liste des formules brutes des sources d'azote présentées en Figure 3-30

Groupe N1 : *Phénylalanine $C_9H_{11}NO_2$ et Tyrosine $C_9H_{11}NO_3$*

Groupe N2 : *Adénosine $C_{10}H_{13}N_5O_4$*

Groupe N3 : *Glucosamine $C_6H_{13}NO_5$*

Groupe N4 : *Lysine $C_6H_{14}N_2O_2$*

Groupe N5 : *Choline $C_5H_{14}NO$*

Groupe N6 : *Arginine $C_6H_{14}N_4O_2$*

Groupe N7 : *Proline $C_5H_9NO_2$*

Groupe N8 : *Glutathione $C_{10}H_{17}N_3O_6S$*

Groupe N9 : *Glutamine $C_5H_{10}N_2O_3$*

Groupe N10 : *Thréonine $C_4H_9NO_3$, L-2-aminoadipate $C_6H_{11}NO_4$ et γ-aminobutyrate $C_4H_9NO_2$*

Groupe N11 : *Hypoxanthine $C_5H_4N_4O$, Xanthine $C_5H_4N_4O_2$*

Groupe N12 : *Asparagine $C_4H_8N_2O_3$*

Groupe N13 : *Guanine $C_5H_{10}N_2O_3$ et Adénine $C_5H_5N_5$*

Groupe N13 : *Alanine $C_3H_7NO_2$, Sérine $C_3H_7NO_3$ et Glutamate $C_5H_9NO_4$*

Groupe N15 : *Ammonium NH_4^+, Aspartate $C_4H_6NO_4$, Glycine $C_2H_5NO_2$ et Éthyl-nitronate $C_2H_5NO_3$*

Groupe N16 : *Urée CH_4N_2O*

Groupe N17 : *Ammoniac NH_3*

Groupe N18 : *Cyanate CNO^-*

Groupe N19 : *Nitrate NO_3^-*

Les interprétations des profils de robustesse fournissent des informations précieuses sur la sensibilité et les limites du réseau métabolique aux variations des importations de différents éléments. En identifiant les points de rupture, les plateaux de saturation et les phases descendantes, il est possible de mieux comprendre les exigences en nutriments et les limites physiologiques du système étudié. En outre, ces informations, relatives à la tolérance et à la résilience du système, peuvent guider les stratégies d'optimisation en identifiant les conditions optimales pour maximiser la production de biomasse ou d'autres produits d'intérêt.

À partir d'un flux cible, ici celui de la biomasse, et d'une liste de flux à contrôler, les flux d'import, l'algorithme `robustnessAnalysis` réalise une boucle d'optimisation pour 50 valeurs (*i.e.* nombre que nous avons défini arbitrairement) de flux contrôlé, uniformément espacées entre le minimum et le maximum du flux selon les conditions définies dans le modèle. Pour chaque valeur fixée de flux contrôlé (*i.e.* bornes inférieure et supérieure de la réaction d'*uptake* figée sur la une des 50 valeurs optimales), le modèle est optimisé pour maximiser le flux de la réaction cible lors d'une FBA (*i.e.* biomasse). La fonction renvoie ensuite les valeurs optimales de biomasse pour chaque valeur du flux de contrôle. Ce sont ces données, qui une fois tracées permettent d'obtenir les profils de robustesse sous forme graphique représentant la production de biomasse en fonction des valeurs des flux d'import.

De manière générale, une relation linéaire entre les axes indique que le flux cible augmente ou diminue proportionnellement avec le flux analysé. Une courbe décroissante dans une analyse de robustesse indique que l'augmentation du flux de la réaction analysée entraîne une diminution du flux cible. Cet événement traduit des interactions complexes dans le réseau métabolique, où des niveaux élevés d'un flux inhibent la production de biomasse en raison de divers mécanismes, tels que l'inhibition toxique (*e.g.* accumulation de métabolites intermédiaires toxiques) ou la surconsommation de ressources induisant un déséquilibre du bilan énergétique (*i.e.* consommation disproportionnée d'énergie ou de cofacteurs, détournant ces ressources de la biosynthèse de biomasse). Une relation non-linéaire signale des interactions plus complexes. Si la courbe présente des plateaux, cela signifie que dans ces régions, les variations du flux analysé n'affectent pas, ou peu, le flux cible. Ces zones sont donc des régions de tolérance dans le réseau. Enfin l'existence d'un point critique (*i.e.* valeur pour laquelle la courbe connaît un changement de pente abrupt) témoigne d'une forte sensibilité, ou vulnérabilité, du réseau à cette condition. Sachant cela, voici comment peuvent être interprétés les divers résultats présentés en **Figure 3-30** :

- Les profils des imports de **phosphate**, de **thiamine** et de **riboflavine** sont linéaires, signifiant que le flux de biomasse augmente proportionnellement avec les flux de ces imports. En revanche, les quantités de ces éléments nécessaires à la fonctionnalité du modèle possèdent des ordres de grandeurs différents puisque les imports de thiamine et de riboflavine, bien qu'indispensables, le sont en quantités infimes. Nous notons toutefois la présence d'un point de rupture concernant les imports de ces deux éléments avant d'atteindre un seuil plateau.
- Le profil de l'import d'**oxygène** révèle trois phases de croissance distinctes : une courbe ascendante avec une pente abrupte, un plateau de saturation et une courbe descendante. La première section nous montre ainsi une forte sensibilité du réseau à cette condition puisque de très faibles variations dans l'import d'oxygène induisent une modification conséquente du flux de biomasse. La seconde phase, une phase de saturation signifie que l'augmentation de l'import d'oxygène ne conduit pas à une augmentation significative de la production de biomasse et par conséquent l'oxygène n'est plus un facteur limitant. En revanche, la dernière phase présente une pente décroissante et indique que des niveaux élevés d'oxygène ont un effet inhibiteur sur la croissance, possiblement en raison d'un stress oxydatif ou d'autres effets toxiques.

- À l'exception de l'existence du plateau de saturation, les profils du **fer ferrique** et du **soufre** sont relativement similaires à celui de l'oxygène et les mêmes conclusions peuvent alors être tirées. Au début, l'importation accrue de fer ou de soufre améliore la croissance, indiquant que ces éléments sont limitants. Le système métabolique est sensible à l'augmentation de leur disponibilité. Une pente descendante signifie que des niveaux élevés d'importation de fer ou de soufre ont un effet négatif sur la croissance. Plusieurs mécanismes peuvent expliquer ce comportement tel que la toxicité (*i.e.* à des concentrations élevées le fer peut provoquer la formation de radicaux libres causant des dommages oxydatifs aux composants cellulaires), le déséquilibre métabolique (*i.e.* une surabondance de ces éléments peut perturber l'équilibre métabolique en surchargeant les systèmes de régulation et de stockage intracellulaires) ou la déplétion de cofacteurs (*i.e.* l'excès de fer ou de soufre peut entraîner une consommation excessive de cofacteurs ou d'autres éléments nécessaires à leur incorporation et leur utilisation, ce qui pourrait indirectement limiter la croissance).
- Parmi les éléments substituables, les profils de la **riboflavine** et de la **FMN** sont superposables, tandis que nous notons une légère différence entre ceux de l'**oxygène** et du **peroxyde d'hydrogène**. En conséquence, ces profils suggèrent que la riboflavine et la FMN peuvent être substituées l'une par l'autre dans le réseau métabolique sans affecter la production de biomasse. En revanche, à de faibles concentrations, le flux de biomasse est plus sensible à l'import d'oxygène qu'à celui du peroxyde d'hydrogène.
- En testant l'ensemble des **sources de carbone et d'azote** disponibles dans le réseau métabolique, les profils mettent en évidence l'existence respectivement de huit et seize groupes similaires aux comportements similaires (*i.e.* mécanismes du métabolisme et probable distribution des flux communs au sein de chaque groupe). Concernant les sources de carbone, à l'exception du groupe C1, les conditions testées sont insuffisantes pour atteindre le plateau de saturation. En revanche, la relation entre disponibilité en carbone et croissance est proportionnelle pour toutes les sources testées. Concernant les sources d'azote testées les groupes N1 à N15 présentent une courbe ascendante avec deux pentes, la première étant plus raide que la seconde, le groupe N16 montre une courbe ascendante aboutissant à un plateau de saturation et pour les groupes N17 à N19, nous observons une phase ascendante puis descendante. Dans les groupes N1 à N15, La première pente plus raide signale une forte sensibilité initiale à l'import d'azote, tandis que la seconde pente plus douce suggère une diminution de cette sensibilité à mesure que la disponibilité en azote augmente. Le plateau observé au sein du groupe N15 indique une saturation où l'augmentation supplémentaire d'azote ne conduit plus à une augmentation de la biomasse. Enfin, la phase descendante observée dans les groupes N16 à N19 montre que l'import excessif d'azote peut entraîner des effets inhibiteurs, réduisant ainsi la croissance.

2.2.3.8. Comparaison des phénotypes simulés par *i*Prub22 et des données expérimentales

Désormais, sauf mention contraire, les informations relatives aux données issues d'expérimentations biologiques présentées en **Tableau 3-11** sont tirées du livre : *Physiological Engineering Aspects of Penicillium chrysogenum* (Nielsen 1997). En raison du fort intérêt pharmacologique de *P. rubens*, espèce productrice d'antibiotiques, ses milieux optimaux de croissance et de production de pénicilline ont largement été étudiés et sont connus depuis longtemps. À l'instar de nombreux champignons filamenteux, *P. rubens* possède une grande capacité de biosynthèse et peut, à ce titre, se développer sur des milieux définis simples. Ainsi, la qualité prédictive d'***i*Prub22** dépend de sa capacité à simuler les divers phénotypes observés et répertoriés lors des expérimentations.

De primes abords, c'est-à-dire qualitativement, les résultats des diverses simulations d'***i*Prub22** semblent en adéquation avec ceux de la littérature. En revanche, quantitativement, ces résultats sont à nuancer et nécessitent des approfondissements comme en témoignent notamment les résultats de la croissance en utilisant les acides aminés. Rappelons que selon leur origine, les acides aminés sont divisés en deux classes, les acides aminés protéinogènes au nombre de 20 qui sont les constituants des protéines et les acides aminés non protéinogènes qui sont principalement utilisés dans les voies de biosynthèse des métabolites spécialisés (*e.g.* NRPS). Potentiellement, les acides aminés sont des sources de carbone et d'azote pouvant servir de composés nutritifs. Néanmoins, en fonction de leur structure (*e.g.* rapport du nombre de carbone et d'azote) ou de leur stabilité (*e.g.* acides aminés contenant du soufre) certains seront à privilégier. Au travers de diverses simulations effectuées sur ***i*Prub22**, nous ne retrouvons pas la granularité de « performance » attendue. Par exemple, nous constatons en **Figure 3-3** (page 151) que l'utilisation de l'aspartate, de l'alanine ou du glutamate en tant que source de carbone et d'azote sont parmi les acides aminés qui fournissent le taux de croissance le plus faible, alors que la lysine, un des acides aminés dont la dégradation en intermédiaires du métabolisme du carbone s'effectue avec le plus grand nombre d'étapes, permet d'obtenir l'un des taux de croissance les plus élevés. Ces résultats, en contredisant les observations expérimentales, viennent souligner le fait que des ajustements sont encore nécessaires pour affiner la qualité prédictive d'***i*Prub22**.

Selon les travaux de Allam et Elzainy (1969), la croissance de *P. chrysogenum* a été testée sur différentes purines et autres composés azotés. Leurs résultats montrent que l'hypoxanthine, suivie de l'adénine, sont parmi les meilleures sources d'azote pour favoriser la croissance. Viennent ensuite la xanthine, l'urée et le nitrate de sodium, que nous considérons équivalents à l'ammoniac dans nos simulations. La **Figure 3-31** reprend les données exposées en **Figure 3-30** pour vérifier si les comportements observés sont également présents dans les simulations d'***i*Prub22**. Nous constatons alors qu'à l'exception de la xanthine et de l'hypoxanthine, qui partagent le même profil de robustesse (groupe N10), les tendances observées expérimentalement semblent effectivement reproduites par les simulations.

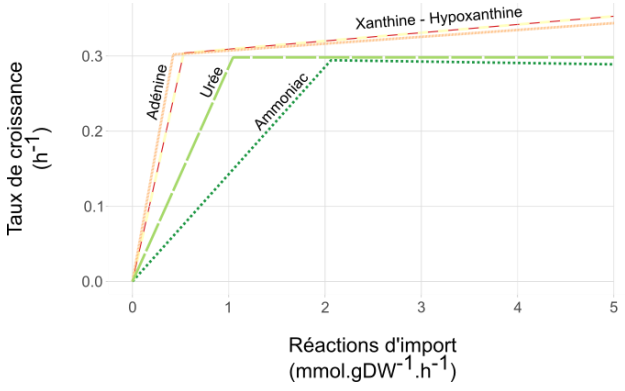


Figure 3-31 : Extrait des analyses de robustesse sur les uptakes minimaux nécessaires pour assurer la fonctionnalité d'iPrub22 centré sur les imports d'hypoxanthine, d'adénine, de xanthine, d'urée et d'ammoniac.

Tableau 3-11 : Adéquation qualitative entre les comportements observés de *Penicillium rubens* et ceux simulés avec *i*Prub22.

Les intitulés précédés d'une coche (✓) représentent des conditions simulées avec ***i*Prub22**. Ceux précédés d'une croix (×) n'ont pas été reproduits par les simulations, les résultats étant généralement inverses aux données expérimentales. Les intitulés précédés d'un cercle (O) nécessitent une vérification supplémentaire, soit parce que la simulation n'a pas été testée, soit parce que le résultat obtenu mérite de plus amples investigations.

Informations et observations extraites de *(Nielsen 1997)*		
Taux de croissance		
O	Le taux de croissance de *P. rubens*, déterminé à partir des mesures de la morphologie microscopique, est d'environ 0,20 h^{-1}	
Sources de carbone		
✓	Croissance utilisant le **glucose** comme seule source de carbone	
✓	Croissance utilisant le **sucrose** comme seule source de carbone	
✓	Croissance utilisant le **fructose** comme seule source de carbone	
✓	Croissance utilisant le **lactose** comme seule source de carbone	
✓	Croissance utilisant le **galactose** comme seule source de carbone	
O	Croissance utilisant le **gluconate** comme seule source de carbone	
✓	Croissance utilisant le **l'acétate** comme seule source de carbone	
Sources d'azote		
✓	Croissance utilisant l'**ammoniaque** comme seule source d'azote.	
✓	Croissance utilisant le **nitrate** comme seule source d'azote.	
✓	Croissance utilisant le **nitrite** comme seule source d'azote.	
✓	Croissance utilisant l'**adénine** comme seule source d'azote.	(Allam et Elzainy 1969)
✓	Croissance utilisant l'**urée** comme seule source d'azote.	
✓	Croissance utilisant le **nitrite** comme seule source d'azote.	
✓	Croissance utilisant la **xanthine** comme seule source d'azote.	
✓	Croissance utilisant l'**hypoxanthine** comme seule source d'azote.	Bonne source d'azote pour la croissance ✓
✓	Croissance utilisant l'**asparagine** comme seule source d'azote.	
✓	Croissance utilisant l'**arginine** comme seule source d'azote.	
✓	Croissance utilisant la **proline** comme seule source d'azote.	
✓	Croissance utilisant le **glutamate** comme seule source d'azote.	Source d'azote intérmédiare pour la croissance ✓
✓	Croissance utilisant l'**aspartate** comme seule source d'azote.	
✓	Croissance utilisant l'**alanine** comme seule source d'azote.	
O	Croissance utilisant l'**orthine** comme seule source d'azote.	O
O	Croissance utilisant l'**histidine** comme seule source d'azote.	Mauvaise source d'azote pour la croissance O
✓	Croissance utilisant la **glycine** comme seule source d'azote.	×
O	Croissance utilisant l'**isoleucine** comme seule source d'azote.	O
✓	Croissance utilisant la **lysine** comme seule source d'azote.	×
O	Croissance utilisant la **leucine** comme seule source d'azote.	O
O	Croissance utilisant la **méthionine** comme seule source d'azote.	×
✓	Croissance utilisant la **phénylalanine** comme seule source d'azote.	×
✓	Croissance utilisant la **tyrosine** comme seule source d'azote.	×
✓	Croissance utilisant la **thréonine** comme seule source d'azote.	×
O	Croissance utilisant la **valine** comme seule source d'azote.	O

Tableau 3-11 (suite)

Sources de carbone et d'azote

Les champignons filamenteux peuvent utiliser les acides aminés comme seule source d'azote et de carbone, mais de façon très médiocre dans la plupart des cas. En général, les acides aminés qui sont dégradés en intermédiaires du métabolisme du carbone en une ou deux étapes, par exemple le L-glutamate, sont de meilleurs candidats. NB. Les acides aminés suivants sont donc classés des meilleures sources théoriques aux moins bonnes.

✓	Croissance utilisant le **glutamate** comme source de carbone et d'azote.	Dégradation en 1 étape	✗
✓	Croissance utilisant l'**aspartate** comme source de carbone et d'azote.		✗
✓	Croissance utilisant l'**alanine** comme source de carbone et d'azote.		✗
✓	Croissance utilisant la **glutamine** comme source de carbone et d'azote.	Dégradation en 2 étapes	✓
✓	Croissance utilisant l'**asparagine** comme source de carbone et d'azote.		✓
✓	Croissance utilisant la **glycine** comme source de carbone et d'azote.		✓
✓	Croissance utilisant la **thréonine** comme source de carbone et d'azote.		✓
✓	Croissance utilisant la **proline** comme source de carbone et d'azote.	Dégradation en 3 étapes	✗
O	Croissance utilisant l'**orthine** comme source de carbone et d'azote.		O
✓	Croissance utilisant la **phénylalanine** comme source de carbone et d'azote.		✗
✓	Croissance utilisant la **tyrosine** comme source de carbone et d'azote.		✗
✓	Croissance utilisant l'**arginine** comme source de carbone et d'azote.	Dégradation en 4 étapes	✗
O	Croissance utilisant l'**histidine** comme source de carbone et d'azote.		O
O	Croissance utilisant la **leucine** comme source de carbone et d'azote.		O
O	Croissance utilisant la **méthionine** comme source de carbone et d'azote.		O
O	Croissance utilisant l'**isoleucine** comme source de carbone et d'azote.	Dégradation en 5 étapes	O
✓	Croissance utilisant la **lysine** comme source de carbone et d'azote.		✗
O	Croissance utilisant la **valine** comme source de carbone et d'azote.		O

Composés ignorganiques

✓	Le **soufre** est essentiel à la croissance (*i.e.* composant de certaines vitamines et des acides aminés cystéine et méthionine)
✓	Le **phosphore** est essentiel à la croissance (*i.e.* constituant des nucléotides, des phospholipides et de plusieurs vitamines)
✗	Le **potassium** est essentiel à la croissance (*i.e.* régule le potentiel osmotique cellulaire, joue un rôle central dans divers processus de transport et sert de cofacteur pour de nombreuses enzymes)
✗	Le **magnésium** est essentiel à la croissance (*i.e.* cofacteur essentiel de plusieurs enzymes impliquées par exemple dans le cycle TCA - rôle majeur dans le transfert des groupes phosphates où il forme des complexes avec l'ATP et l'ADP)
✓	Le **fer** est essentiel à la croissance (*i.e.* activateur enzymatique et composant des porphyrines héminiques, qui sont impliquées dans la chaîne de transport des électrons - absorption au moyen de chélateurs appelés sidérophores, qui peuvent également être impliqués dans le stockage du fer)
✗	Le **cuivre** est essentiel à la croissance (*i.e.* activateur essentiel de plusieurs enzymes – peut inhiber la croissance à des concentrations sous-optimales).
✗	Le **calcium** favorise la croissance (*i.e.* composant essentiel pour diverses structures cellulaires et régule plusieurs fonctions cellulaires spécifiques telles que l'induction de la conidiation)
O	Le **manganèse** favorise la croissance (*i.e.* constituant pouvant servir de substitut au magnésium dans de nombreuses enzymes – une carence en ce minéral peut entraîner une réduction des niveaux de protéines, de lipides et d'acides nucléiques)
O	Le **zinc** favorise la croissance (*i.e.* constituant de nombreuses enzymes essentielles – en cas de carence la croissance des champignons filamenteux est ralentie significativement)
✗	Le **cobalt** favorise la croissance (*i.e.* participe à de nombreuses réactions cellulaires, mais les enzymes activées par le cobalt le sont généralement aussi par d'autres cations divalents).

Tableau 3-11 (suite)

Effets des conditions environnementales et des substrats sur le métabolisme et la croissance

▲Croissance et métabolisme

o	**Machinerie métabolique** : À des taux de croissance spécifiques faibles, les cellules possèdent une machinerie métabolique inutilisée ou inefficacement utilisée qui peut être rapidement activée lorsque les conditions environnementales changent.
o	**ARN et croissance** : Le pool d'ARN stable et le taux de croissance sont positivement corrélés.
o	**Niveaux de nucléotides** : Pendant la croissance en surface de *P. chrysogenum*, le pool de GTP, d'ATP et d'UTP libres sont respectivement compris entre [3,0 - 3,6], [19,0 - 25,6] et [1,4 - 1,5] µmoles.gDW^{-1}.
o	**Rendement biomasse** : Le rendement de la biomasse à partir du gluconate est inférieur à celui de la croissance sur sucres (*i.e.* approximativement 50 %).

▲Conditions de croissance spécifique

o	**Croissance sur lactose :** La croissance de *P. chrysogenum* sur le lactose comme seule source de carbone est lente. Ce milieu a été utilisé avec succès dans la production de pénicilline avant l'introduction du procédé fed-batch.
o	**Métabolites précurseurs** : Les besoins minimaux en métabolites précurseurs, exprimés en µmol.gDW^{-1} pour la synthèse d'une cellule de Penicillum *chrysogenum* sont les suivants : Glucose-6-P = 1 110 ; Fructose-6-P = 356 ; Ribose-5-P = 498 ; Erythtose-4-P = 315 ; Glycéraldéhyde-3-P =50 ; 3-Phospoglycérates = 880 ; Phosphoenolpyruvate = 589 ; Pyruvate = 1 923 ; Acetyl-Coa = 2 306 ; α-cétoglutarate = 1 215 ; Oxaloacetate = 1 175.

▲Effets métaboliques du glucose

o	**Formation d'acide gluconique** : La croissance sur glucose entraîne la formation de grandes quantités d'acide gluconique.
o	**Flux glycolytiques :** Lors de la glycolyse (*i.e.* somme des trois voies métaboliques par lesquelles le glucose est converti en pyruvate : la voie d'Embden-Meyerhof-Parnas (EMP), la voie des pentoses-phosphates (PP) et la voie d'Entner-Doudoroff (ED)) les flux à travers la voie d'EMP sont nettement supérieurs à ceux de la voie des PP.
o	**Métabolisme du pyruvate :** Le pyruvate est le dernier métabolite de la glycolyse et il peut suivre plusieurs voies en fonction de l'état redox et énergétique des cellules. En croissance oxydative normale, le pyruvate entre dans le cycle de Krebs *via* l'acétyl-CoA et est oxydé en dioxyde de carbone et en eau. En conditions d'oxygène limitées, le pyruvate est converti en acétate ou en éthanol.
o	**Inhibition du métabolisme du lactate :** En présence de glucose, le métabolisme du lactate est intégralement inhibé.
o	**Croissance en trois phases sur glucose** : Lors de la croissance sur un mélange de glucose et de fructose, le glucose réprime l'absorption du fructose. Une fois le glucose épuisé, après une courte phase de latence, le fructose est absorbé. La croissance s'effectue en trois phases avec un taux de croissance spécifique de 0,20 h^{-1} lors de la première phase et de 0,12 h^{-1} lors de la troisième phase.

▲Métabolisme des acides aminés et autres intermédiaires

o	**Précurseurs biosynthétiques :** L'α-cétoglutarate et l'oxaloacétate, deux intermédiaires du cycle de Krebs, servent de métabolites précurseurs pour la biosynthèse des acides aminés et des nucléotides.
o	**Catabolisme des acides aminés :** Important à faible concentration de glucose, mais très faible à forte concentration de glucose..
o	**Toxicité de la L-cystéine** : À des concentrations élevées, la L-cystéine est toxique et se lie au glutathione, le thiol intracellulaire le plus abondant.

Tableau 3-11 (suite et fin)

▲ Métabolisme de l'acétate et des acides gras

o	**Gluconéogenèse** : Lors de la croissance sur acétate et acides gras, la gluconéogenèse (*i.e.* inverse de la voie EMP, nécessaire lorsque les nutriments du milieu de culture sont insuffisants pour assurer la production des précurseurs de la glycolyse) est initiée avec l'oxaloacétate formé à partir du succinate. En croissance sur lactate, la première étape de la gluconéogenèse est la formation d'oxaloacétate à partir de pyruvate.
o	**Synthèse de l'acétyl-CoA** : Chez *P. chrysogenum*, il est peu probable que l'acétyl-CoA soit synthétisé par l'acétate. Sa synthèse s'effectue probablement *via* la citrate lyase, qui clive le citrate cytoplasmique en oxaloacétate et en acétyl-CoA.

▲Inhibiteurs de croissance

✖	**Effet du CO_2** : À des concentrations élevées, le CO_2 inhibe la croissance des champignons filamenteux, entraînant une diminution de la viabilité des spores, un retard de germination, un prolongement de la phase de dormance, une augmentation des ramifications des hyphes et une augmentation de l'activité de la chitinase.
✖	**Toxicité du peroxyde d'hydrogène :** Le peroxyde d'hydrogène est toxique pour les cellules et est rapidement décomposé en eau et en oxygène par la catalase.

3 - Focus sur le métabolisme spécialisé : descriptions de diverses voies de biosynthèse de métabolites spécialisés

Pour clore ce chapitre, nous allons désormais nous intéresser plus particulièrement au métabolisme spécialisé de *P. rubens* Wisconsin 54-1255. Au cours de la section *1 - Génération d'une nouvelle reconstruction, iPrub22* (page 127), nous avons déjà constaté que la notion de métabolite spécialisé varie selon l'axe employé (*i.e.* biologique, chimique, bio-informatique), revêtant ainsi des définitions légèrement différentes. En résumé, un métabolite spécialisé est, d'un point de vue biologique, tout composé conférant un avantage sélectif à l'organisme qui le produit. D'un point de vue chimique, et pour les champignons filamenteux, un métabolite spécialisé se caractérise par sa structure et son mode de production (*i.e.* composés tels que les NRPS, les PKS, les terpènes ou les éléments hybrides résultant d'un cluster de gènes – cf. encart : *Classification des métabolites spécialisés*, page 79). Enfin, et notamment en raison des annotations présentes dans diverses bases de données, nous avons remarqué que la notion de métabolite spécialisé en bio-informatique est bien plus confuse. Ces divers degrés de précision sont alors des vecteurs potentiels de malentendus lors de l'interprétation des résultats d'analyses.

Outre ces éléments, les thèmes liés à la nécessité d'une vérification manuelle et aux limites de l'automatisation de la reconstruction pour le métabolisme spécialisé ont également été abordés, tout comme le fait qu'***i*Prub22** soit capable de simuler des comportements de croissance différents selon la nature de l'environnement fourni. De plus, ***i*Prub22** semble capable de produire un ensemble de métabolites spécialisés en fonction des contraintes initiales du modèle. Rappelons à cet effet que l'objectif principal des travaux que nous présentons ici était de fournir une ressource capable de simuler les réponses d'un organisme lorsque son milieu de croissance varie, une approche que nous qualifions d'**OSMAC *in silico***.

Ainsi, nous nous attarderons sur **(1)** les raisons qui nous ont motivés à sélectionner les huit voies de biosynthèse de métabolites que nous présentons ici, et **(2)** nous détaillerons ensuite la topologie de ces voies au sein d'***i*Prub22**. Nous questionnerons ensuite la robustesse d'***i*Prub22** et sa capacité à réagir à différentes perturbations, d'une part, **(3)** en analysant l'impact potentiel de modifications quantitatives des précurseurs de la réaction de biomasse et, d'autre part, **(4)** nous conclurons en explorant l'impact des conditions environnementales sur la production de ces métabolites, pour, le cas échéant, identifier les conditions optimales permettant de maximiser leur biosynthèse.

3.1. Comment sélectionner les métabolites spécialisés à étudier ?

En préambule, nous soulignons que la pertinence de la production de tel ou tel autre composé est discutable, et est discutée au sein même de la communauté scientifique. La reconstruction de ce modèle est un instantané des connaissances incluses dans les bases de données. Le **GSMN** proposé exprime les potentialités de production portée par le génome de la souche Wisconsin 54-1255, sans tenir compte de nombreux facteurs tels que la compartimentation intracellulaire ou les processus de régulation.

La première interrogation que nous nous sommes posée est la suivante : est-il possible de quantifier de manière globale la proportion de métabolites spécialisés contenus dans ***i*Prub22** ? En tirant profit des annotations présentes dans MetaCyc (v24.5), une approche générique pour caractériser l'enrichissement du modèle en métabolites spécialisés a été effectuée. La sélection des données a été guidée à l'aide des annotations des composés et des pathways intrinsèques à cette base de données. À la date du 17 mars 2021, il existait 2 519 métabolites étiquetés comme étant « *a secondary metabolite* ». À titre d'exemple, les diverses pénicillines (*i.e.* G, K, N, et V) n'appartenaient pas à cet ensemble de composés. En revanche, les pathways qui les contenaient étaient eux associés au métabolisme secondaire. Ainsi, les 1 397 produits finaux des 899 pathways « *Secondary metabolite Biosynthesis* » et les 162 réactants des 125 pathways « *Secondary metabolite Degradation* » ont été utilisés pour maximiser l'étendue de la recherche. Les pénicillines K et V étaient alors effectivement présentes dans le jeu de données considéré. En revanche, les pénicillines G et N n'appartenaient toujours pas à cet ensemble.

Au vu de ces résultats, il nous a alors semblé plus opportun d'axer et de restreindre nos tests de capacité de productibilité du modèle sur les métabolites spécialisés de *P. rubens* Wisconsin 54-1255 les plus documentés. La revue de Guzmán-Chávez et *al.* *(2018)* rapporte l'existence de 50 clusters de gènes biosynthétiques, comprenant 33 gènes cœur : 10 NRPS, 20 PKS, 2 hybrides NRPS-PKS et 1 DMAT. Parmi ces BGCs, seuls 14 (*i.e.* 28 %)sont associés à des produits naturels clairement identifiés. Les gènes cœurs de chacun de ces clusters et leurs informations associées sont présentés dans le **Tableau 3-12**.

De manière plus détaillée, parmi les peptides non-ribosomaux identifiés issus des NRPS, nous trouvons la fungisporine *(Ali et al. 2014; Salo et al. 2015; Nielsen et al. 2017; Guzmán-Chávez et al. 2018)*, les roquefortines C *(Houbraken et al. 2012; Ali et al. 2013; Salo et al. 2015; Nielsen et al. 2017; Guzmán-Chávez et al. 2018)*, D *(Houbraken et al. 2012; Ali et al. 2013; Salo et al. 2015)*, F *(Salo et al. 2015)*, M *(Salo et al. 2015)*, N *(Salo et al. 2015)* ainsi que les produits associés tels que la méléagrine *(Houbraken et al. 2012; Ali et al. 2013; Salo et al. 2015; Nielsen et al. 2017)*, les glandicolines A et B et les précurseurs histidyl-tryptophanyl-dicétopipérazine (HTD) et dehydrohistidyl-tryptophanyl-dicétopipérazine (DHTD) *(Ali et al. 2013; Salo et al. 2015)*. Cette liste de peptides non-ribosomaux est également enrichie par les trois sidérophores suivants *(Samol 2015; Guzmán-Chávez et al. 2018)*, le coprogène, le ferrichrome et les composés de la famille des fusarinines. Parmi les polycétides identifiés, nous pouvons citer les composés appartenant à la famille des sorbicillinoïdes *(Houbraken et al. 2012; Meng et al. 2016; Guzmán-Chávez et al. 2018)* et la chrysogine *(Houbraken et al. 2012; Guzmán-Chávez et al. 2018; Viggiano et al. 2018)*.

Tableau 3-12 : Séquences génomiques des enzymes clés dans la synthèse des métabolites spécialisés de *Penicillium rubens* Wisconsin 54-1255. Ce tabeau répertorie les identfiants des séquences essentielles des BGCs, indique leur présence dans ***i*Prub22** et, le cas échéant, le nombre de réactions qui leurs sont associées. Lorsqu'ils sont connus les noms des gènes, les protéines associées à chaque séquence identifiée *(Guzmán-Chávez et al. 2018)* ou *putative (Van den Berg et al. 2008)* et le produit résultant du cluster sont indiqués.

NRPS	*Pc13g05250*	✓	10	pssC	Sidérophore synthétase	Sidérophore
	Pc13g14330	✓	17	NRPS2	Tétrapeptide synthétase	-
	Pc16g03850	✓	9	pssA	Sidérophore synthétase	Coprogène
	Pc16g04690	✓	15	hcpA	Cyclique tétrapeptide synthétase	Fungisporine
	Pc21g01710	✓	12	nrpsA	Dipeptide synthétase	Brevianamide
	Pc21g10790	✓	17	NRPS6	Hexapeptide synthétase	-
	Pc21g12630	✓	22	chyA	2-Aminobenzamide synthétase	Chrysogine
	Pc21g15480	✓	21	roqA	Hisdidyl-tryptophanyldiketo-piperazine synthétase	Roquefortine/ Méléagrine
	Pc21g21390	✓	1	pcbAB	α-Aminoadipyl-cysteninyl-valine synthétase	β-lactames
	Pc22g20400	✓	9	pssB	Sidérophore synthétase	Fusarinines
PKS	*Pc12g05590*	✓	62	pks1	*Similitude avec l'équisétine synthase*	-
	Pc13g04470	✓	60	pks2	*Similitude avec la lovastatine diketide synthase*	-
	Pc13g08690	✓	60	pks3	*Similitude avec la lovastatine diketide synthase*	-
	Pc16g00370	✓	60	YanA	6-MSA synthase	6-MSA / Yanuthones
	Pc16g03800	✓	60	pks5	-	-
	Pc16g04890	✓	61	pks6	-	-
	Pc16g11480	✓	60	pks7	*Similitude avec la lovastatine diketide synthase*	-
	Pc21g00960	✓	74	pks8	*Similitude avec la lovastatine diketide synthase*	-

Tableau 3-12 (suite et fin)

	Pc21g03930	✓	61	pks9	*Similitude avec la lovastatine diketide synthase*	-
	Pc21g03990	✓	60	pks10	-	-
	Pc21g04840	✓	60	pks11	-	
	Pc21g05070	✓	7	SorB	Sorbicilline synthase	Sorbicillinoides
	Pc21g05080	✓	62	SorA	Sorbicilline synthase	Sorbicillinoides
	Pc21g12440	✓	60	pks14	*Similitude avec la lovastatine diketide synthase*	-
	Pc21g12450	✓	68	pks15	-	-
	Pc21g15160	✓	61	pks16	*Similitude avec la lovastatine diketide synthase*	-
	Pc21g16000	✓	68	alb1	YWA1 synthase	YWA1 / DHN-Mélanine
	Pc22g08170	✓	60	PatK	6-MSA synthase	6-MSA
	Pc22g22850	✓	2	AdrD	DMOA synthase	DMOA / Andrastine
	Pc22g23750	✓	59	pks20	*Similitude avec la lovastatine diketide synthase*	-
NRPS-like	*Pc06g01540*	✓	15	NRPS-like1	*Similitude avec la saframycine Mx1 synthase*	-
	Pc12g09980	✓	1	NRPS-like2	*Similitude avec l'acide-CoA ligase*	-
	Pc12g13170	✗	-	NRPS-like3	-	-
	Pc13g12570	✓	15	NRPS-like4	*Similitude avec la saframycine Mx1 synthase*	-
	Pc14g01790	✓	2	NRPS-like5	-	-
	Pc16g09930	✓	13	NRPS-like6	-	-
	Pc18g00380	✓	7	NRPS-like7	*Similitude avec la saframycine Mx1 synthase*	-
	Pc20g02260	✗	-	NRPS-like8	*Similitude avec l'enzyme aminoadipate réductase*	-
	Pc20g02590	✓	15	NRPS-like9	*Similitude avec la saframycine Mx1 synthase*	-
	Pc20g09690	✓	2	NRPS-like10	-	-
	Pc20g12670	✓	9	NRPS-like11	-	-
	Pc21g22530	✓	9	NRPS-like12	-	-
	Pc21g22650	✓	2	NRPS-like13	-	-
	Pc22g06310	✓	18	NRPS-like14	*Similitude avec la L-aminoadipate-semialdehyde déshydrogénase*	-
	Pc22g09430	✓	2	NRPS-like15	-	-
NRPS-PKS	*Pc14g00080*	✓	69	NRPS-PKS 1	*Similitude avec la nonakétide synthase*	-
	Pc16g13930	✓	71	NRPS-PKS 2	*Similitude avec la lovastatine nonaketide synthase*	-
	Pc22g09030	✓	73	NRPS-PKS 3	*Similitude avec la 3-oxoacyl-[acyl-carrier-protein]-synthase*	-
PKS-Like	*Pc12g02940*	✓	1	PKS-Like 1	*Similitude avec la lovastatine dikétide synthase*	-
	Pc16g03760	✗	-	PKS-Like 2	-	-

Il apparaît que les peptides non-ribosomaux bénéficient d'une caractérisation nettement supérieure comparée aux autres groupes de métabolites. Plus surprenant, la quasi-totalité des séquences génomiques du **Tableau 3-12** sont présentes dans ***i*Prub22**. Dans un premier temps, nous avons alors examiné le nombre et la nature des réactions associées à ces séquences dans la reconstruction ***i*Prub22**, par le biais des associations GPR associées aux clusters de gènes listés dans le **Tableau 3-12**. La **Figure 3-32** résultante présente le nombre de réactions partagées entre différents BGCs pour les voies des NRPS et des PKS. Un nombre relativement important de réactions est commun à plusieurs voies, ce qui suggère que les mêmes réactions sont présentes dans chacune d'elles, un phénomène particulièrement marqué pour les polycétides. Ces associations GPR pourraient refléter une faible spécificité des enzymes impliquées. Cependant, il est plus probable que ces chevauchements résultent de faux positifs ou de réactions génériques d'activation des précurseurs. Cette observation soulève néanmoins des questions quant à la qualité de la reconstruction automatique du métabolisme spécialisé.

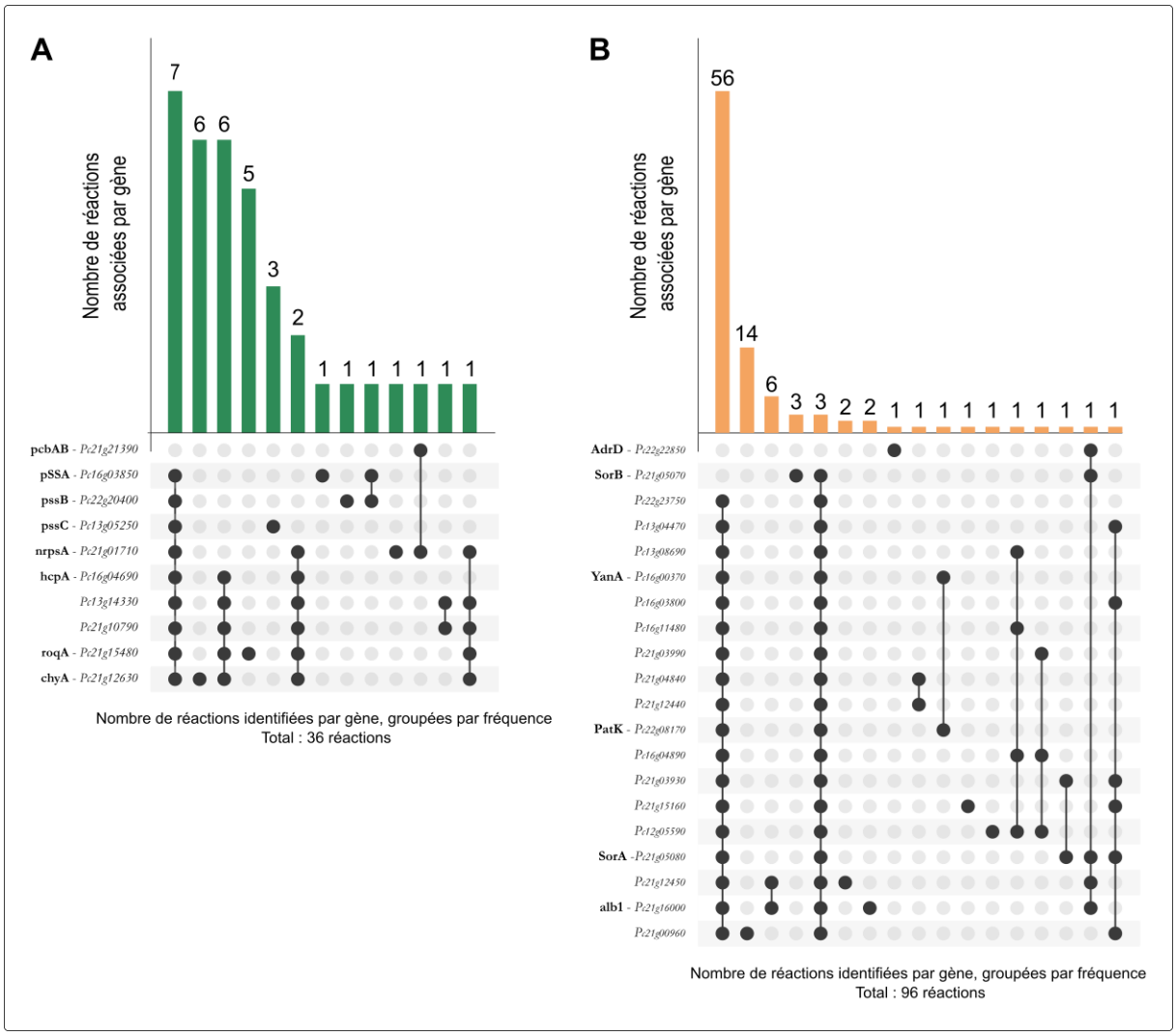


Figure 3-32 : Nombre de réactions associées aux gènes cœurs des BGCs de NRPS (A) et de PKS (B) au sein d'iPrub22.

Au cours de la section « *What about specialised metabolites?* » (page 175), nous avions expliqué que les listes de métabolites spécialisés à rechercher avaient été enrichies et consolidées à l'aide des outils de *genome mining* SMASH. Cependant, nous avions constaté une variation des résultats en fonction de l'utilisation de FungiSMASH ou AntiSMASH (**Tableau 3-13** et **Tableau 3-14**), tant pour l'identification des clusters de gènes biosynthétiques que pour leurs produits finaux. Il est également important de mentionner que les pourcentages de similarité indiquant la fiabilité de la présence du BGC dans la souche étudiée pouvaient différer entre AntiSMASH et FungiSMASH. En outre, ces outils évoluent rapidement, avec trois mises à jour majeures déployées entre novembre 2019 et novembre 2023, et les résultats obtenus entre les versions 5, 6 et 7 sont peu similaires. Les résultats exposés en **Tableau 3-13** et **Tableau 3-14** proviennent des versions 6.0.0 (juillet 2021) des outils.

Tableau 3-13 : Comparaison de l'identification des clusters de gènes biosynthétiques (BGC) entre AntiSMASH et FungiSMASH (version 6.0.0) sur la base du génome de *Penicillium rubens* Wisconsin 54-1255. Les nombres entre parenthèses correspondent aux produits naturels identifiés par chacun des outils et sont détaillés en **Tableau 3-14**.

Clusters de gènes biosynthétiques	AntiSMASH	FungiSMASH
Nombre de BGCs (composés identifiés)	45 (16)	42 (12)
Terpènes (composés identifiés)	11 (2)	5 (1)
NRPS (composés identifiés)	21 (3)	19 (2)
T1PKS (composés identifiés)	7 (6)	14 (6)
Hybrides T1PKS-NRPS (composés identifiés)	5 (5)	4 (3)
Sidérophore (composé identifié)	1 (0)	-

Tableau 3-14 : Détection des clusters de gènes de *Penicillium rubens* Wisconsin 54-1255 effectuée avec la suite d'outils SMASH (version 6.0.0). Les composés inscrits en gras sont ceux dont les voies de biosynthèse sont étudiées dans cette section.

	Clusters de gènes connus	Similarité détectée par antiSMASH	Similarité détectée par fungiSMASH	Identifiant MetaCyc
	Sidérophores			
	Pistillarine	Non détecté	Non détecté	-
	Coprogène	Non détecté	Non détecté	CPD0-2262
	Fusarinine	Non détecté	Non détecté	CPD-19128
●	**Ferrichrome**	Non détecté	Non détecté	CPD0-2241
	NRPS			
	Aspercryptines	26 %	Non détecté	-
	Nidulanine A	75 %	75 %	-
●	**Benzylpénicilline (pénicilline G)** **Isopénicilline N** **Phénoxyméthylpénicilline (pénicilline V)** **δ-(L-α-aminoadipyl)-L-cystéine-D-valine (ACV)**	87 %	62 %	PENICILLIN-G ISOPENICILLIN-N CPD-9196 N-5S-5-AMINO-5-CARBOXYPENTANOYL-L-CY
	Terpènes			
●	**PR-toxine**	100 %	Non détecté	CPD-13303
	Squalestatine S1	60 %	60 %	-
	T1PKS			
●	**Andrastine**	Non détecté	Non détecté	-
	Naphthopyrone	Non détecté	100 %	-

Tableau 3-14 (suite et fin)

	Chrysoxanthone A	Non détecté		-
	Chrysoxanthone B	Non détecté	47 %	-
	Chrysoxanthone C	Non détecté		-
	Dépudecine	Non détecté	33 %	-
	Neurosporine A	13 %	Non détecté	-
	ACT-Toxin II	100 %	Non détecté	-
●	**Patuline ***	46 %	33 %	CPD-16726
	***** Chez *P. rubens,* la voie de la patuline est incomplète, seules les réactions menant à la synthèse de l'isoepoxydon (*i.e.* n-2) sont conservées *(Van den Berg et al. 2008; Nielsen et al. 2017)*			
●	**Mélanine**	100 %	Non détecté	MELANIN
	Sorbicilline	100 %	57 %	-
●	**Yanuthone D**	100 %	80 %	-
	T1PKS, NRPS			
	Chrysogine	83 %	83 %	-
	Diméthylcoprogène	100 %	100 %	-
	Chaetoglobosines	28 %	Non détecté	-
	NG-391	33 %	Non détecté	-
	Déshydrohistidyl-tryptophanyl-dicétopipérazine (DHTD)			CPD-17379
	Glandicoline A			CPD-17389
	Glandicoline B			CPD-17390
●	**Histidyl-tryptophanyl-dicétopipérazine (HTD)**	100 %	100 %	CPD-17378
	Méleagrine			CPD-17391
	Roquefortine C			CPD-17381
	Roquefortine D			CPD-17380

Une fois cette liste de voies de biosynthèse et de métabolites d'intérêt établie, une seconde interrogation a émergé : est-ce que les voies métaboliques menant à la synthèse des composés sont connues et décrites dans la littérature ? Le cas échéant, ces voies sont-elles indexées dans MetaCyc ? Il en a résulté la création de deux listes, l'une contenant les métabolites pour lesquels il existe un identifiant dans MetaCyc, et l'autre, composée de métabolites orphelins pour lesquels il sera nécessaire, entre autres, de déterminer les plus proches précurseurs connus. La découverte de nouveaux produits naturels et l'accessibilité informatique de leurs données ne sont pas concomitantes. Ainsi, malgré les mises à jour récurrentes des ressources, des lacunes concernant les métabolites ciblés existent. Dans un premier temps, il nous a alors semblé plus judicieux de nous concentrer uniquement sur les précurseurs nécessaires à la biosynthèse de ces composés et identifiables dans la base de données MetaCyc.

Enfin, en fonction de l'existence des voies relatives à la biosynthèse des métabolites sélectionnés dans MetaCyc et de leur niveau de complétion, quels sont les divers traitements envisagés pour l'amélioration de la modélisation ? Si les voies métaboliques sont intégralement présentes, la curation consiste essentiellement à vérifier la présence ou l'absence effective des réactions dans la reconstruction, et, si les informations issues de la littérature le permettent, de vérifier les associations GPR correspondantes, car l'exactitude de ces dernières est primordiale pour l'interprétation biologique et informatique des « *knock-out* » *in silico*. Nous remarquerons toutefois que les cas particuliers sont nombreux, il est alors délicat d'implémenter actuellement un processus général et automatique pour l'intégration de ces données.

En revanche, si les voies métaboliques sont semi-complètes ou absentes, les analyses portent sur les plus proches précurseurs connus. Dans le cas des métabolites orphelins, il est nécessaire de créer à la fois les métabolites et les réactions soutenues par des séquences génomiques. Dans l'optique de garder une traçabilité optimale, l'ajout de ces entités doit être soutenu par le maximum de métadonnées (*e.g.* références bibliographiques pour la constitution des voies de biosynthèse, identifiants InChIKey, InChI, PubChem, masse moléculaire, charge, SMILES, *etc.*). Ce travail pourra être approfondi lorsque le moyen le plus fidèle possible d'intégrer et de générer les informations manquantes sera déterminé. Ces lacunes pourraient potentiellement être résolues par l'ajout de modules ou d'une réaction unique, équilibrée et soutenue par l'ensemble des gènes impliqués dans la biosynthèse des composés d'intérêt. Outre la pertinence biologique et la création d'un modèle au plus proche de la réalité, l'intérêt de recréer précisément ces voies réside dans la possibilité d'analyser ultérieurement, plus finement, la distribution et la réallocation des flux à partir des précurseurs de l'ensemble des métabolites ciblés.

Au vu des informations présentées ci-dessus, nous avons choisi de restreindre nos analyses à huit voies de biosynthèse différentes pour illustrer au mieux les capacités de production de *P. rubens* Wisconsin 54-1255 (mentionnées en gras dans le **Tableau 3-14**). Cette sélection se base sur des critères incluant la nature des molécules, la présence d'éléments orphelins, et la complétion des voies, afin de représenter au mieux le métabolisme spécialisé du champignon filamenteux.

Naturellement, la première voie de production de métabolites spécialisés étudiée est celle des pénicillines. Nous avons ensuite sélectionné les voies de la roquefortine/méléagrine clairement identifiés chez notre organisme. Cette liste a été enrichie par les produits ferrichrome et ferrichrome A issus de NRPS, afin d'inclure une voie de biosynthèse relative aux sidérophores.

En revanche, l'étude des PKS est plus complexe en raison du caractère partiel, voire sommaire, des voies de biosynthèse. Notre choix s'est alors porté sur l'exploitation de la voie de biosynthèse de l'isoepoxydon, dont les étapes initiales sont communes à la biosynthèse des yanuthones. Ensuite, bien que nous n'ayons pas trouvé de preuves probantes de la capacité de *P. rubens* Wisconsin 54-1255 à produire de l'avérufine ou de la versicolorine B, nous avons souhaité évaluer les flux de production de cette voie, qui a bénéficié de la complémentarité des approches de reconstruction pour sa complétion.

Nous complétons l'analyse des PKS en examinant les composés YWA1 et DMOA, précurseurs respectifs de la mélanine et des andrastines, et nous concluons notre sélection avec l'aristolochène, précurseur de la PR-toxine, seul terpène inclus dans notre étude.

Enfin, avant d'approfondir la composition de ces voies de biosynthèse dans la section suivante, nous présentons en **Figure 3-33** un schéma synthétique des voies de biosynthèse des métabolites spécialisés étudiés. Ce schéma inclut les peptides non ribosomaux, les polycétides, les terpènes et les molécules hybrides, offrant ainsi une visualisation claire de leur spécificité biosynthétique.

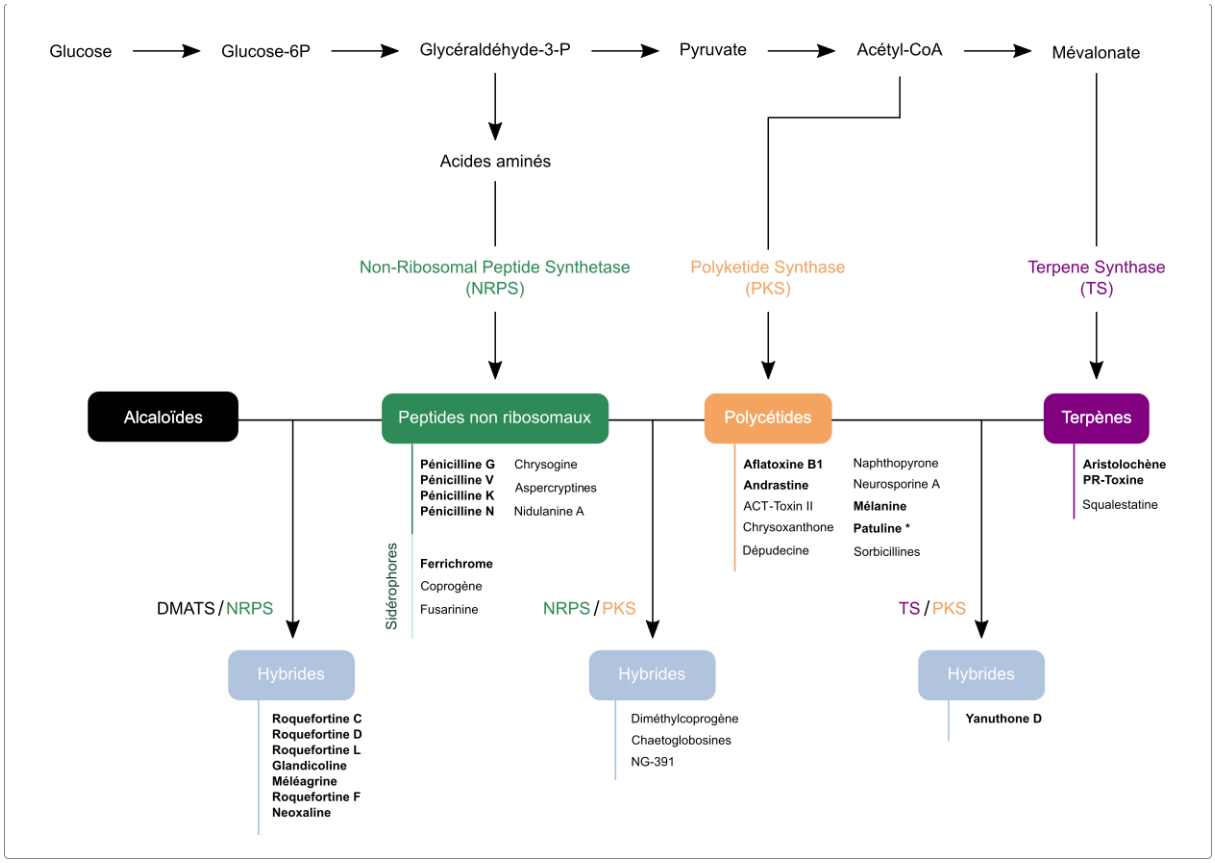


Figure 3-33 : Résumé synthétique des voies de biosynthèse des métabolites spécialisés produits, ou potentiellement produits, par Penicillium rubens Wisconsin 54-1255. *Les composés inscrits en gras sont ceux dont les voies de biosynthèse sont étudiées dans cette section. Inspiré et adapté de (El Hajj Assaf et al. 2020).*

3.2. Reconstruire les voies de biosynthèse des métabolites spécialisés

L'intégration du métabolisme spécialisé dans la reconstruction d'***i*Prub22** a été une tâche complexe, que nous pouvons résumer en trois points clés. Premièrement, en raison du caractère spécifique des informations recherchées et des approches de reconstruction employées, les données relatives au métabolisme spécialisé ne peuvent être obtenues que par l'annotation fonctionnelle ou par l'intégration de données provenant de réseaux récents. La modélisation de ce métabolisme atteint ainsi une des limites de l'automatisation des reconstructions, nécessitant une curation manuelle intensive au cas par cas pour intégrer correctement les métabolites souhaités. Deuxièmement, l'incomplétude des bases de données pose des questions sur le traitement des métabolites orphelins (*i.e.* métabolites et/ou réactions non référencés dans les bases de données). Cela représente un défi majeur pour la reconstruction précise et exhaustive des voies métaboliques. Troisièmement, en raison de phénomènes de catalyses complexes ou de connaissances lacunaires concernant ces voies, de nombreuses réactions non équilibrées y sont recensées. Lors de la recherche de fonctionnalité d'***i*Prub22**, ces réactions ont été initialement fermées et ont dû être réouvertes pour assurer la production des métabolites spécialisés. Ainsi, afin de couvrir les différents cas de figure et aspects cités précédemment, nous nous sommes appuyés sur les huit voies de biosynthèse suivantes :

- Voies de biosynthèse des **pénicillines** (**Figure 3-34**)

Les pénicillines sont des antibiotiques de la famille des *β*-lactames qui inhibent la synthèse de la paroi cellulaire des bactéries, offrant à *P. rubens* un avantage compétitif en limitant la croissance bactérienne environnante *(Fleming 1929; Chain et al. 1940)*. Leur biosynthèse commence par la condensation de trois acides aminés : le L-*a*-aminoadipate, la L-cystéine et la L-valine. Le tripeptide résultant, l'acide *δ*-(L-*a*-aminoadipoyl)-L-cysteinyl-D-valine (ACV), est ensuite cyclisé pour former l'isopénicilline N, le point de branchement vers la biosynthèse des diverses formes de pénicillines : G, K et N. Comme mentionné précédemment, cette voie est médiée par l'action de quatre enzymes différentes : ACVS (*Pc21g21390*), IPNS (*Pc21g21380*), Phl (*Pc21g14900*) et PenDE (*Pc21g21370*) *(Guzmán-Chávez et al. 2018; García-Estrada et al. 2020)*. Le pathway présenté en **Figure 3-34** comprend 8 réactions dont seule l'interconversion de la pénicilline G en pénicilline V, médiée par PenDE, manquait lors de la génération du draft.

- Voies de biosynthèse des **roquefortines** (**Figure 3-35**)

Les roquefortines et la méléagrine sont des alcaloïdes indoliques produits par diverses espèces de *Penicillium* et d'*Aspergillus*. Possédant des propriétés antibactériennes et antifongiques, ces mycotoxines contribuent à la protection contre les micro-organismes. La voie de biosynthèse de ces composés, illustrée en **Figure 3-35**, commence avec les acides aminés tryptophane et histidine et inclut 13 réactions, dont 10 non équilibrées, notamment en raison de l'intervention de l'enzyme NADPH-hémoprotéine réductase dans les réactions d'oxydation et de réduction. Initialement absente lors de la génération du draft (cf. section *1.4.1 The strengths and limitations of reconstruction automation*, page 159), cette voie a été intégrée manuellement à ***i*Prub22** en associant les réactions concernées aux séquences génomiques des gènes roqA (*Pc21g15480*), roqD (*Pc21g15430*), roqR (*Pc21g15470*), roqM (*Pc21g15460*), roqO (*Pc21g15450*) et roqN (*Pc21g15440*) *(Ali et al. 2013; Martín et Liras 2017; Guzmán-Chávez et al. 2018)*.

- Voies de biosynthèse des **ferrichromes** (**Figure 3-36**)

Le ferrichrome et le ferrichrome A sont tous deux des sidérophores, des molécules spécialisées dans la chélation et dans le transport du fer chez les micro-organismes. Leurs différences structurelles influencent la spécificité d'interaction avec les récepteurs et l'efficacité du transport du fer. Leur biosynthèse, étudiée et élucidée notamment chez *Aspergillus fumigatus* *(Haas 2014)*, commence par l'assemblage de précurseurs acétyl-CoA et de dérivés d'ornithine par des enzymes NRPS. Les peptides formés subissent ensuite une cyclisation et des modifications supplémentaires, telles que l'acétylation et l'hydroxylation, pour produire des ferrichromes fonctionnels capables de se lier au fer avec une grande affinité. La voie de biosynthèse présentée en **Figure 3-36**, comprend 7 réactions correspondant à l'intégralité des pathways MetaCyc `PWY-7571` (« *ferrichrome A biosynthesis* » – 5 réactions) et `PWY-7577` (« *ferrichrome biosynthesis* » – 2 réactions). Les réactions menant à la biosynthèse du ferrichrome sont, quant-à-elles, majoritairement catalysées par les produits géniques homologues à ceux identifiés chez *A. fumigatus* : SidA (*Pc13g05260*), SidL (*Pc22g20380*, *Pc16g03860*) et Pssc (*Pc13g05250*) *(Van den Berg et al. 2008; Samol et al. 2015)*. En revanche, nous noterons que roqA, dont le produit génique est associé à la voie des roquefortines, est la seule justification de la présence de la réaction `RXN-15951`, responsable de la production de ferrichrome A.

- Voies de biosynthèse de l'**isoepoxydon** et des **yanuthones** (**Figure 3-37**)

L'isoepoxydon et les yanuthones sont des polycétides qui partagent un précurseur commun, le 3-méthylphénol. L'isoepoxydon est une mycotoxine aidant à la défense et à la compétition pour les ressources, tandis que les yanuthones, considérées comme des antibiotiques, possèdent des activités biologiques diverses souvent associées à l'inhibition de la croissance des compétiteurs microbiens. La biosynthèse de ces composés débute par l'assemblage de précurseurs acétyl-CoA *via* des enzymes PKS et s'effectue en deux réactions dont la première est catalysée par les enzymes YanA (*Pc16g00370*) et PatK (*Pc22g08170*).

Chez *P. rubens*, la voie de la patuline est incomplète et semble s'effectuer jusqu'à la synthèse de l'isoepoxydon *(Van den Berg et al. 2008 ; Nielsen et al. 2017)*. Les intermédiaires formés subissent ensuite des modifications enzymatiques telles que l'époxydation, la cyclisation et l'oxydation pour aboutir à l'isoepoxydon en quatre réactions. Cette voie fait alors intervenir les enzymes PatH (*Pc22g08110*), PatI (*Pc22g08150*) et PatJ (*Pc22g08160*). Ce pathway a bénéficié d'une complétion manuelle par l'ajout des deux réactions finales associées à leurs gènes correspondants en accord avec la littérature *(Holm et al. 2014 ; Nielsen et al. 2017 ; Guzmán-Chávez et al. 2018)*.

Les yanuthones, qui sont des diterpènes, sont formées, quant à elles, à partir de précurseurs isoprénoïdes par une voie distincte impliquant des enzymes terpènes synthases (TS) et diverses modifications post-synthétiques. En revanche, même si l'ensemble des gènes menant à la production des yanuthones est caractérisé *(Holm et al. 2014 ; Guzmán-Chávez et al. 2018)*, ces métabolites sont absents de MetaCyc. Le plus proche précurseur connu et intégré dans MetaCyc est le toluquinol, un métabolite qui ne possède aucune réaction de consommation sur cette base de données et qui nous servira d'ancrage de référence pour l'étude des flux de production de cette voie.

- Voies de biosynthèse des **andrastines** (**Figure 3-38**)

L'andrastine A est un potentiel agent antitumoral dont la biosynthèse commence par l'assemblage de précurseurs acétyl-CoA et malonyl-CoA par une PKS. Les étapes subséquentes incluent des cyclisations et des oxydations catalysées par neuf enzymes spécifiques pour former le noyau tétracyclique de l'andrastine. Des modifications additionnelles, telles que l'hydroxylation et la méthylation, sont nécessaires pour produire les formes actives de l'andrastine. Bien que son cluster de gènes soit identifié *(Matsuda et al. 2014; Martín et Liras 2017 ; Guzmán-Chávez et al. 2018)*, la voie de biosynthèse des andrastines n'est pas encore référencée sous MetaCyc**,** nous nous concentrerons donc sur le plus proche antécédent répertorié, le 3,5-diméthylorsellinate (DMOA).

L'analyse de la topologie d'***i*Prub22** indique l'existence d'une autre voie de biosynthèse de méroterpénoïdes à partir du DMOA, celle de l'anditomine *(Matsuda et al. 2014)*. Sous MetaCyc, cette voie de biosynthèse est référencée sous l'identifiant `PWY-7599` et comporte 15 réactions. Au sein d'***i*Prub22**, nous recensons 14 de ces réactions, dont 8 présentes dans le draft *via* l'apport des sources externes. Les 6 autres ont été apportées par *gap-filling*, et l'une d'entre elles appartient au groupe des réactions spontanées. Le cluster de gènes de ce métabolite spécialisé présente des protéines homologues à celles du cluster des andrastines et, nous retrouvons en effet les gènes adrD, adrI, adrG, adrH et adrF associés à cinq des réactions de biosynthèse de l'anditomine.

- Voie de biosynthèse de la **mélanine** (**Figure 3-39**)

La mélanine est un pigment bleu-vert protègeant contre les stress environnementaux tels que les UV, les radicaux libres et les fluctuations de température. Chez les champignons filamenteux, elle est également responsable de l'épaississement de la paroi cellulaire augmentant la résistance des organismes aux conditions défavorables. La biosynthèse de ce polycétide comprend l'agrégation d'acétyl-CoA et de malonyl-CoA, aboutissant à la formation d'YWA1, un intermédiaire dans la biosynthèse de divers métabolites secondaires tels que l'aurofusarine et les naphtopyrones. Les chaînes polycétidiques subissent ensuite des modifications, telles que la cyclisation, l'oxydation et la polymérisation, pour former un polymère hétérogène, la mélanine *(Guzmán-Chávez et al. 2018)*. Sous MetaCyc, la voie de biosynthèse de ce métabolite est partielle et morcelée comme représenté en **Figure 3-39**. Nous concentrons donc notre étude de production sur le composé YWA1.

En explorant la topologie d'***i*Prub22**, nous remarquons également la présence de deux autres voies de biosynthèse de polycétides qui se branchent sur les précurseurs de la mélanine : celle de la citreoisocoumarine et celle de l'aurofusarine. La citreoisocoumarine, à l'instar de nombreux métabolites spécialisés, peut jouer un rôle dans la communication chimique et dans la défense et la compétition microbienne, tandis que l'aurofusarine, un pigment de couleur rouge foncé à orange aux effets toxiques, protège les champignons filamenteux contre les stress environnementaux. Cependant, les clusters de gènes de ces voies n'ont pas été détectés par les outils de *genome mining* Fungi et AntiSMASH. De surcroît, la présence de ces voies dans ***i*Prub22** possède un poids de confiance relativement faible. En effet, sous

MetaCyc, le pathway PWY-7696 (« *citreoisocoumarin and bikisocoumarin biosynthesis* ») comprend 4 réactions, certes intégralement retrouvées dans ***i*Prub22**, mais dont deux proviennent des étapes du *gap-filling* tandis que les deux autres sont soutenues par la même séquence génomique, le gène alb1 dont l'enzyme résultante est impliquée dans la première étape de la voie de production d'YWA1. À ce titre, nous remarquerons que dans ***i*Prub22** cette séquence génomique n'est pas, comme attendu, associée aux premières réactions présentées en **Figure 3-39.** Concernant la voie de biosynthèse de l'aurofusarine, le pathway PWY-7695 (« *aurofusarin biosynthesis* ») comprend un total de 12 réactions intégralement retrouvées dans ***i*Prub22**. Parmi celles-ci, 8 sont présentes dès la génération du draft et impliquent 4 séquences génomiques différentes, dont l'une, abr2, est associée au BGC de la mélanine. Ainsi, la capacité potentielle de production de citreoisocoumarine et d'aurofusarine par *P. rubens Wisconsin 54-1255* doit donc être traitée avec prudence.

- Voies de biosynthèse de l'**avérufine** et de la **versicolorine B** (**Figure 3-40**)

L'avérufine et la versicolorine B sont des précurseurs dans la biosynthèse des mycotoxines aflatoxines. Ces intermédiaires possèdent également des propriétés antifongiques et antibiotiques, jouant un rôle dans la défense contre d'autres micro-organismes *(Yabe et Nakajima 2004)*. Bien que les clusters de gènes relatifs à ces polycétides n'aient pas été détectés avec la suite d'outils SMASH, nous avons choisi de présenter ces voies de biosynthèse pour illustrer les conséquences des étapes de *gap-filling* liées au métabolisme spécialisé, puisque l'avérufine et la versicolorine appartiennent à la liste des Targets 1 *(Maskey et al. 2003)*. La **Figure 3-40** comprend 21 réactions et représente les pathways suivants : PWY-5954 (« *(1'S,5'S)-averufin biosynthesis* » - 8 réactions sur 8), PWY-5955 (« *versicolorin B biosynthesis* » - 9 réactions sur 9), PWY-5956 (« *sterigmatocystin biosynthesis* » - 2 réactions sur 3) et PWY-5959 (« *aflatoxins B1 and G1 biosynthesis* » - 2 réactions sur 3). Comparées aux voies présentées précédemment, les origines des réactions figurant dans la **Figure 3-40** sont plus variées : 7 réactions proviennent de séquences génomiques issues de l'annotation fonctionnelle, 7 de Prubens, 2 de PchCyc et 5 sont obtenues par *gap-filling*. Nous notons toutefois que l'aflatoxine ne peut être produite par ***i*Prub22** en raison de l'absence de la réaction RXN-9496, qui permet la transformation de la versicolorine B en 6-demethylsterigmatocystine. Cette réaction n'a pu être retrouvée par *gap-filling* car le composé 6-DEMETHYLSTERIGMATOCYSTIN a été classé à tort comme composé d'initiation (cf. encart : *Les risques de la curation manuelle - Des intrus se sont glissés dans la liste des composés d'initiation*, page 245).

- Voie de biosynthèse de la **PR-toxine** (**Figure 3-41**)

La PR-toxine est une mycotoxine dont la biosynthè se implique des enzymes PKS qui assemblent des unités d'acétyl-CoA en une chaîne polycétidique. Cette chaîne est ensuite modifiée par des enzymes de cyclisation, d'oxydation et de réduction pour former la structure époxydée caractéristique de la PR-toxine. Les modifications post-synthétiques incluent l'ajout de groupes fonctionnels spécifiques qui confèrent la toxicité et les propriétés biologiques de la toxine *(Hidalgo et al. 2014 ; Martín et Liras 2017)*. La PR-toxine est un composé orphelin indexé dans MetaCyc, pour lequel aucune réaction n'est associée. En revanche, le chemin menant à la synthèse de son précurseur, l'aristolochène, est présent dans le draft et servira de métabolite de référence pour l'étude de cette voie.

Afin de visualiser ces divers éléments, nous présentons ci-contre les caractéristiques liées à la reconstruction des voies de biosynthèse des métabolites spécialisés telles que référencées dans ***i*Prub22**. Dans ces figures, les noms des métabolites sont écrits en rouge et leurs représentations des structures des molécules proviennent de MetaCyc. Les identifiants de réactions sont indiqués en noir, et les cofacteurs ou enzymes et éléments annexes sont inscrits en vert.Graphiquement, les métabolites dont les flux de production peuvent être monitorés sont associés à des « usines de production », les « panneaux voies sans issue » indiquent les *dead-ends*, les « murs de briques en construction » correspondent à des éléments qui, au moment de la génération d'***i*Prub22**, n'étaient pas référencés dans la base de données MetaCyc, les livres représentent des éléments soutenus par la littérature scientifique (*i.e.* généralement associés à des séquences génomiques au sein des GPR), enfin, les « panneaux noirs *closed* » indiquent des réactions qui ont été fermées initialement mais qui ont dû être réouvertes pour assurer un flux de production.

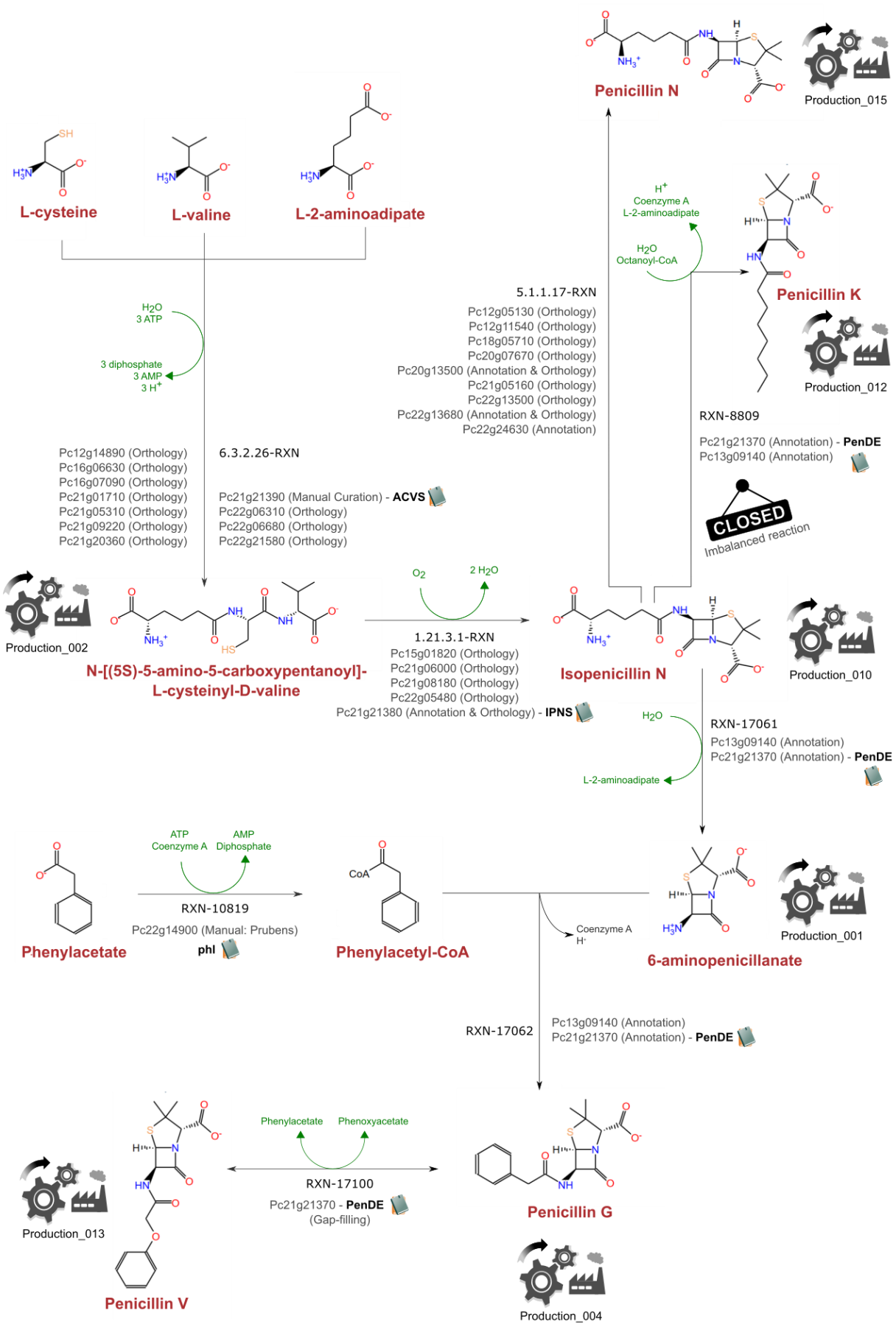
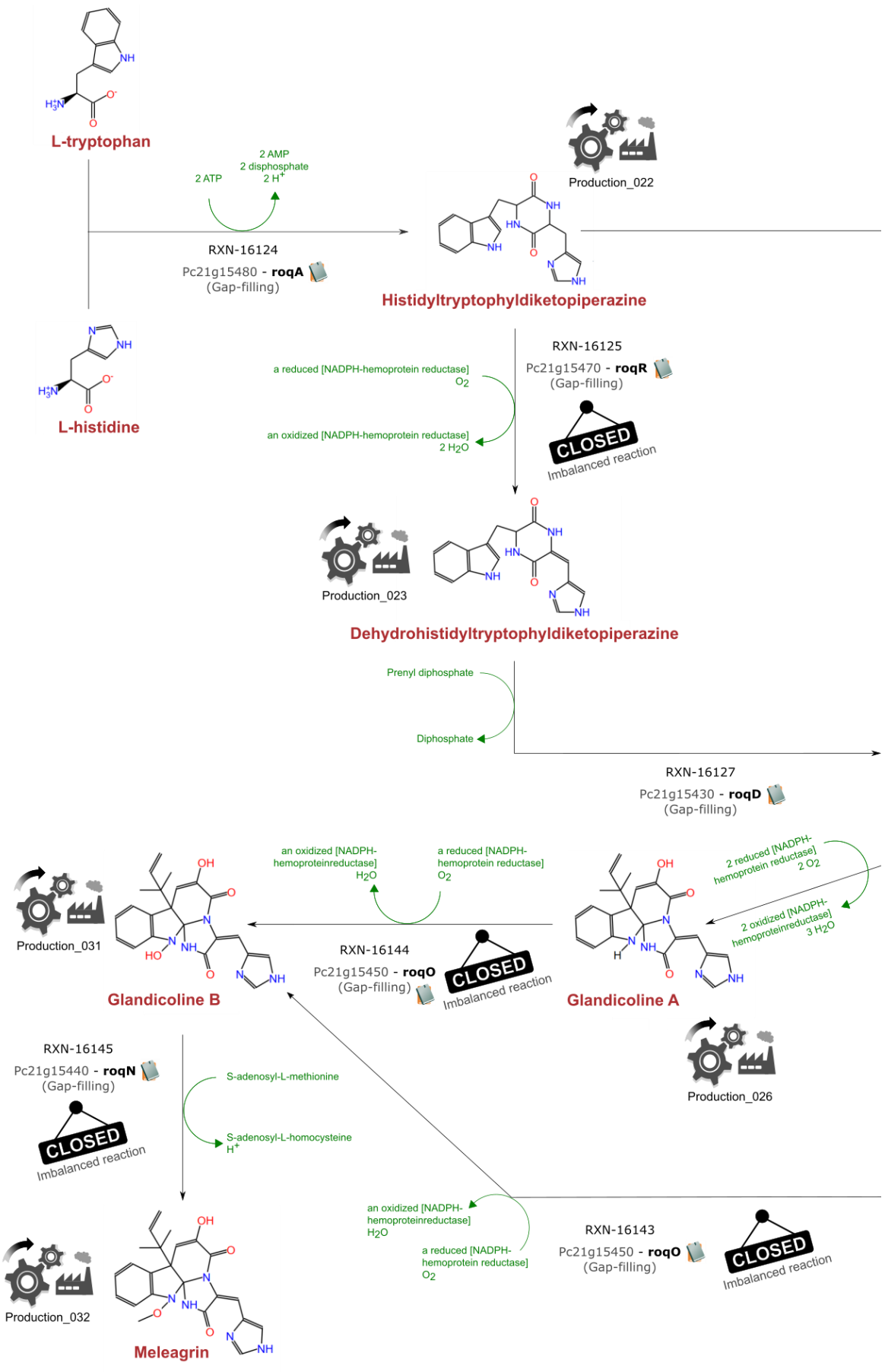


Figure 3-34 : Voie de biosynthèse des pénicillines.


L-tryptophan
2 ATP
2 AMP
2 disphosphate
2 H+
RXN-16124
Pc21g15480 - roqA
(Gap-filling)
Production_022
Histidyltryptophyldiketopiperazine
L-histidine
RXN-16125
Pc21g15470 - roqR
(Gap-filling)
a reduced [NADPH-hemoprotein reductase]
O2
an oxidized [NADPH-hemoprotein reductase]
2 H2O
CLOSED
Imbalanced reaction
Production_023
Dehydrohistidyltryptophyldiketopiperazine
Prenyl diphosphate
Diphosphate
RXN-16127
Pc21g15430 - roqD
(Gap-filling)
Production_031
Glandicoline B
an oxidized [NADPH-hemoproteinreductase]
H2O
a reduced [NADPH-hemoprotein reductase]
O2
RXN-16144
Pc21g15450 - roqO
(Gap-filling)
CLOSED
Imbalanced reaction
Glandicoline A
2 reduced [NADPH-hemoprotein reductase]
2 O2
2 oxidized [NADPH-hemoproteinreductase]
3 H2O
Production_026
RXN-16145
Pc21g15440 - roqN
(Gap-filling)
CLOSED
Imbalanced reaction
S-adenosyl-L-methionine
S-adenosyl-L-homocysteine
H+
Production_032
Meleagrin
an oxidized [NADPH-hemoproteinreductase]
H2O
a reduced [NADPH-hemoprotein reductase]
O2
RXN-16143
Pc21g15450 - roqO
(Gap-filling)
CLOSED
Imbalanced reaction

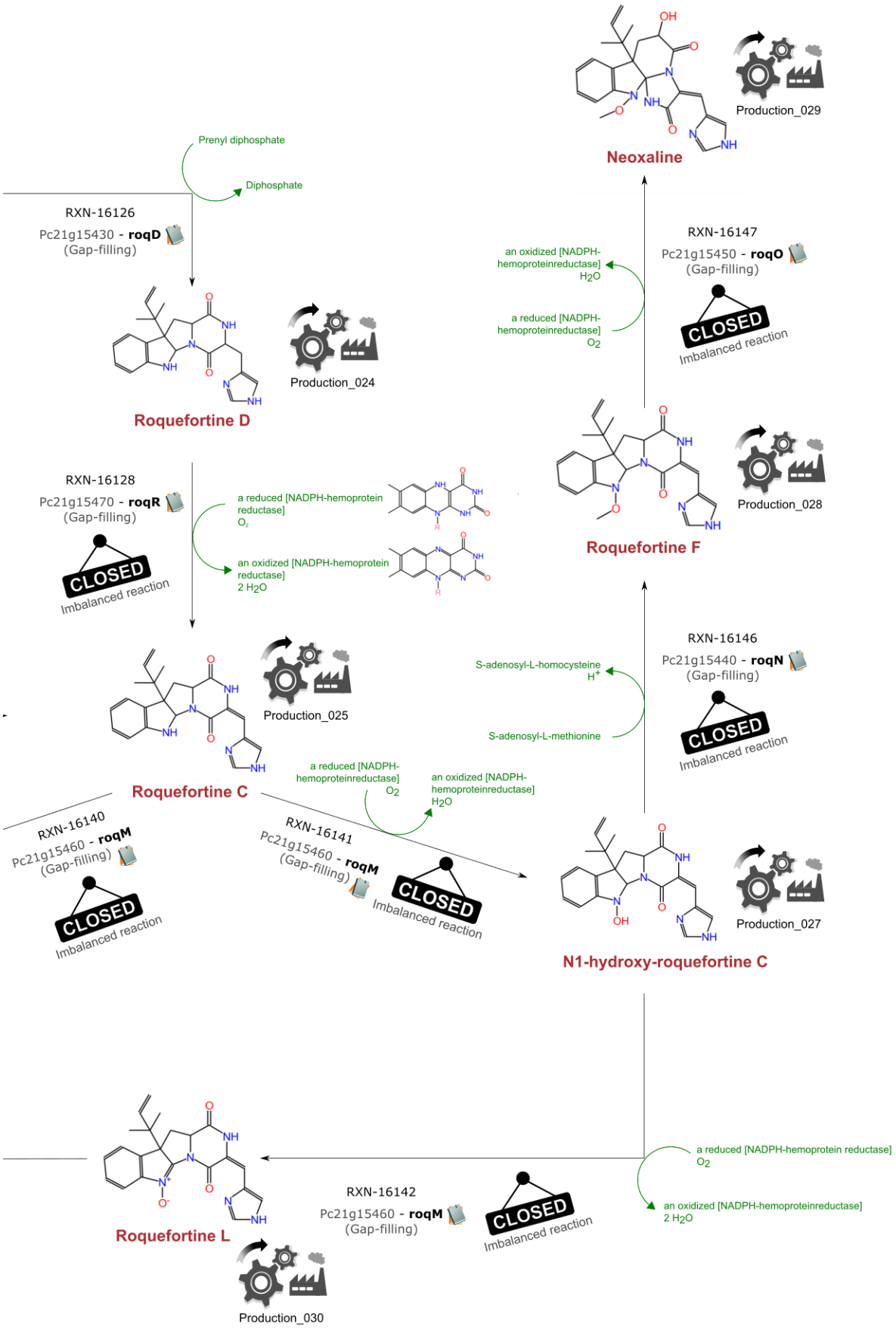
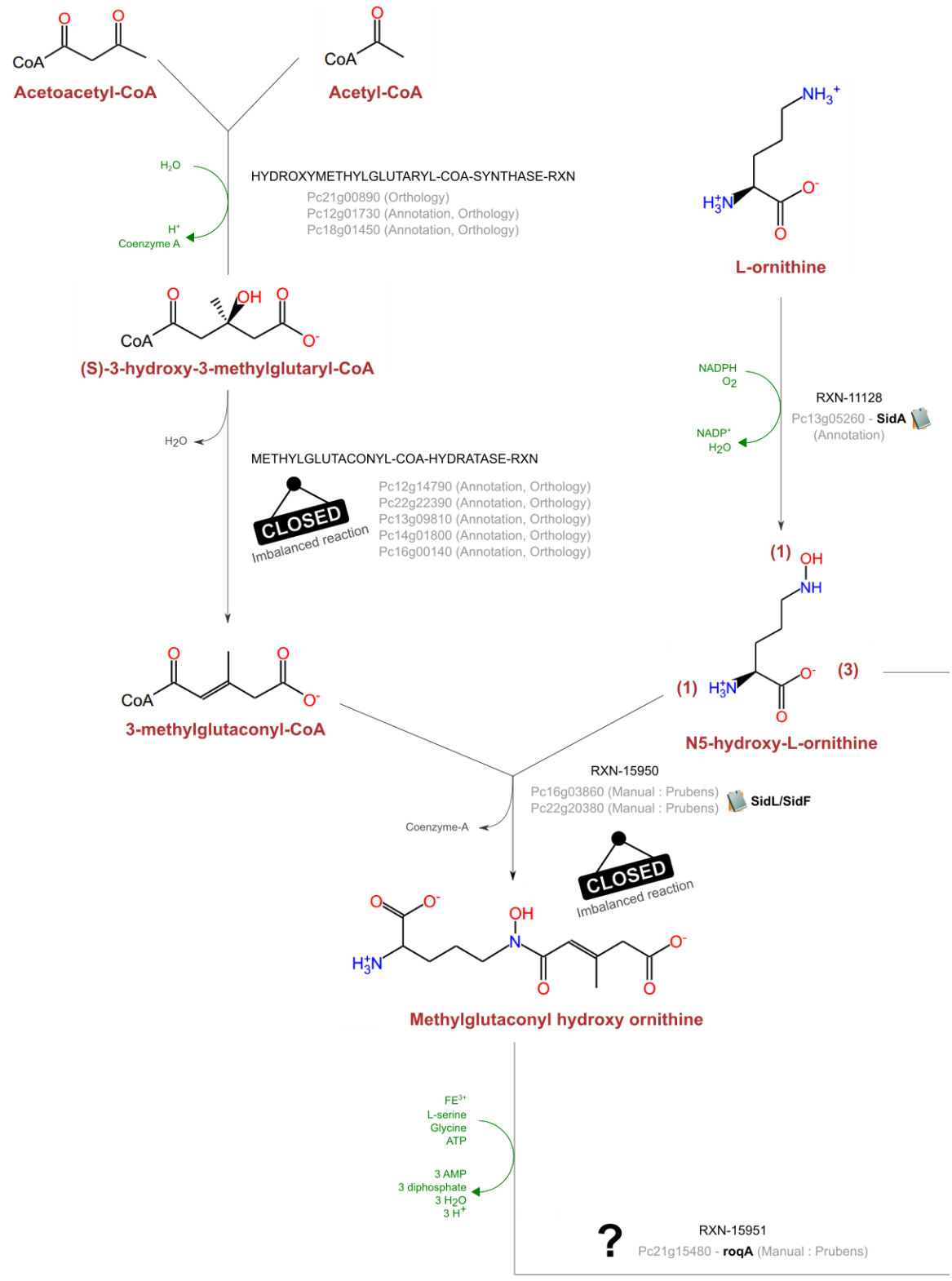


Figure 3-35 : Voies de biosynthèse des roquefortines.


Acetoacetyl-CoA
Acetyl-CoA
H2O
H+
Coenzyme A
HYDROXYMETHYLGLUTARYL-COA-SYNTHASE-RXN
Pc21g00890 (Orthology)
Pc12g01730 (Annotation, Orthology)
Pc18g01450 (Annotation, Orthology)
L-ornithine
(S)-3-hydroxy-3-methylglutaryl-CoA
NADPH
O2
NADP+
H2O
RXN-11128
Pc13g05260 - SidA
(Annotation)
H2O
METHYLGLUTACONYL-COA-HYDRATASE-RXN
CLOSED
Imbalanced reaction
Pc12g14790 (Annotation, Orthology)
Pc22g22390 (Annotation, Orthology)
Pc13g09810 (Annotation, Orthology)
Pc14g01800 (Annotation, Orthology)
Pc16g00140 (Annotation, Orthology)
(1)
(3)
3-methylglutaconyl-CoA
N5-hydroxy-L-ornithine
RXN-15950
Pc16g03860 (Manual : Prubens)
Pc22g20380 (Manual : Prubens)
SidL/SidF
Coenzyme-A
CLOSED
Imbalanced reaction
Methylglutaconyl hydroxy ornithine
FE3+
L-serine
Glycine
ATP
3 AMP
3 diphosphate
3 H2O
3 H+
?
RXN-15951
Pc21g15480 - roqA (Manual : Prubens)

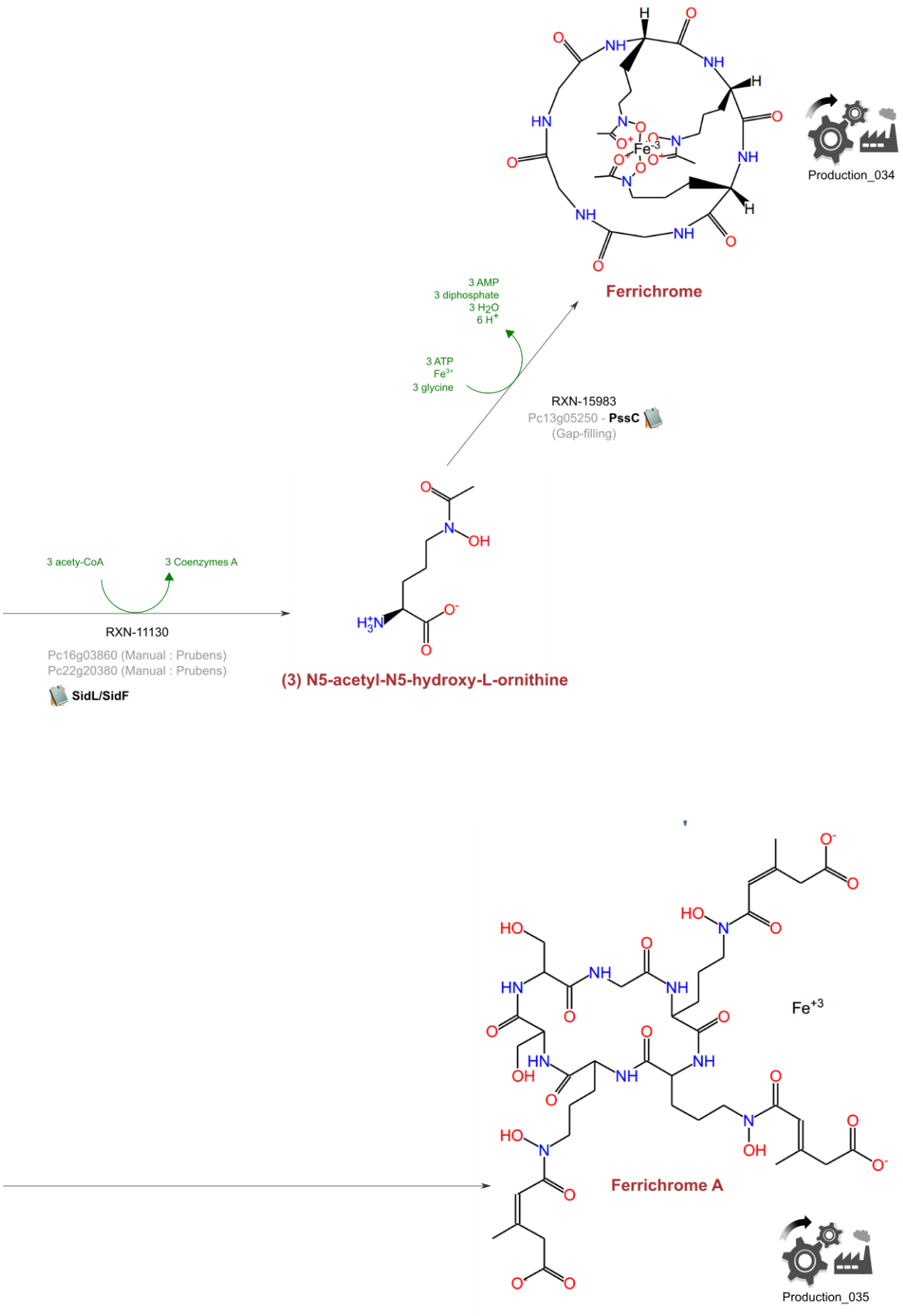


Figure 3-36 : Voies de biosynthèse des ferrichromes.

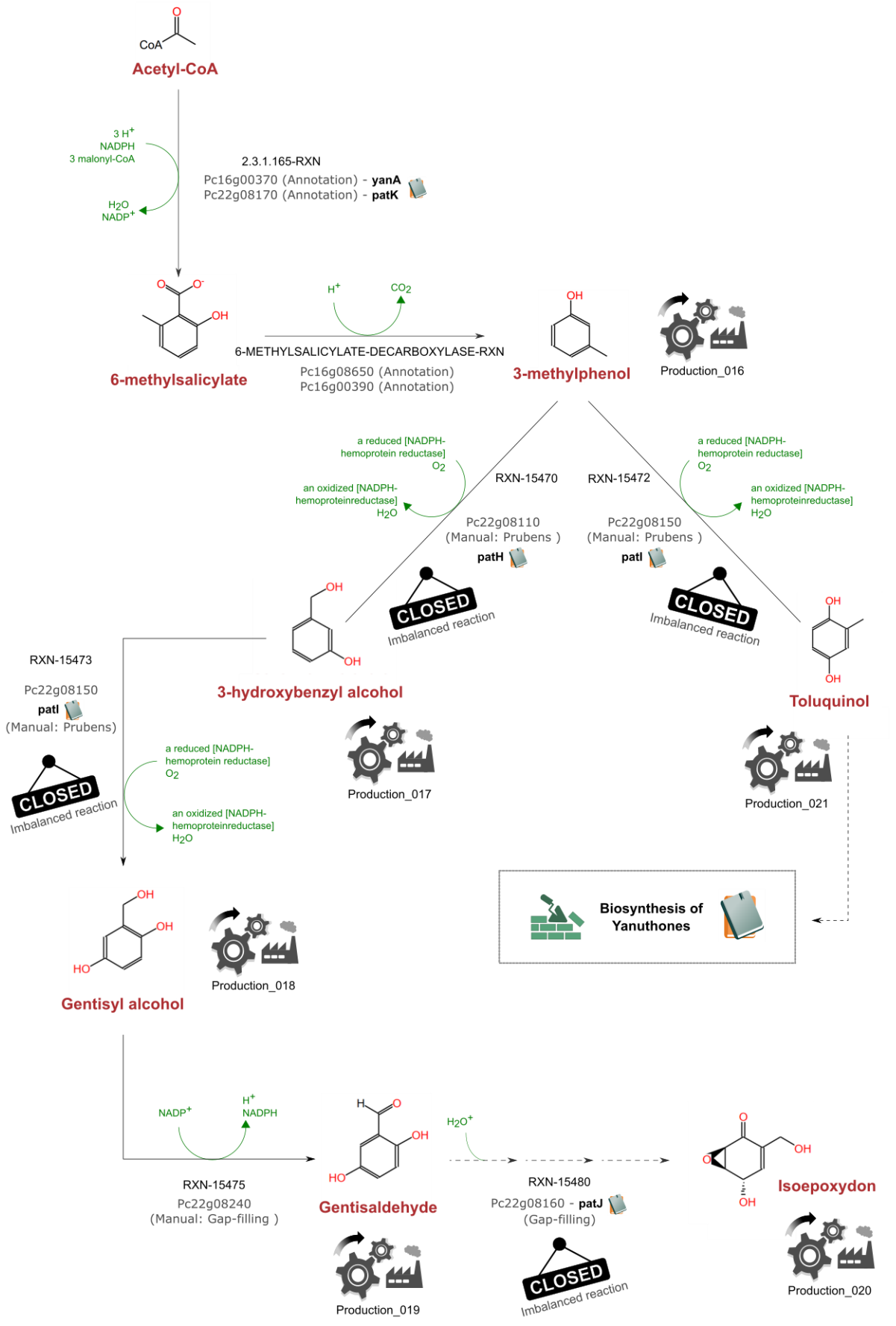


Figure 3-37 : Voies de biosynthèse de l'isoepoxydon et des yanuthones.

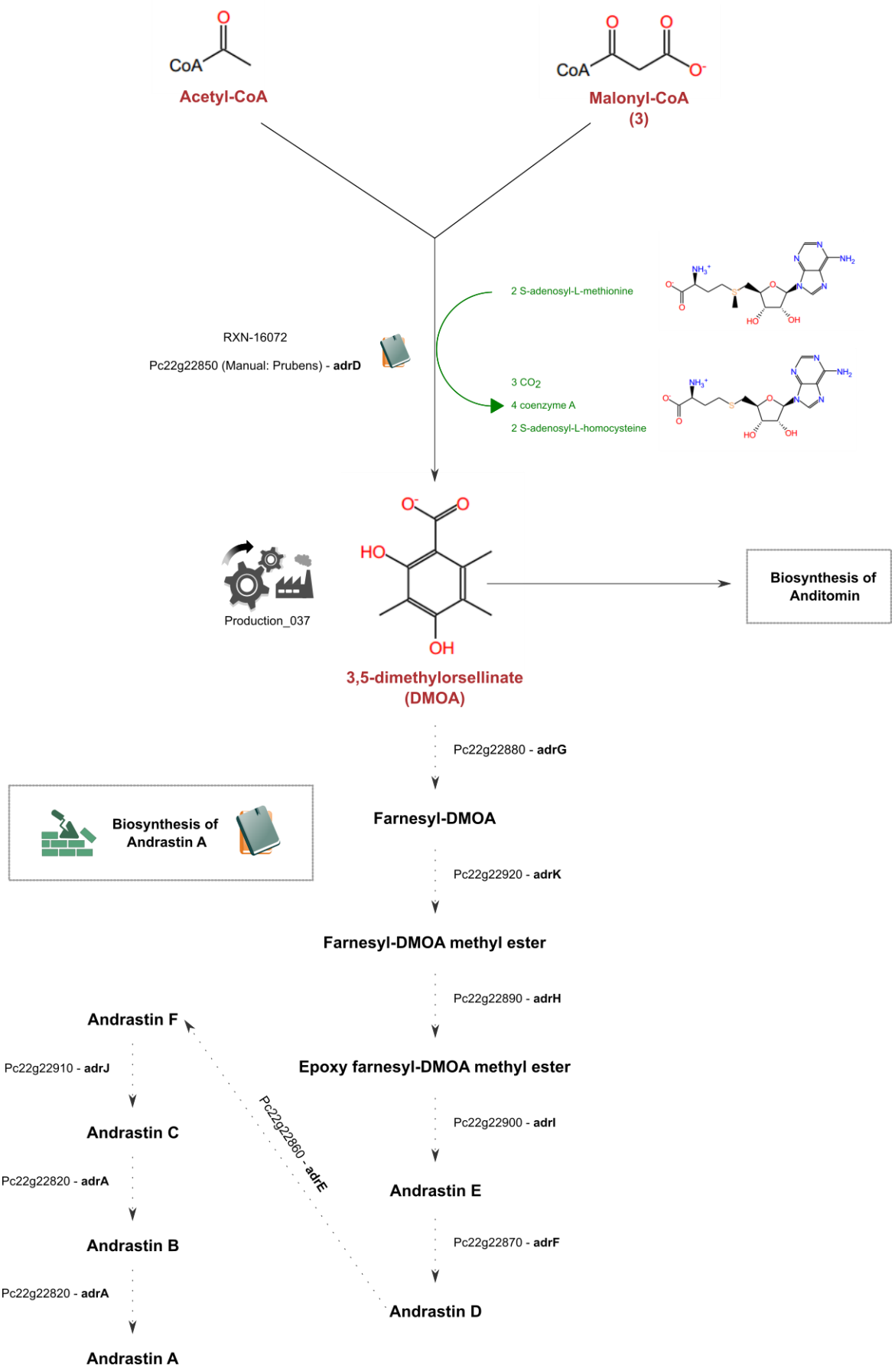
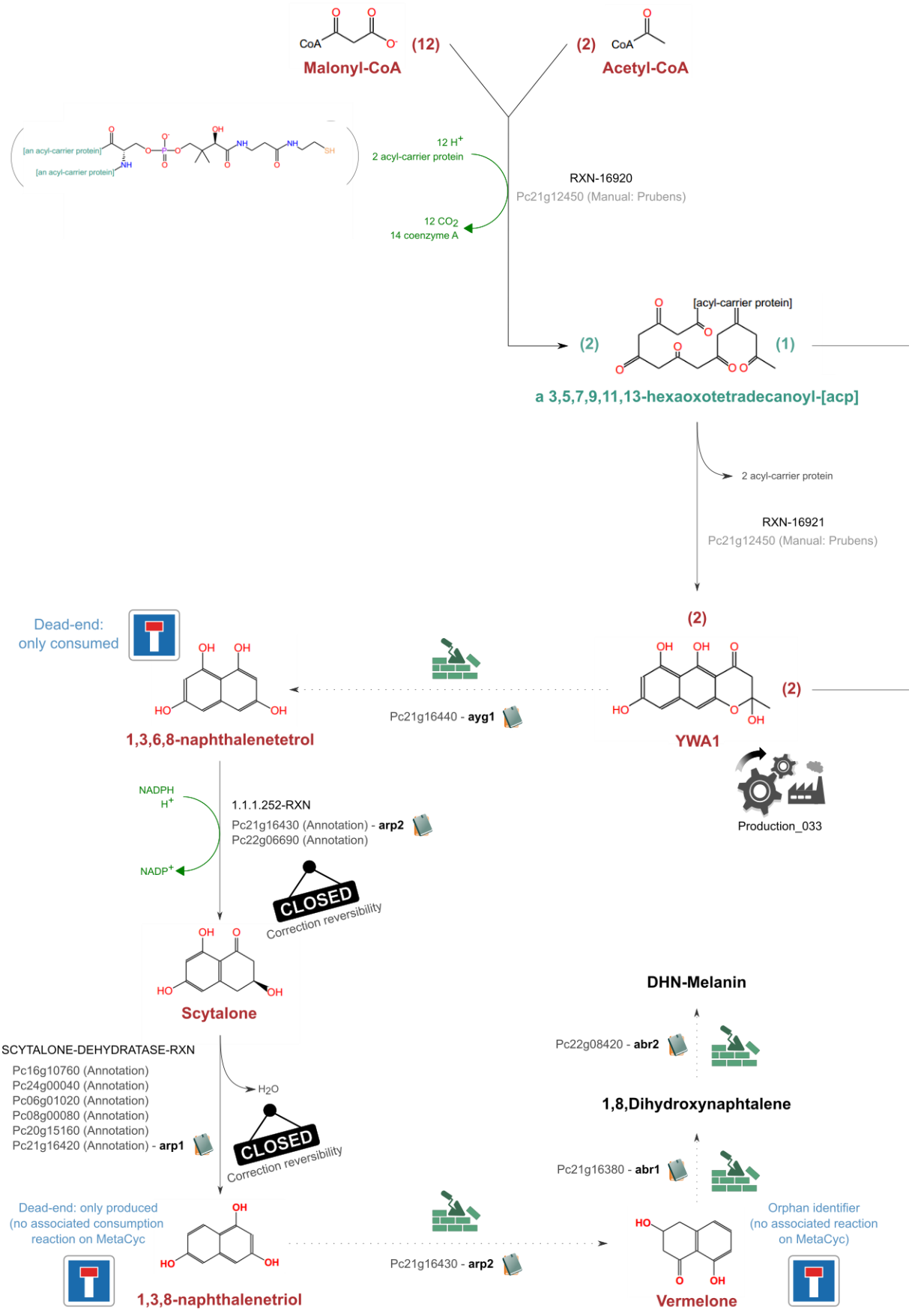


Figure 3-38 : Voie de biosynthèse des andrastines.


Malonyl-CoA
(12)
(2)
Acetyl-CoA
12 H+
2 acyl-carrier protein
12 CO2
14 coenzyme A
RXN-16920
Pc21g12450 (Manual: Prubens)
[acyl-carrier protein]
(2)
(1)
a 3,5,7,9,11,13-hexaoxotetradecanoyl-[acp]
2 acyl-carrier protein
RXN-16921
Pc21g12450 (Manual: Prubens)
(2)
YWA1
(2)
Production_033
Pc21g16440 - ayg1
Dead-end:
only consumed
1,3,6,8-naphthalenetetrol
NADPH
H+
NADP+
1.1.1.252-RXN
Pc21g16430 (Annotation) - arp2
Pc22g06690 (Annotation)
CLOSED
Correction reversibility
Scytalone
SCYTALONE-DEHYDRATASE-RXN
Pc16g10760 (Annotation)
Pc24g00040 (Annotation)
Pc06g01020 (Annotation)
Pc08g00080 (Annotation)
Pc20g15160 (Annotation)
Pc21g16420 (Annotation) - arp1
H2O
CLOSED
Correction reversibility
Dead-end: only produced
(no associated consumption
reaction on MetaCyc
1,3,8-naphthalenetriol
Pc21g16430 - arp2
DHN-Melanin
Pc22g08420 - abr2
1,8,Dihydroxynaphtalene
Pc21g16380 - abr1
Orphan identifier
(no associated reaction
on MetaCyc)
Vermelone

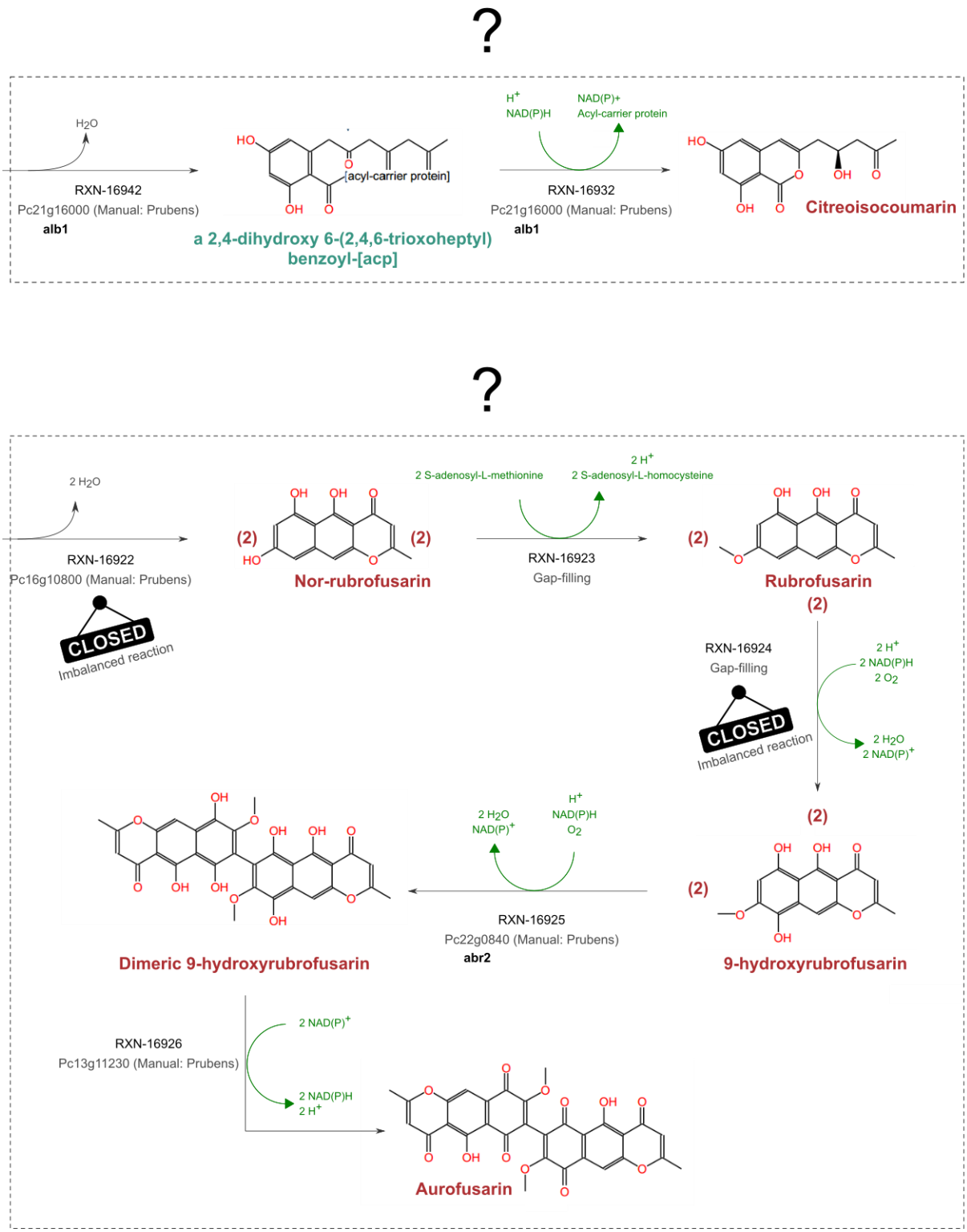
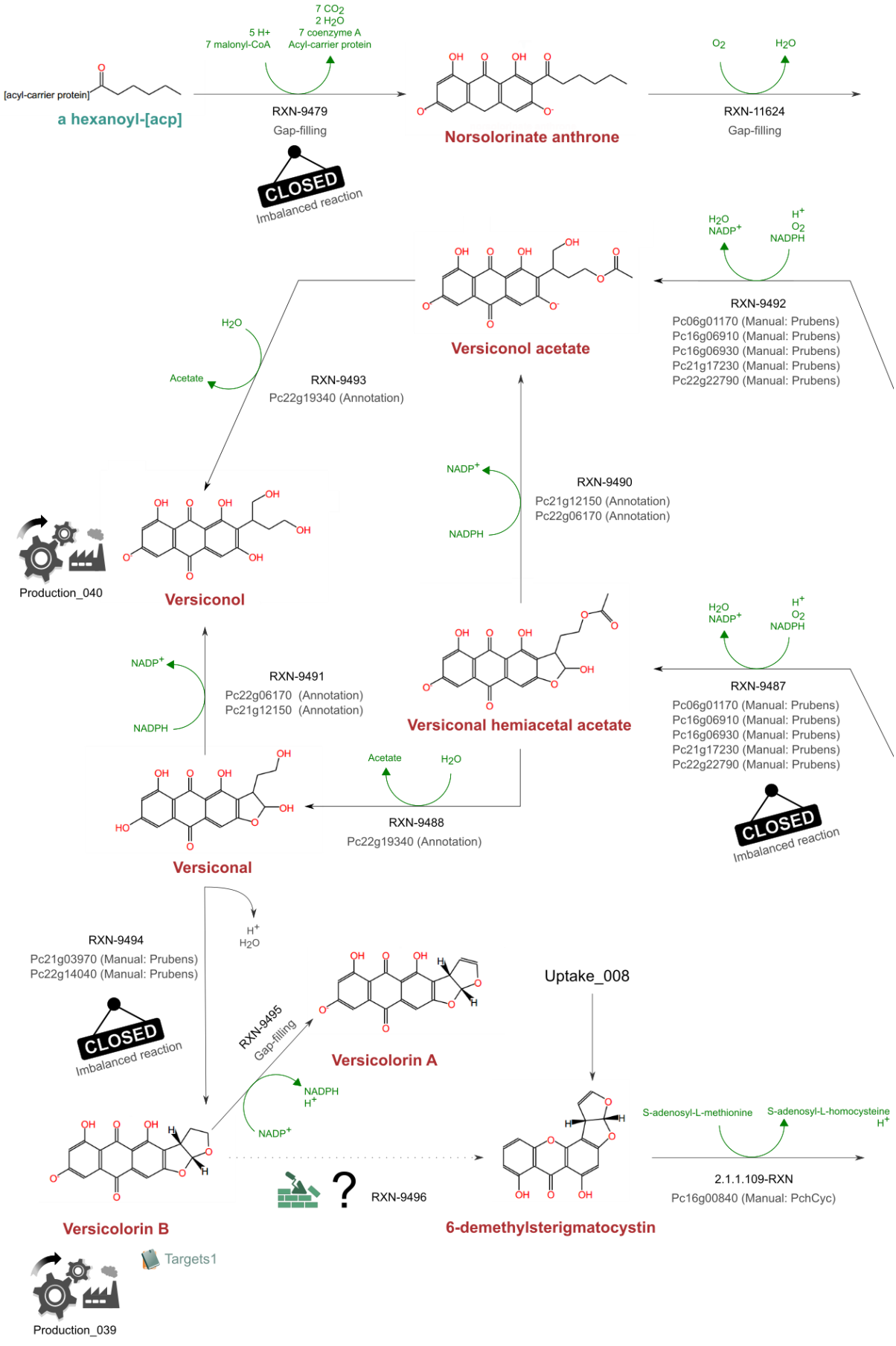


Figure 3-39 : Voie de biosynthèse de la mélanine.


[acyl-carrier protein]
a hexanoyl-[acp]
5 H+
7 malonyl-CoA
7 CO2
2 H2O
7 coenzyme A
Acyl-carrier protein
RXN-9479
Gap-filling
CLOSED
Imbalanced reaction
Norsolorinate anthrone
O2
H2O
RXN-11624
Gap-filling
Versiconol acetate
H2O
NADP+
H+
O2
NADPH
RXN-9492
Pc06g01170 (Manual: Prubens)
Pc16g06910 (Manual: Prubens)
Pc16g06930 (Manual: Prubens)
Pc21g17230 (Manual: Prubens)
Pc22g22790 (Manual: Prubens)
H2O
Acetate
RXN-9493
Pc22g19340 (Annotation)
NADP+
NADPH
RXN-9490
Pc21g12150 (Annotation)
Pc22g06170 (Annotation)
Production_040
Versiconol
Versiconal hemiacetal acetate
NADP+
NADPH
RXN-9491
Pc22g06170 (Annotation)
Pc21g12150 (Annotation)
Acetate
H2O
RXN-9488
Pc22g19340 (Annotation)
Versiconal
H2O
NADP+
H+
O2
NADPH
RXN-9487
Pc06g01170 (Manual: Prubens)
Pc16g06910 (Manual: Prubens)
Pc16g06930 (Manual: Prubens)
Pc21g17230 (Manual: Prubens)
Pc22g22790 (Manual: Prubens)
CLOSED
Imbalanced reaction
H+
H2O
RXN-9494
Pc21g03970 (Manual: Prubens)
Pc22g14040 (Manual: Prubens)
CLOSED
Imbalanced reaction
RXN-9495
Gap-filling
Versicolorin A
Uptake_008
NADPH
H+
NADP+
Versicolorin B
RXN-9496
6-demethylsterigmatocystin
S-adenosyl-L-methionine
S-adenosyl-L-homocysteine
H+
2.1.1.109-RXN
Pc16g00840 (Manual: PchCyc)
Targets1
Production_039

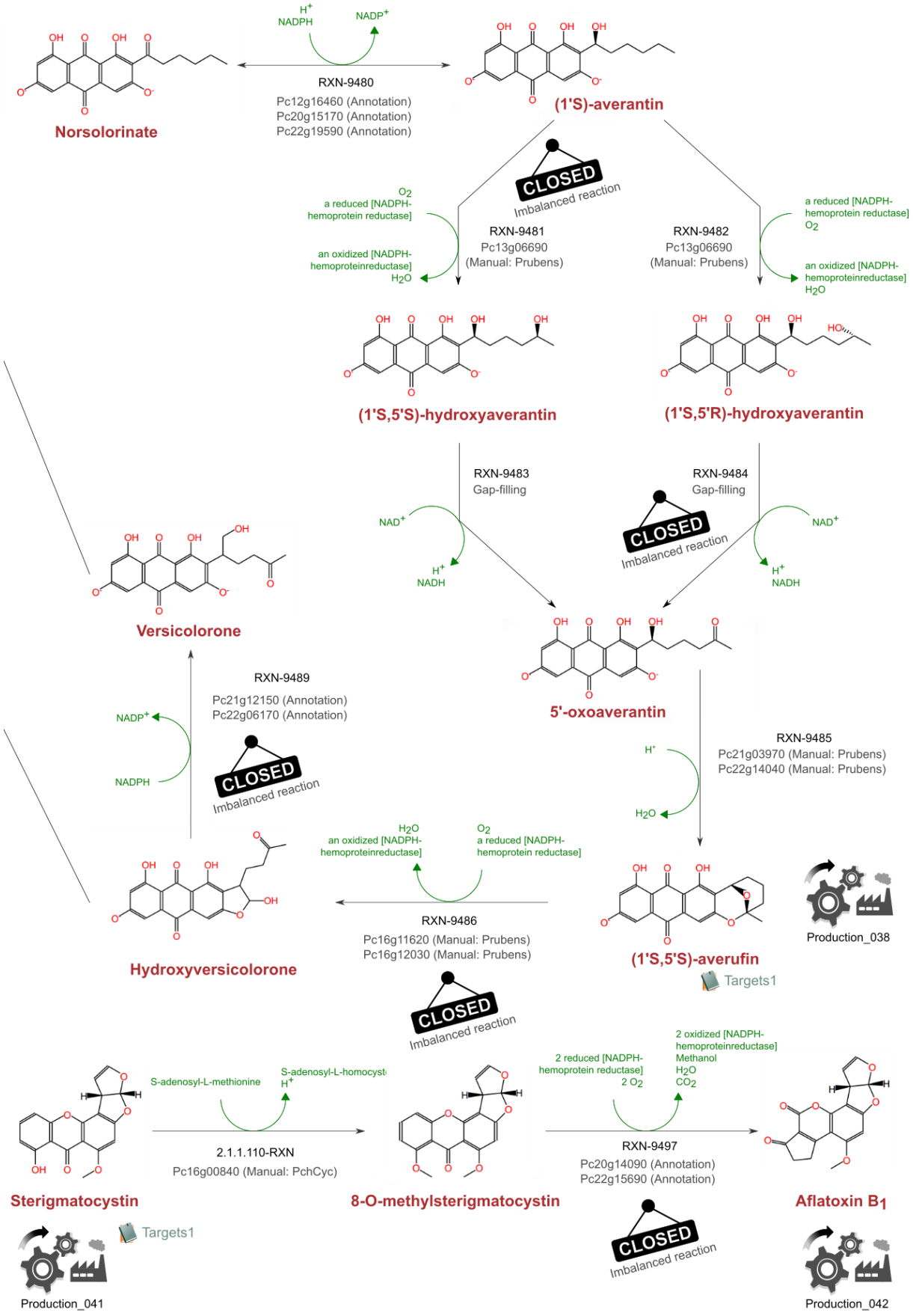


Figure 3-40 : Voies de biosynthèse de l'averufine et de la versicolorine B.

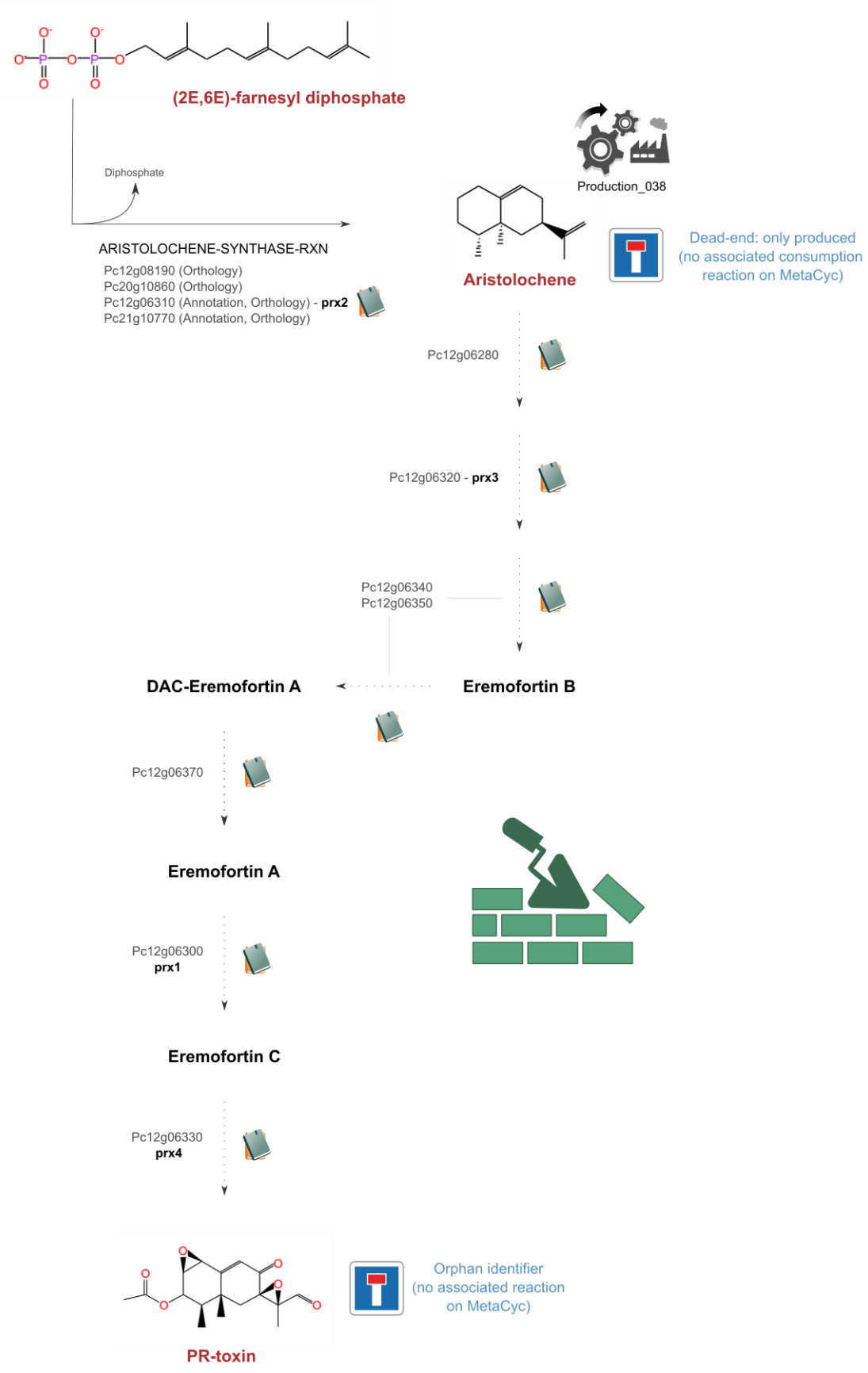


Figure 3-41 : Voie de biosynthèse de la PR-toxine.

3.3. Une production de métabolites spécialisés sensible aux variations de la composition de la réaction de biomasse ?

Avant de tester les potentielles différences de production de métabolites spécialisés en fonction des conditions environnementales, nous souhaitons évaluer la robustesse du modèle autour de la réaction de production de biomasse. Le premier élément à évaluer est donc le degré d'indépendance du modèle vis-à-vis de cette réaction, afin de déterminer si les variations environnementales influencent ou non la production de métabolites spécialisés. Pour ce faire, nous comparons les flux maximaux de production de 10 métabolites spécialisés ou précurseurs dans les conditions standards du modèle, puis, en variant de plus ou moins 50 % la valeur des coefficients des précurseurs de la réaction de biomasse (**Figure 3-43**). Cette analyse de robustesse sur la production des précurseurs des métabolites spécialisés nous sert ainsi d'indicateur sur la stabilité des phénotypes.

Nous observons alors deux types de profils qualitativement distincts : d'une part, les productions des métabolites des voies de biosynthèse des pénicillines (**Figure 3-43.A**), des roquefortines (**Figure 3-43.C**) et des ferrichromes (**Figure 3-43.E**), sensibles aux variations de certains coefficients stœchiométriques des précurseurs de la biomasse, et d'autre part, celles qui ne le sont pas. En effet, avec des variations de l'ordre de 10^{-5} mmol.gDW^{-1}.h^{-1} pour les métabolites des voies de biosynthèse de l'isoepoxydon et des yanuthones (**Figure 3-43.B**), de la mélanine (**Figure 3-43.D**), de la PR-toxine (**Figure 3-43.F**), de l'andrastine (**Figure 3-43.G**) et de l'avérufine et de la versicolorine (**Figure 3-43.H**), nous pouvons considérer que les variations des coefficients des précurseurs de la réaction de biomasse ont peu ou pas d'incidence sur la production maximale de ces métabolites.

Nous noterons toutefois que les variations des précurseurs `COF`, `AAPOOL`, `CELLWALL` et `PLIPIDS` entraînent des valeurs maximales de flux supérieures à celles déterminées sans modification des coefficients des précurseurs (*i.e.* valeurs de référence selon les contraintes définies par défaut). Il est fort plausible que ces légères variations d'amplitude soient liées à la précision du solveur utilisé puisque le seuil de tolérance de gurobi est fixé par défaut à 1.10^{-6} et peut donc occasionner des incertitudes du même ordre sur les calculs.

En revanche, les modifications des valeurs des précurseurs `PROTEIN`, `RNA` et `AAPOOL` provoquent des variations notables dans la production des métabolites premièrement cités (*i.e.* pénicillines, roquefortines et ferrichromes). Essentiellement, les variations les plus marquées sont observées lorsque les valeurs du coefficient du sous-système `PROTEIN` sont modifiées, entraînant des différences maximales de production de flux de plus de 0,5 mmol.gDW^{-1}.h^{-1} pour la 6-aminopénicilline, de 0,3 mmol.gDW^{-1}.h^{-1} pour l'isopénicilline N, de 0,2 mmol.gDW^{-1}.h^{-1} pour le HTD, et de 0,1 mmol.gDW^{-1}.h^{-1} pour les ferrichromes.

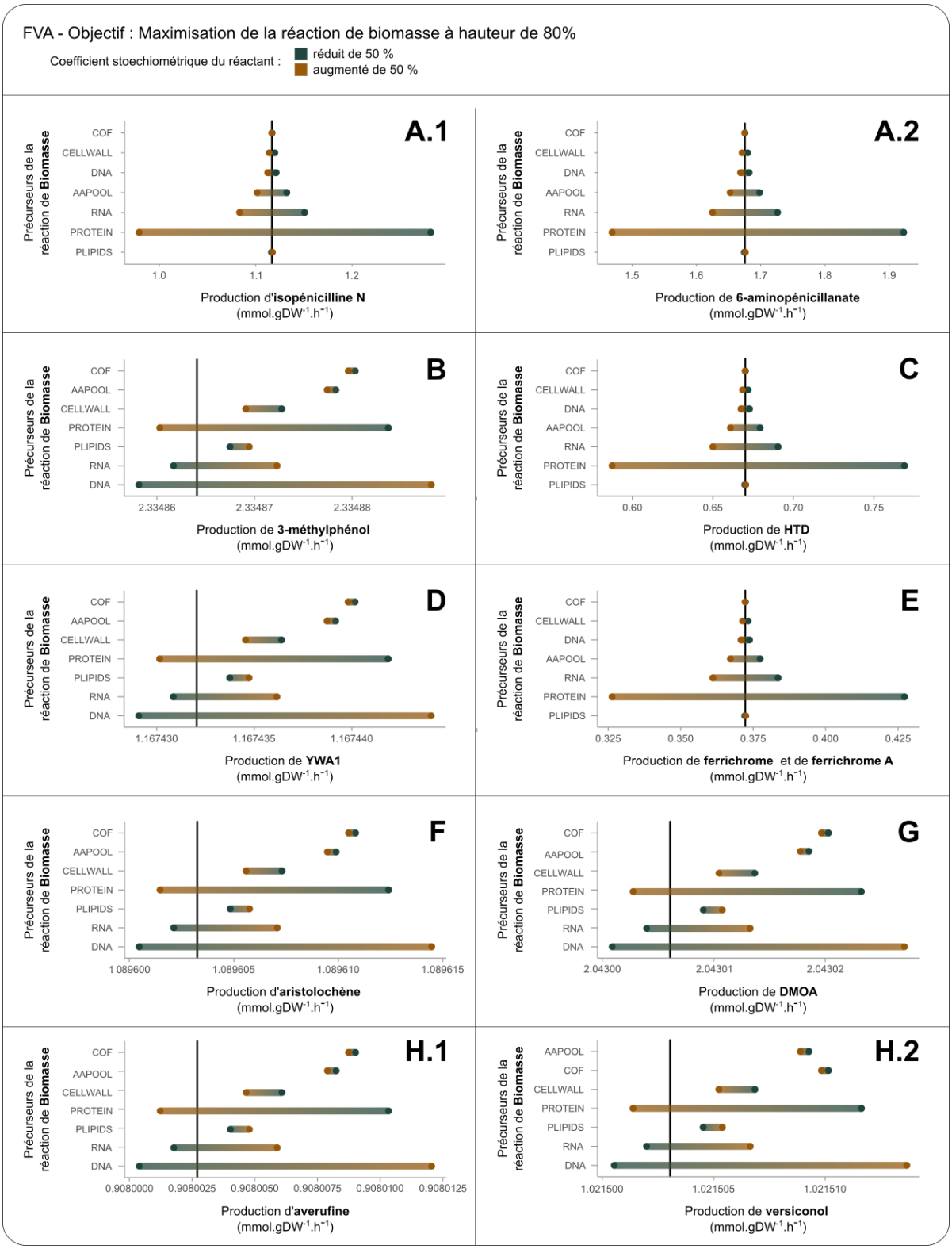


Figure 3-42 : Analyse de sensibilité des flux de production des métabolites spécialisés en fonction des variations des coefficients des précurseurs de la réaction de biomasse. *Une FVA maximisant la réaction de biomasse à 80 % est réalisée pour chaque modification de coefficient de la réaction de biomasse. Les points marron (●) et verts (●) représentent respectivement le résultat maximal de la FVA avec une augmentation et une réduction de 50 % de la valeur du coefficient des divers précurseurs, identifiés par leur identifiant MetaCyc. Ces simulations utilisent les contraintes définies précédemment (cf.* ***Tableau 3-10****) et autorisent un import infini pour la réaction Uptake_155 (i.e. Red-NADPH-Hemoprotein-Reductases - classe d'enzymes impliquées dans des réactions redox où elles utilisent le NADPH pour réduire des hémoprotéines). Pour effectuer ces tests, nous avons choisi les métabolites précurseurs (i.e. points de branchements) les plus représentatifs des voies de biosynthèse* ***(A)*** *des pénicillines,* ***(B)*** *de l'isoepoxydon et des yanuthones,* ***(C)*** *des roquefortines,* ***(D)*** *de la mélanine,* ***(E)*** *des ferrichromes,* ***(F)*** *de la PR-toxine,* ***(G)*** *des andrastines et* ***(H)*** *de l'avérufine et de la versicolorine B. Enfin, la ligne verticale noire représente la valeur maximale du flux de production testé sans modification des valeurs des coefficients des précurseurs de la biomasse.*

De prime abord, les comportements observés semblent cohérents avec ceux attendus. Logiquement, lors de l'augmentation des valeurs des coefficients des précurseurs de la réaction de biomasse, nous diminuons l'espace disponible et requis pour la production de métabolites spécialisés, d'où une valeur de production observée inférieure à celle de référence et inversement. De surcroît, les pénicillines, les roquefortines et les ferrichromes sont des peptides non-ribosomaux dont les voies de biosynthèse consomment des acides aminés. Il est donc cohérent de constater que les variations les plus marquées se trouvent sur les sous-systèmes `PROTEIN` et `AAPOOL`. Enfin, ces variations raisonnables sont le témoin de la sensibilité du modèle et expriment la variabilité biologique potentielle de production des métabolites spécialisés en conditions expérimentales *in vitro*. Ces variations devront toutefois être mises en perspective par rapport aux flux observés expérimentalement. Ainsi, si nous venions à observer des différences spécifiques de production sur ces métabolites lors des simulations de changements d'environnement, elles devront être interprétées avec vigilance.

En conclusion, la composition quantitative de la réaction de biomasse, sous les contraintes de modélisation précédemment définies, ne présente pas d'impact sur la production des métabolites spécialisés autres que les peptides non-ribosomaux. Pour ces derniers, les variations observées dans des conditions relativement extrêmes (*i.e.* variations calculées sur la base d'une augmentation ou d'une diminution des coefficients de 50 % de leur valeur initiale) ne semblent pas éloignées des variations biologiques intrinsèques. Ces résultats sont donc un appui supplémentaire permettant de mettre en évidence les diverses réponses d'***i*Prub22** aux perturbations témoignant ainsi de la sensibilité des modèles. Nous observons également des différences de comportement selon les voies de biosynthèse considérées et, outre une légère variabilité pouvant être attribuée à la variabilité biologique, nous pouvons considérer que les résultats présentés dans la partie suivante sont majoritairement indépendants de la composition quantitative de la réaction de biomasse.

3.4. Une production de métabolites spécialisés sensible aux conditions environnementales ?

Le dernier aspect que nous souhaitons aborder dans ce chapitre concerne la sensibilité de la production de métabolites spécialisés d'***i*Prub22**. Les modèles de ce système offrent-ils un cadre de simulations pour une méthode **OSMAC *in silico*** ? Pour répondre à cette question, nous développons les résultats présentés en **Figure 3-6** (page 157) en testant les diverses conditions environnementales décrites en section *2.2.2 Modélisation de différentes conditions de culture*, (page 256). Ainsi, l'impact de la variation des sources de carbone puis d'azote sur les flux maximaux potentiels de production de divers métabolites spécialisés, de leurs intermédiaires ou de leurs précurseurs est évalué par FVA en maximisant la réaction de biomasse à 80 %. Nous disposons de 27 composés différents pouvant servir de source de carbone et de 29 composés pour simuler la source d'azote. *Penicillium rubens* Wisconsin 54-1255 ayant été largement étudié pour sa capacité de production de *β*-lactames (cf. encart : *La production des pénicillines dans la littérature*, page 339), nous commençons par étudier les variations sur la production de pénicillines, dont les résultats sont présentés en **Figure 3-43**.

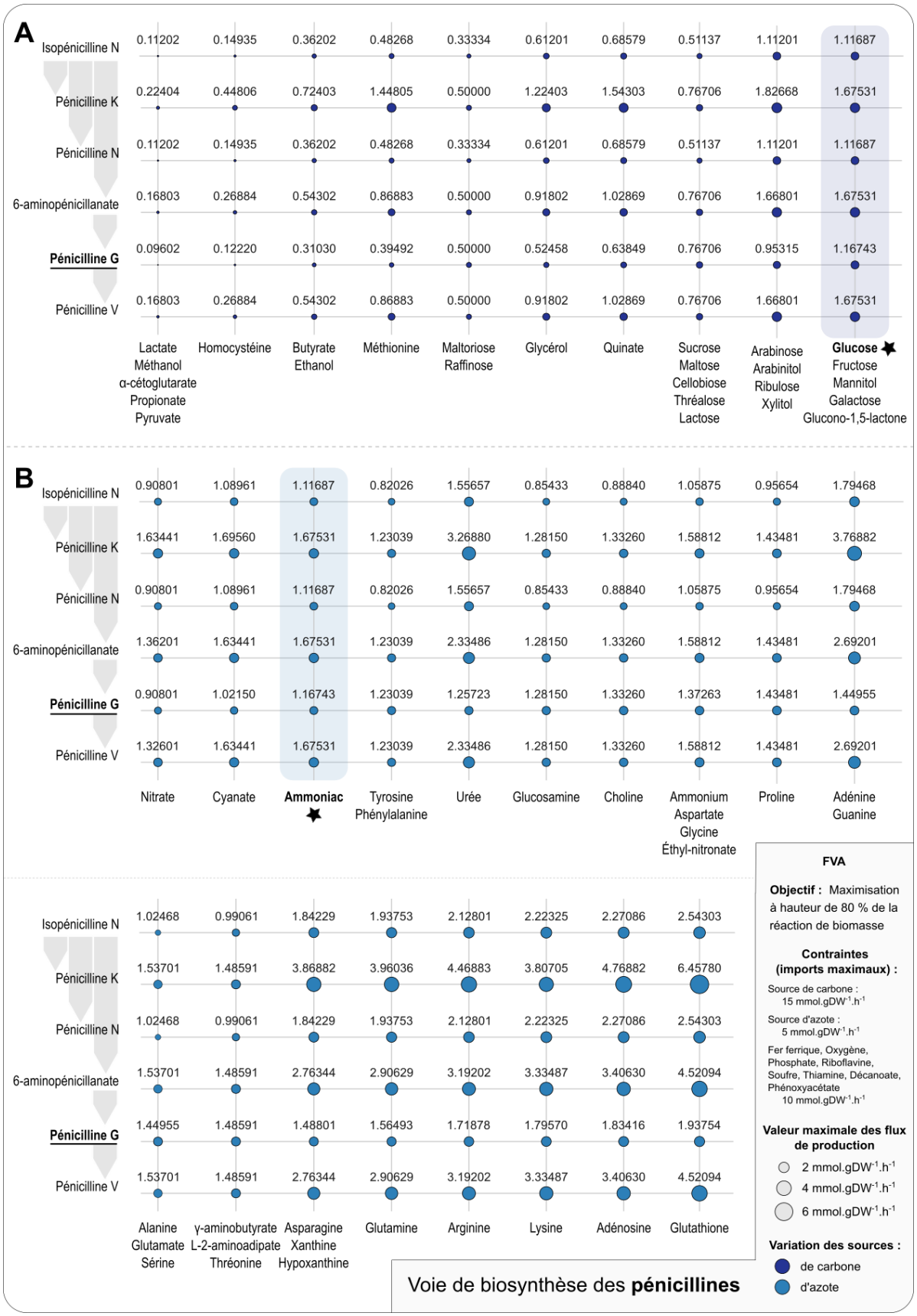


Figure 3-43 : Évaluations des productions maximales potentielles des métabolites de la voie de biosynthèse des pénicillines, testées sous différentes conditions environnementales. (A) *Variation de la source de carbone.* ***(B)*** *Variation de la source d'azote. Les conditions encadrées « ★ » correspondent aux valeurs par défaut obtenues lors du chargement d'**iPrub22** (i.e. utilisation du glucose et de l'ammoniac comme source de carbone et d'azote), présentées en **Figure 3-6**. Le classement des conditions est effectué selon les flux maximaux de production potentielle croissants de pénicilline G. Les flèches grises entre les noms des molécules représentent l'enchaînement des molécules dans la voie de biosynthèse.*

La production des pénicillines dans la littérature

À l'instar des éléments présentés dans le **Tableau 3-11** relatifs aux observations expérimentales liées à la croissance de *P. rubens*, diverses vérifications d'expression phénotypique devront être effectuées. Cette mise en parallèle avec les résultats des simulations d'***i*Prub22** permettra de renforcer le modèle ou de l'affiner. À ce titre, nous énonçons ci-dessous quelques comportements clés de la biosynthèses des pénicillines à vérifier en priorité *(Nielsen 1997 ; Prauße et al. 2016)*.

Nous avons vu précédemment que le soufre, le phosphore et le fer sont des éléments inorganiques essentiels à la croissance de *P. rubens*. Ils sont également cruciaux pour la production de pénicilline, nécessitant des concentrations supérieures à celles requises pour la croissance, avec une augmentation respective d'un facteur de 1,5, 3 et 25 des concentrations pour une activité maximale.

Lors de cultures en fed-batch, *P. rubens* suit un processus dynamique de croissance en trois phases distinctes suggérant une redistribution des flux à travers diverses voies métaboliques : une phase de croissance rapide, une phase linéaire et une phase ralentie. Pendant la troisième phase, les taux spécifiques d'absorption du glucose restent constants à 0,22 mmol glucose.gDW^{-1}.h^{-1}, légèrement au-dessus du taux d'entretien de 0,12 mmol glucose.gDW^{-1}.h^{-1}. Le rendement moyen de pénicilline V est alors estimé à 0,045 mol par mol de glucose, mais peut atteindre 0,083 mol par mol de glucose en présence des acides aminés précurseurs cystéine, valine et aminoadipate, soit un rendement six fois supérieur.

L'acide pénicilloïque, l'acide pénilloïque et la *p*-OH-pénicilline V, sous-produits de la biosynthèse des pénicillines, peuvent représenter 10 à 20 % de la pénicilline sécrétée initialement par le champignon. Lors de cultures en fed-batch, environ 30 % de l'ACV (*i.e.* précurseur biologique de la pénicilline) accumulé reste intracellulaire, tandis que 70 % est sécrété. Le taux initial de production de pénicilline, dès lors relativement élevé, augmente encore après la phase de croissance rapide, atteignant un maximum de 13,7.10^{-3} mol de pénicilline V.gDW^{-1}.h^{-1}, avant de diminuer rapidement au cours de la troisième phase. Lors d'une production rapide de pénicilline, une production accrue d'*a*-cétobutyrate (*i.e.* intermédiaire dans la voie de dégradation de la thréonine et dans les voies de biosynthèse d'acides aminés soufrés) est observée alors que dans des conditions de croissance normales, l'ensemble de sa production est dirigée vers la voie de biosynthèse de l'isoleucine.

Outre les paramètres physiques (*i.e.* température, pH, lumière, *etc.*), les concentrations de glucose et d'acides aminés sont des variables déterminantes pour la production de pénicilline. À des taux de croissance qualifiés de faibles (*i.e.* inférieurs à 0,20 h^{-1}), la production de pénicilline observée diminue. Ainsi, un taux minimal de croissance doit être conservé pour autoriser la production de pénicilline signifiant alors qu'une activité minimale du métabolisme basal doit être maintenue afin de soutenir le métabolisme spécialisé. Il a également été observé que l'acétate a un effet positif sur la production de pénicilline avec une amélioration du rendement de 25 % entre l'utilisation du glucose seul et du glucose couplé à l'acétate.

Le soufre, élément essentiel dans la biosynthèse des pénicillines, est apporté par la L-cystéine synthésisé par deux voies distinctes : la sulfhydrylation directe et la transsulfuration. Cependant, la transsulfuration, énergétiquement plus coûteuse, réduit d'environ 15 % le rendement théorique maximal de pénicilline sur glucose par rapport à la sulfhydrylation directe. La majorité de la L-cystéine cellulaire n'est pas présente sous forme d'acide aminé libre, mais est liée en peptide dans la glutathione, le thiol intracellulaire le plus abondant qui sert donc de réserve à cet acide aminé. La glutathione est présente dans l'organisme à des concentrations 10 à 15 fois supérieures à celles de la L-cystéine et représente ainsi environ 0,4 % du poids sec cellulaire. Pendant une culture en fed-batch, le niveau de glutathione reste constant mais augmente avec la concentration en ammoniaque, en raison probablement de l'augmentation du pool de L-glutamate.

La **Figure 3-43** révèle deux types de variations de flux : les variations entre les sources de carbone/azote utilisées (*i.e.* lecture horizontale) et celles au sein de la voie de biosynthèse étudiée (*i.e.* lecture verticale). Étant donné que la précision du solveur est fixée à 10^{-6}, les comparaisons ont été effectuées avec une précision de l'ordre de 10^{-5}.

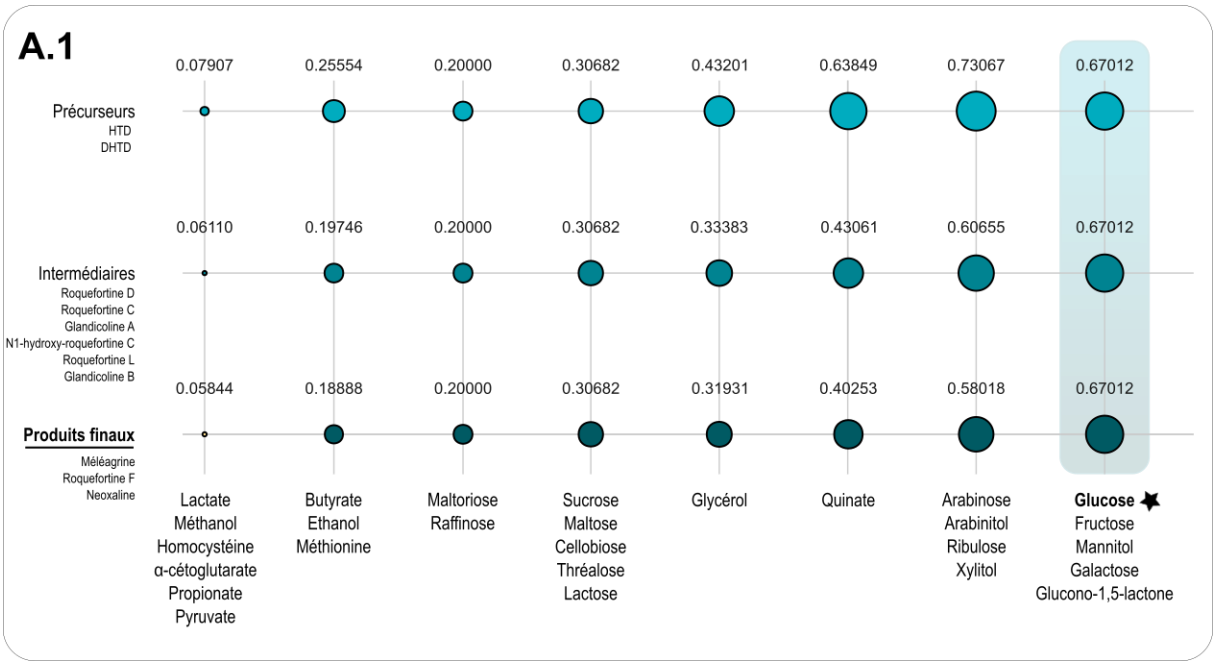
Selon la source de carbone employée, 10 profils différents sont observés. Au sein de la voie de biosynthèse, l'isopénicilline N et la pénicilline N partagent les mêmes valeurs de flux maximaux potentiels, indiquant que la production d'isopénicilline N est un facteur limitant pour la production de pénicilline N. Un constat similaire est établi pour l'intermédiaire 6-aminopénicillanate et la pénicilline V. Entre les meilleures sources de carbone potentielles (*i.e.* glucose, fructose, mannitol, galactose, ou glucono-1,5-lactone) et celles où la production théorique de pénicilline G est la plus faible (*i.e.* lactate, méthanol, *a*-cétoglutarate, propionate, ou pyruvate), nous observons une différence d'un facteur de 12,2. La pénicilline G est le métabolite dont les variations de flux sont le plus marquées en fonction du changement de la source de carbone. À l'exception de la pénicilline K, qui voit son flux maximal théorique multiplié par 8,2, les autres variations de flux témoignent d'une amélioration théorique d'un facteur 10. De surcroît, le classement des conditions présentées en **Figure 3-43.A** favorise la production de pénicilline G, mais cet ordre diffère pour les autres produits finaux de la voie de biosynthèse. Par exemple, la méthionine pourrait être envisagée pour l'enrichissement du milieu de culture en vue d'une production améliorée de pénicilline K, mais serait de moindre intérêt pour les autres pénicillines. Ceci, cependant, suggère qu'il semble possible de privilégier la production de certains métabolites par rapport à d'autres au sein d'une même voie de biosynthèse ; au moins, quand celle-ci n'est pas linéaire.

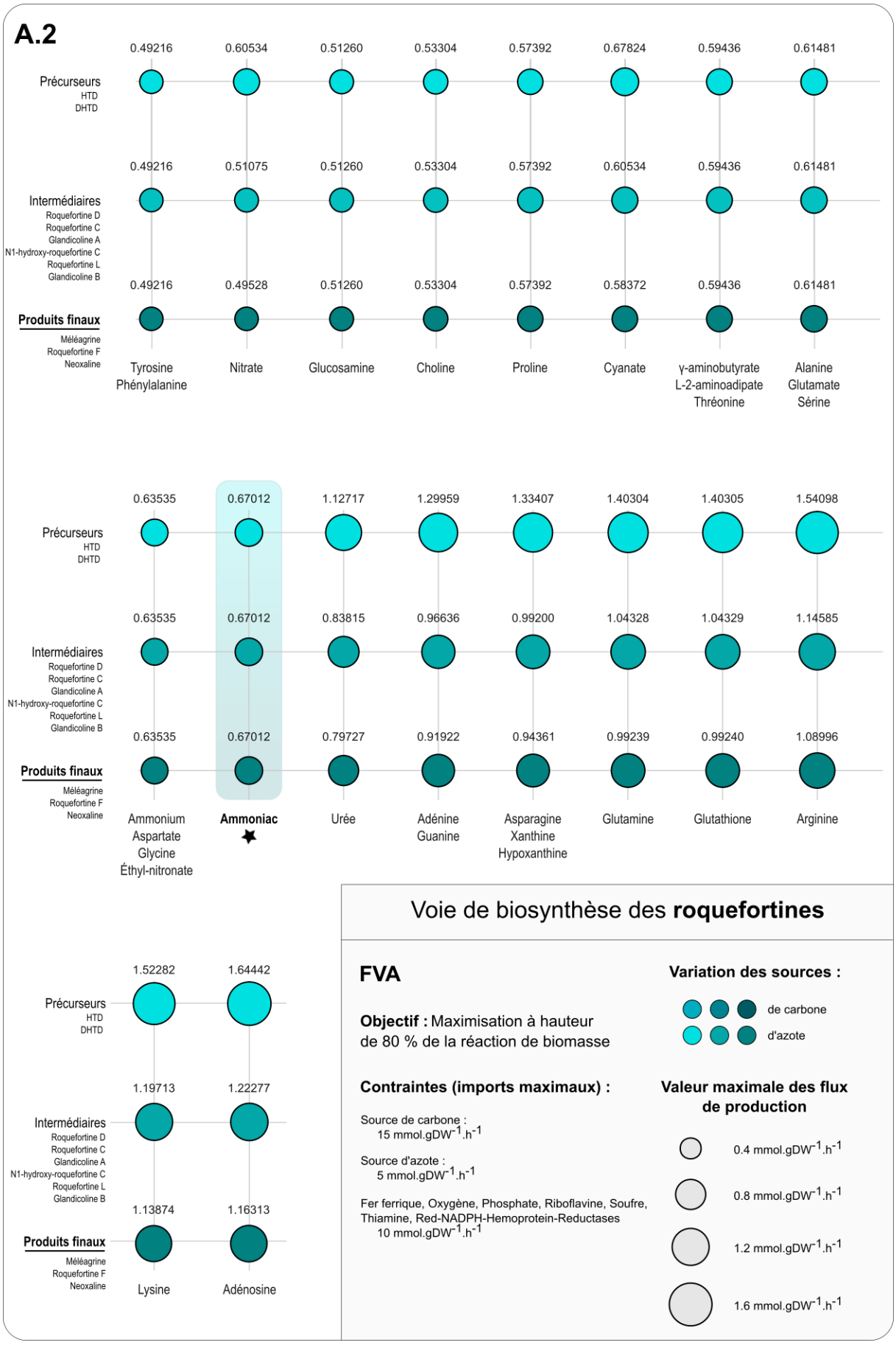
Concernant les variations liées au changement de la source d'azote, nous observons en **Figure 3-43.B** l'existence de 18 comportements différents avec un flux de pénicilline G multiplié par 2,1 entre les deux conditions extrêmes. En revanche, et à l'inverse de l'impact de la source de carbone, les autres métabolites de la voie présentent des variations plus importantes avec un facteur multiplicatif de 3,1 pour l'isopénicilline N et la pénicilline N, de 3,7 pour le 6-aminopénicillanate et la pénicilline V, et de 5,2 pour la pénicilline K. Toutefois, selon ces simulations, il semblerait que le choix de la source d'azote ait moins d'impact que celui de la source de carbone pour l'optimisation de la production des produits de cette voie de biosynthèse.

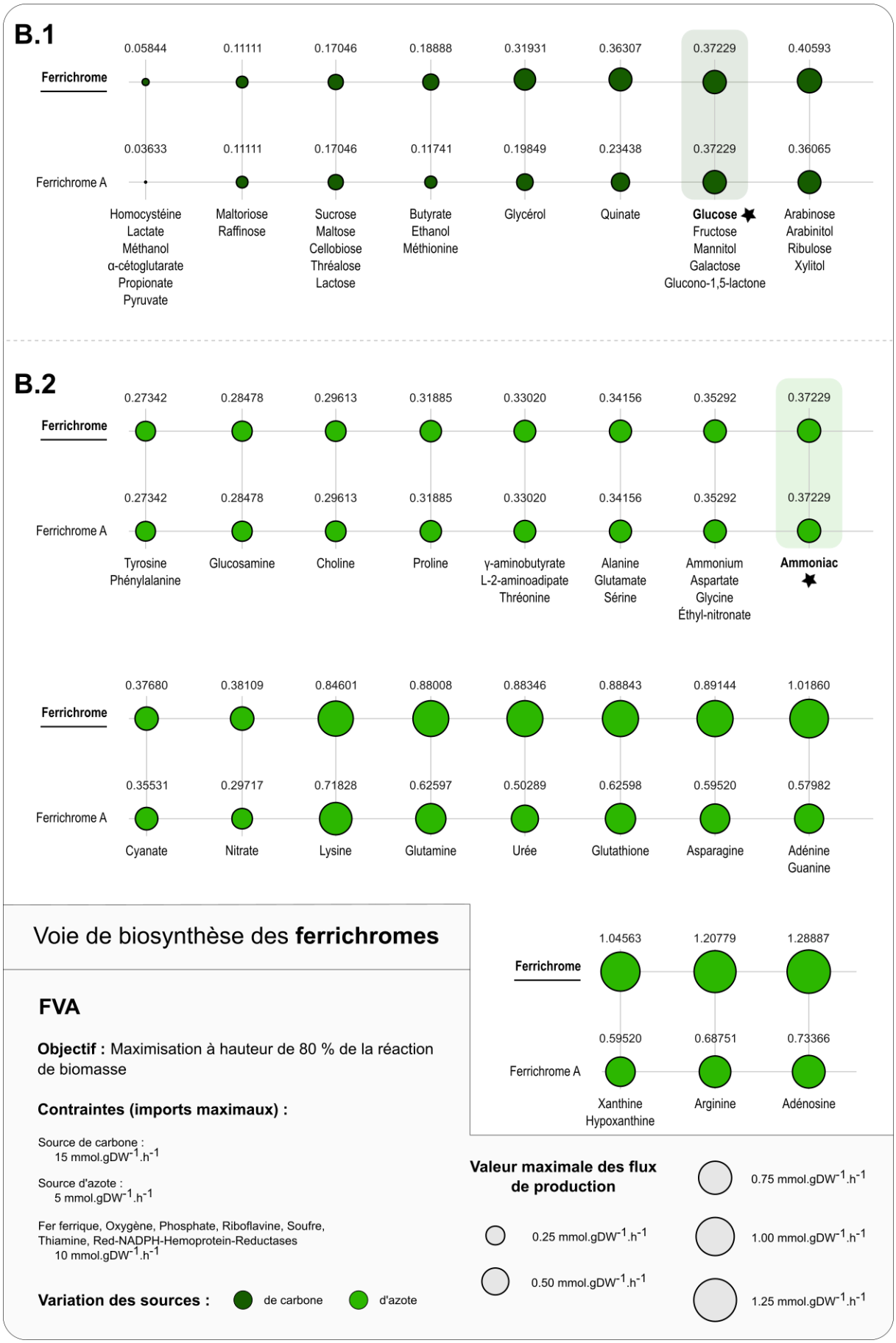
Enfin, nous constatons que l'ammoniac comme seule source d'azote associé au glucose comme source de carbone ne semble pas être une condition optimale de production de pénicilline G. En revanche, la glutathione semble être la meilleure condition potentielle pour tous les métabolites de cette voie, selon les conditions testées.

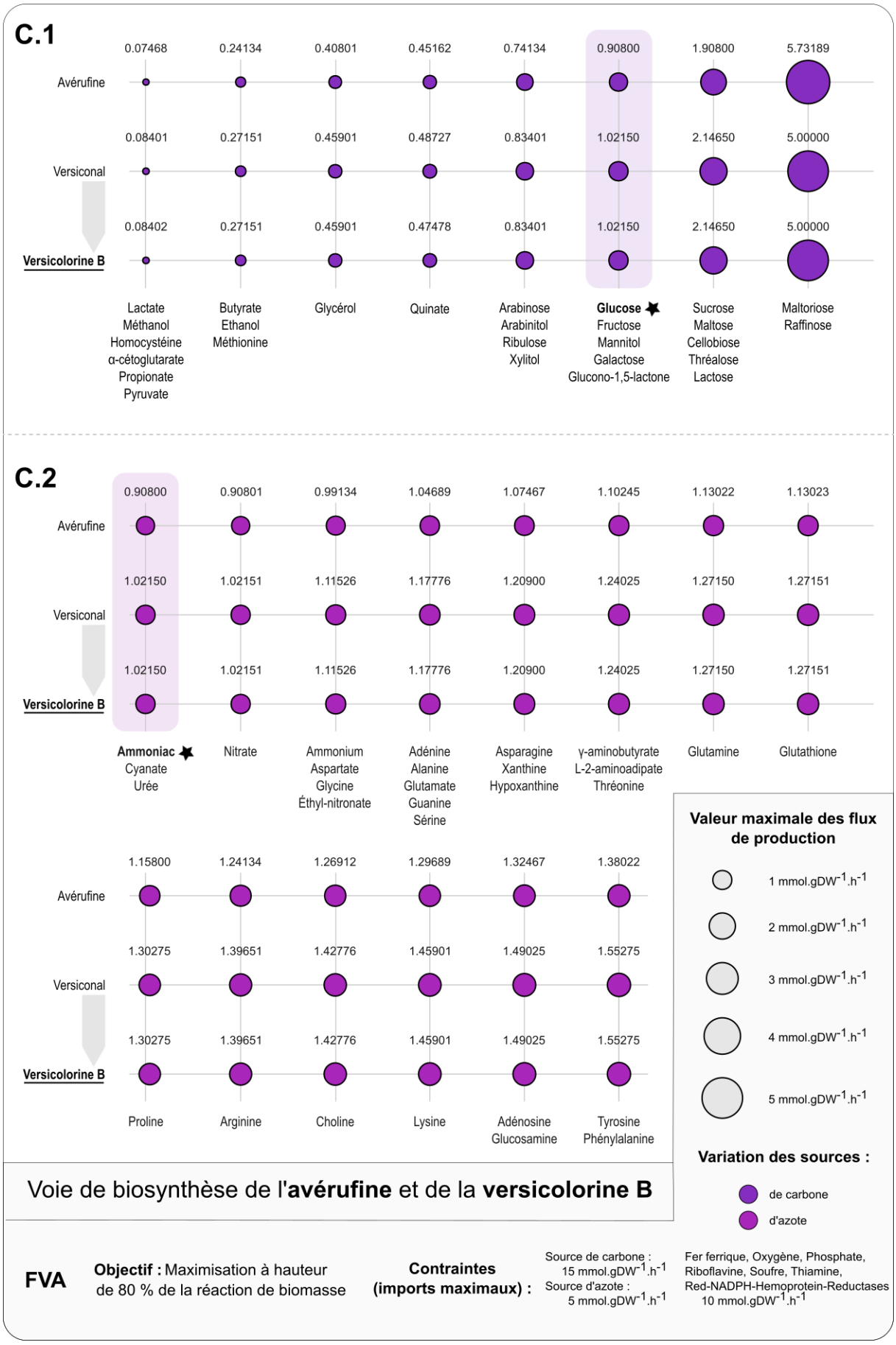
Les divers constats énoncés ci-dessus témoignent une fois de plus de la sensibilité d'***i*Prub22** aux perturbations. Néanmoins, afin de consolider les capacités prédictives des modèles, il serait nécessaire, d'une part, de les mettre en perspective avec les comportements connus (*i.e.* vérifier l'adéquation avec les phénotypes recensés), et d'autre part, d'approfondir les simulations. En effet, les résultats présentés ici demeurent sommaires puisque nous avons modifié une seule contrainte à la fois en nous focalisant exclusivement sur les sources de carbone puis d'azote. À l'avenir, nous pourrions envisager de tester des combinaisons de nutriments différentes et plus complexes afin de détecter, par exemple, de potentielles interactions entre conditions.

Les constatations de variations intra-voie de biosynthèse et inter-conditions de culture observées pour les composés de la voie de biosynthèse des pénicillines sont-elles généralisables à l'ensemble des métabolites spécialisés ? Les résultats présentés en **Figure 3-6** (page 157) semblaient suggérer une uniformité des flux maximaux potentiels au sein des voies de biosynthèse autre que celle des pénicillines, cet élément se vérifie-t-il sur l'ensemble des conditions testées ? Notons toutefois que l'activation de la voie de biosynthèse des pénicillines requiert des contraintes différentes (*i.e.* phénoxyacétate et de décanoate ou octanoate) de celles utilisées pour la production des autres métabolites spécialisés sélectionnés. Ces importations ont donc été fermées pour les simulations présentées en **Figure 3-44**.




A.2
Précurseurs
HTD
DHTD
Intermédiaires
Roquefortine D
Roquefortine C
Glandicoline A
N1-hydroxy-roquefortine C
Roquefortine L
Glandicoline B
Produits finaux
Méléagrine
Roquefortine F
Neoxaline
0.49216 0.60534 0.51260 0.53304 0.57392 0.67824 0.59436 0.61481
0.49216 0.51075 0.51260 0.53304 0.57392 0.60534 0.59436 0.61481
0.49216 0.49528 0.51260 0.53304 0.57392 0.58372 0.59436 0.61481
Tyrosine
Phénylalanine
Nitrate
Glucosamine
Choline
Proline
Cyanate
γ-aminobutyrate
L-2-aminoadipate
Thréonine
Alanine
Glutamate
Sérine
0.63535 0.67012 1.12717 1.29959 1.33407 1.40304 1.40305 1.54098
0.63535 0.67012 0.83815 0.96636 0.99200 1.04328 1.04329 1.14585
0.63535 0.67012 0.79727 0.91922 0.94361 0.99239 0.99240 1.08996
Ammonium
Aspartate
Glycine
Éthyl-nitronate
Ammoniac
Urée
Adénine
Guanine
Asparagine
Xanthine
Hypoxanthine
Glutamine
Glutathione
Arginine
1.52282 1.64442
1.19713 1.22277
1.13874 1.16313
Lysine
Adénosine
Voie de biosynthèse des roquefortines
FVA
Objectif : Maximisation à hauteur de 80 % de la réaction de biomasse
Contraintes (imports maximaux) :
Source de carbone : 15 mmol.gDW-1.h-1
Source d'azote : 5 mmol.gDW-1.h-1
Fer ferrique, Oxygène, Phosphate, Riboflavine, Soufre, Thiamine, Red-NADPH-Hemoprotein-Reductases 10 mmol.gDW-1.h-1
Variation des sources :
de carbone
d'azote
Valeur maximale des flux de production
0.4 mmol.gDW-1.h-1
0.8 mmol.gDW-1.h-1
1.2 mmol.gDW-1.h-1
1.6 mmol.gDW-1.h-1


B.1
Ferrichrome
0.05844
0.11111
0.17046
0.18888
0.31931
0.36307
0.37229
0.40593
Ferrichrome A
0.03633
0.11111
0.17046
0.11741
0.19849
0.23438
0.37229
0.36065
Homocystéine
Lactate
Méthanol
α-cétoglutarate
Propionate
Pyruvate
Maltoriose
Raffinose
Sucrose
Maltose
Cellobiose
Thréalose
Lactose
Butyrate
Ethanol
Méthionine
Glycérol
Quinate
Glucose
Fructose
Mannitol
Galactose
Glucono-1,5-lactone
Arabinose
Arabinitol
Ribulose
Xylitol
B.2
Ferrichrome
0.27342
0.28478
0.29613
0.31885
0.33020
0.34156
0.35292
0.37229
Ferrichrome A
0.27342
0.28478
0.29613
0.31885
0.33020
0.34156
0.35292
0.37229
Tyrosine
Phénylalanine
Glucosamine
Choline
Proline
γ-aminobutyrate
L-2-aminoadipate
Thréonine
Alanine
Glutamate
Sérine
Ammonium
Aspartate
Glycine
Éthyl-nitronate
Ammoniac
Ferrichrome
0.37680
0.38109
0.84601
0.88008
0.88346
0.88843
0.89144
1.01860
Ferrichrome A
0.35531
0.29717
0.71828
0.62597
0.50289
0.62598
0.59520
0.57982
Cyanate
Nitrate
Lysine
Glutamine
Urée
Glutathione
Asparagine
Adénine
Guanine
Ferrichrome
1.04563
1.20779
1.28887
Ferrichrome A
0.59520
0.68751
0.73366
Xanthine
Hypoxanthine
Arginine
Adénosine
Voie de biosynthèse des ferrichromes
FVA
Objectif : Maximisation à hauteur de 80 % de la réaction de biomasse
Contraintes (imports maximaux) :
Source de carbone :
15 mmol.gDW-1.h-1
Source d'azote :
5 mmol.gDW-1.h-1
Fer ferrique, Oxygène, Phosphate, Riboflavine, Soufre, Thiamine, Red-NADPH-Hemoprotein-Reductases
10 mmol.gDW-1.h-1
Variation des sources :
de carbone
d'azote
Valeur maximale des flux de production
0.25 mmol.gDW-1.h-1
0.50 mmol.gDW-1.h-1
0.75 mmol.gDW-1.h-1
1.00 mmol.gDW-1.h-1
1.25 mmol.gDW-1.h-1


C.1
0.07468
0.24134
0.40801
0.45162
0.74134
0.90800
1.90800
5.73189
Avérufine
0.08401
0.27151
0.45901
0.48727
0.83401
1.02150
2.14650
5.00000
Versiconal
0.08402
0.27151
0.45901
0.47478
0.83401
1.02150
2.14650
5.00000
Versicolorine B
Lactate
Méthanol
Homocystéine
α-cétoglutarate
Propionate
Pyruvate
Butyrate
Ethanol
Méthionine
Glycérol
Quinate
Arabinose
Arabinitol
Ribulose
Xylitol
Glucose
Fructose
Mannitol
Galactose
Glucono-1,5-lactone
Sucrose
Maltose
Cellobiose
Thréalose
Lactose
Maltoriose
Raffinose
C.2
0.90800
0.90801
0.99134
1.04689
1.07467
1.10245
1.13022
1.13023
Avérufine
1.02150
1.02151
1.11526
1.17776
1.20900
1.24025
1.27150
1.27151
Versiconal
1.02150
1.02151
1.11526
1.17776
1.20900
1.24025
1.27150
1.27151
Versicolorine B
Ammoniac
Cyanate
Urée
Nitrate
Ammonium
Aspartate
Glycine
Éthyl-nitronate
Adénine
Alanine
Glutamate
Guanine
Sérine
Asparagine
Xanthine
Hypoxanthine
γ-aminobutyrate
L-2-aminoadipate
Thréonine
Glutamine
Glutathione
1.15800
1.24134
1.26912
1.29689
1.32467
1.38022
Avérufine
1.30275
1.39651
1.42776
1.45901
1.49025
1.55275
Versiconal
1.30275
1.39651
1.42776
1.45901
1.49025
1.55275
Versicolorine B
Proline
Arginine
Choline
Lysine
Adénosine
Glucosamine
Tyrosine
Phénylalanine
Valeur maximale des flux de production
1 mmol.gDW-1.h-1
2 mmol.gDW-1.h-1
3 mmol.gDW-1.h-1
4 mmol.gDW-1.h-1
5 mmol.gDW-1.h-1
Variation des sources :
de carbone
d'azote
Voie de biosynthèse de l'avérufine et de la versicolorine B
FVA
Objectif : Maximisation à hauteur de 80 % de la réaction de biomasse
Contraintes (imports maximaux) :
Source de carbone : 15 mmol.gDW-1.h-1
Source d'azote : 5 mmol.gDW-1.h-1
Fer ferrique, Oxygène, Phosphate, Riboflavine, Soufre, Thiamine, Red-NADPH-Hemoprotein-Reductases 10 mmol.gDW-1.h-1


D.1
3-méthylphénol
0.19203 0.62060 1.04916 1.08920 1.90630 2.33486 4.90630 14.73914
Toluquinol
Précurseur yanuthones
0.19203 0.62060 1.04916 1.08920 1.90630 2.33486 4.90630 10.00000
Isoepoxydon
0.19203 0.62060 1.04916 1.15727 1.90630 2.33486 4.90630 5.00000
YWA1
Précurseur mélanine
0.09602 0.31003 0.42458 0.66130 0.95315 1.16743 2.45315 7.36957
Aristolochène
Précurseur PR-Toxine
0.08961 0.28961 0.48961 0.44087 0.88961 1.08960 2.28960 6.64357
DMOA
Précurseur andrastines
0.16803 0.54302 0.91802 0.97454 1.66801 2.04301 4.29301 12.89675
Lactate
Méthanol
Homocystéine
α-cétoglutarate
Propionate
Pyruvate
Butyrate
Ethanol
Méthionine
Glycérol
Quinate
Arabinose
Arabinitol
Ribulose
Xylitol
Glucose
Fructose
Mannitol
Galactose
Glucono-1,5-lactone
Sucrose
Maltose
Cellobiose
Thréalose
Lactose
Maltoriose
Raffinose
D.2
3-méthylphénol
2.33486 2.33487 2.54916 2.69201 2.76344 2.83486 2.90629 2.90632
Toluquinol
Précurseur yanuthones
2.33486 2.33487 2.54916 2.69201 2.76344 2.83486 2.90629 2.90632
Isoepoxydon
2.33486 2.33487 2.54916 2.69201 2.76344 2.83486 2.90629 2.90632
YWA1
Précurseur mélanine
1.16743 1.16744 1.27458 1.34601 1.38172 1.41734 1.45314 1.45316
Aristolochène
Précurseur PR-Toxine
1.08960 1.08961 1.18961 1.24627 1.28961 1.32294 1.35627 1.35628
DMOA
Précurseur andrastines
2.04301 2.04302 2.23052 2.35551 2.41801 2.48051 2.54300 2.54303
Ammoniac
Cyanate
Urée
Nitrate
Ammonium
Aspartate
Glycine
Éthyl-nitronate
Adénine
Alanine
Glutamate
Guanine
Sérine
Asparagine
Xanthine
Hypoxanthine
γ-aminobutyrate
L-2-aminoadipate
Thréonine
Glutamine
Glutathione

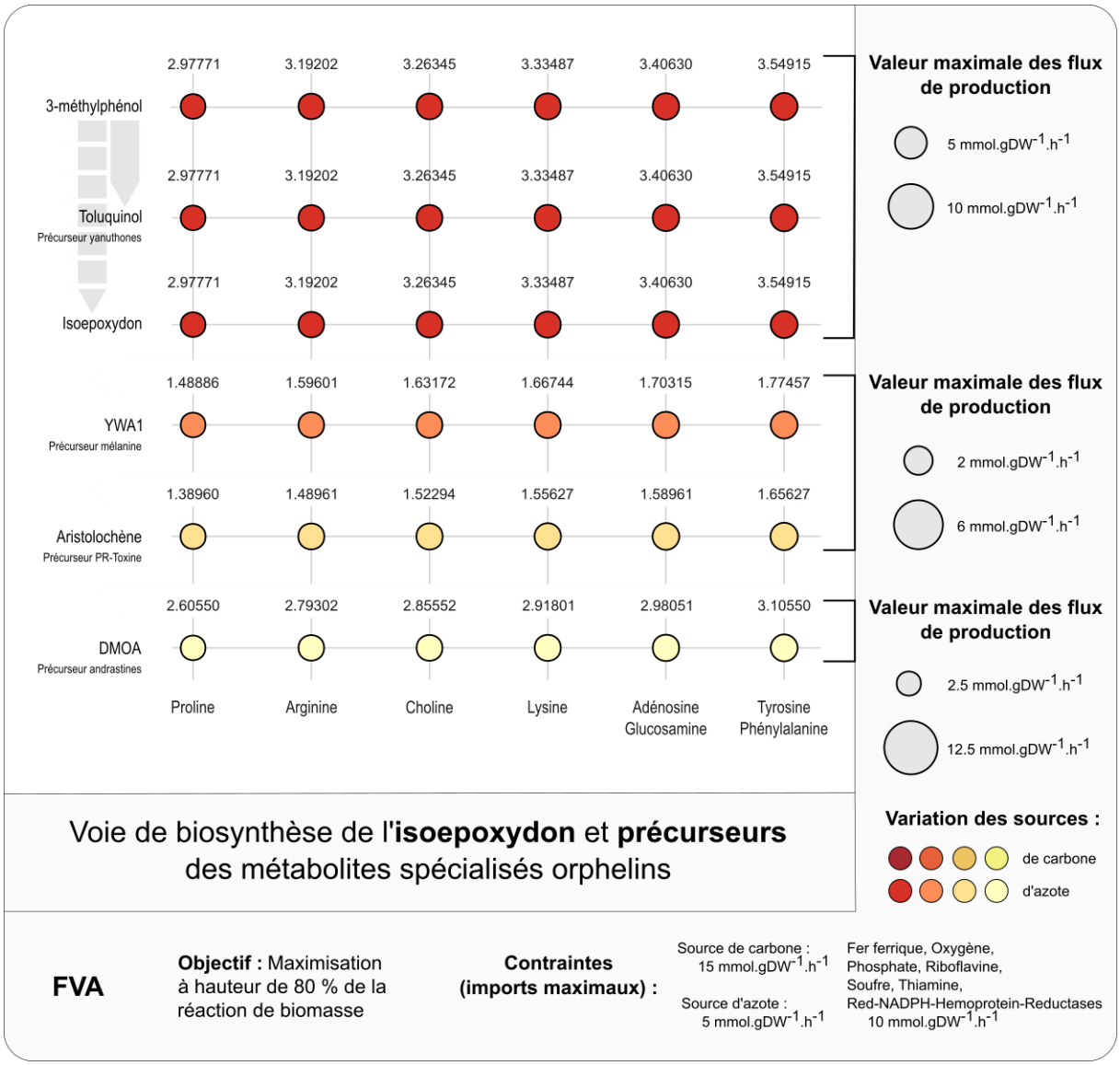


Figure 3-44 : Evaluations, testées sous différentes conditions environnementales, des productions maximales potentielles (A) de l'ensemble des composés de la voie de biosynthèse des roquefortines, (B) des produits finaux de la voie de biosynthèse des ferrichromes, (C) de l'avérufine, de la versicolorine B et de son antécédent direct et (D) des précurseurs des voies de biosynthèse des yanuthones, de la mélanine, de la PR-toxine et des andrastines. (1) *Variation de la source de carbone.* ***(2)*** *Variation de la source d'azote. Les conditions encadrées correspondent aux valeurs par défaut obtenues lors du chargement d'****iPrub22*** *(i.e. utilisation du glucose et de l'ammoniac comme source de carbone et d'azote), présentées en* ***Figure 3-6****. Le classement des conditions est effectué selon les flux maximaux de production potentielle croissants des métabolites écrits en gras et soulignés.*

Contrairement à la voie des pénicillines nous observons lors de la variation des sources de carbone, 8 comportements différents puisque désormais l'utilisation de l'homocystéine ou de la méthionine n'exprime plus des flux différents. En revanche, l'ordre des conditions (*i.e.* classement par flux croissant) est relativement similaire pour les produits finaux de la voie de biosynthèse des roquefortines et pour la production de ferrichromes. Pour les PKS étudiés, tels que la versicolorine B et l'avérufine, ainsi que pour les précurseurs 3-méthylphénol, YWA1, DMOA, et l'aristolochène (*i.e.* précurseurs de divers PKS pour les trois premiers éléments et de terpènes pour le dernier), le classement des quatre dernières conditions les plus performantes s'effectue par nombre de carbones croissant (*i.e.* $C_5H_{5|10}O_5$; $C_6H_{10|12|14}O_6$; $C_{12}H_{22}O_{11}$; $C_{18}H_{22}O_{16}$).

L'utilisation du glucose, la source de carbone que nous utilisons comme référence, appartient aux conditions les plus favorables pour stimuler l'ensemble des voies étudiées et le métabolisme basal. En revanche, tout comme pour la maximisation de la biomasse, l'utilisation du lactate, du méthanol, de l'homocystéine, de l'α-cétoglutartate du propionate et du pyruvate sont des sources de carbone de très faible intérêt pour la sollicitation du métabolisme spécialisé. De plus, les sucres maltotriose et raffinose ne sont pas des éléments à considérer pour l'optimisation de la production de peptides non-ribosomaux, mais seront en revanche à privilégier, d'une part, pour maximiser la production de biomasse, et d'autre part, la production de polycétides ou de leurs précurseurs.

Toutefois, et de manière générale, lorsqu'il est couplé au glucose, l'ammoniac ne semble pas être une source d'azote optimale pour la production des métabolites spécialisés étudiés. En revanche, les acides aminés tyrosine et phénylalanine sont les sources d'azote pour lesquelles les productions maximales potentielles des polycétides et de leurs précurseurs est maximale, tandis que les productions des peptides non-ribosomaux semblent être favorisées par la glutathione, l'arginine, la lysine ou l'adénosine. Selon les voies étudiées, nous observons entre 14 et 19 groupes de comportements différents en fonction de la source d'azote usitée.

En fonction de la modification de la source de carbone et de la source d'azote, les productions maximales théoriques de flux sont améliorées d'un facteur respectif de :

		Facteur d'augmentation (×)	
Voies de biosynthèse	**Type de molécule**	**Variation de la source de Carbone**	**Variation de la source d'Azote**
Roquefortine	Précurseurs	× 9,2	× 1,4
Roquefortine	Intermédiaires	× 11,0	× 1,2
Roquefortine	Produit finaux	× 11,5	× 1,2
Ferrichrome	Produit final	× 6,9	× 4,7
Ferrichrome A	Produit final	× 10,2	× 2,7
Avérufine	Produit final	× 76,8	× 1,5
Versicolorine B	Produit final	× 59,5	× 1,5
Toluquinol	Précurseurs	× 26,0	× 1,5
Isoepoxydon	Produit final	× 76,8	× 1,5
Aristolochène	Précurseurs	× 74,1	× 1,2
DMOA	Précurseurs	× 76,8	× 1,2

Ces divers chiffres indiquent que la production de PKS et de leurs précurseurs est hautement influencée par le choix de la source de carbone. À titre d'exemple, remplacer le glucose par le maltotriose ou le raffinose permettrait d'augmenter les flux potentiels d'un facteur compris entre 2,1 et 6,3.

Enfin et contrairement aux résultats présentés en **Figure 3-6** qui laissaient penser qu'il n'existait pas de variations intra voies pour les composés de la voie des roquefontines, des yanuthones et de l'isopeoxydon et des ferrichromes, nous constatons l'existence d'au moins une condition (*e.g.* quinate) où les flux maximaux potentiels de ces métabolites diffèrent.

En conclusion, ces analyses démontrent que la production de métabolites spécialisés par le modèle ***i*Prub22** est, d'une part, sensible aux variations des conditions environnementales et, d'autre part, exprime des comportements différents selon les différentes voies de biosynthèse étudiées. En effet, selon la classe de métabolites considérée (*i.e.* NRPS, PKS, hybrides), nous observons des conditions préférentielles pour maximiser la production de chacun de ces types de métabolites. Ces travaux ouvrent donc la voie à une exploration rationnelle **OSMAC *in silico*** des conditions de culture pour maximiser la production de métabolites spécialisés, et ce à grande échelle

.

CHAPITRE 4 :

Le modèle *i*Prub22, Évolution, Pérennité ou Obsolescence Programmée ?

Pour ce dernier chapitre, nous souhaitons clore notre réflexion en abordant deux aspects indépendants relatifs, d'une part, à l'évolution potentielle d'***i*Prub22**, et d'autre part, à la pérennisation des méthodes présentées dans ce manuscrit. Plus précisément, **(1)** nous discuterons de la modélisation de la compartimentation intracellulaire au sein des **GSMNs** eucaryotes, une réflexion ayant été entamée à cet égard lors de la reconstruction d'***i*Prub22**, mais n'ayant pu être finalisée. Nous commenterons ensuite brièvement **(2)** la possibilité d'intégrer automatiquement les voies du métabolisme spécialisé dans les futures reconstructions et nous nous intéresserons **(3)** au pouvoir prédictif des comparaisons de **GSMNs**, notamment au travers des reconstructions. Enfin, ayant eu à notre disposition deux assemblages d'une même souche de *Penicillium rubens/chrysogenum* marin, **(4)** nous tâcherons d'évaluer la faisabilité, et le cas échant, la pertinence d'une reconstruction de **GSMN** pour ce nouvel organisme.

1 - Un GSMN d'eucaryote sans compartimentation intracellulaire

Nous avons désormais établi que le **GSMN** que nous présentons, ***i*Prub22**, exprime pleinement les potentialités du génome de *P. rubens* et correspond, de ce fait, à un réseau global du métabolisme de l'organisme. Nous avons veillé à ce que notre modèle **GSMN** réponde aux prérequis actuels, à court, moyen et long terme afin qu'il puisse être considéré comme un modèle de qualité. Néanmoins, des améliorations, tant sur la reconstruction que sur le modèle qui en découle, demeurent nombreuses. Ce type de modélisation constitue certes une abstraction de la réalité biologique, mais les questions et conséquences relatives à l'absence de modélisation intracellulaire doivent être soulevées.

1.1. L'organisation modulaire des eucaryotes

La compartimentation intracellulaire chez les eucaryotes est une caractéristique fondamentale qui permet à la cellule d'isoler différentes réactions métaboliques dans des structures spécialisées, appelées organites (**Figure 4-1**). Chaque organite possède une fonction spécifique et un environnement biochimique unique, favorisant l'efficacité des processus métaboliques et protégeant la cellule des réactions potentiellement nocives. Ce cloisonnement assure une régulation précise des activités cellulaires, tout en permettant l'interaction coordonnée entre les différents compartiments *via* des mécanismes de transport. Ainsi, présenter un modèle non compartimenté sacrifie le réalisme biologique, mais simplifie également le système en traitant la cellule comme une entité homogène unique.

Notons que dans cette section l'emploi du terme compartiment fait référence, tantôt à l'élément biologique (*i.e.* une partie spécialisée d'une cellule délimitée par un réseau de membranes), tantôt à la représentation abstraite de ce concept tel qu'encodé dans un **GSMN** (*e.g.* limites du système, milieu extracellullaire, espace intermembrannaire, *etc.*). Par conséquent, la question de la compartimentation subcellulaire ne se limite pas exclusivement aux organismes eucaryotes, puisque, dans une moindre mesure certes, les **GSMNs** de procaryotes sont également compartimentés *(Liu et al. 2014)*.

La compartimentation intracellulaire dans un **GSMN** peut être considérée comme une couche d'annotation supplémentaire essentielle pour affiner la modélisation métabolique. *Saccharomyces cerevisiae* est le premier organisme eucaryote pour lequel une reconstruction tenant compte de la compartimentation des réactions a été réalisée *(Förster et al. 2003)*. Paradoxalement, cet aspect reste rarement mis en avant dans les publications, et lorsque c'est le cas, les informations accessibles sont souvent limitées *(Mintz-Oron et al. 2009)*. Pour souligner toutefois l'importance de cette caractéristique, nous citons quelques exemples de **GSMNs** développés pour des organismes modèles historiques : *H. sapiens (Swainston et al. 2016)*, *A. thaliana (de Oliveira Dal'Molin et al. 2010; Mintz-Oron et al. 2012)*, *M. musculus (Sigurdsson et al. 2010)*, *C. elegans (Gebauer et al. 2016; Yilmaz et Walhout 2016)* et *D. rerio*. Il est important de noter que, pour ce dernier sa première version était dépourvue d'associations GPR *(Bekaert 2012)*, une couche essentielle qui y sera ajoutée dans sa version suivante, soit sept ans plus tard *(van Steijn et al. 2019)*.

Outre la volonté de présenter un modèle au plus proche de la réalité, nous supposions que cette amélioration aurrait pu nous aider lors des analyses futures, notamment en abordant le concept des modules *(Ma et al. 2004; Müller et Bockmayr 2014; Reimers 2015)*. Cependant, peu d'études se sont concentrées sur la comparaison de modèles compartimentés ou non afin de quantifier l'impact de leurs différences *(Klitgord et Segrè 2010; Waller et al. 2020)*.

La modularité, considérée comme un pilier de l'organisation biologique est un concept abstrait visant à réconcilier les différents aspects structurels et fonctionnels d'un organisme *(Wagner et al. 2007)*. Par essence, ce concept intègre une large gamme de données hétérogènes dont la compréhension permet de mettre en évidence des sous-systèmes distincts qui fonctionnent de manière relativement indépendante tout en contribuant au fonctionnement global de l'organisme. Les modèles de connectivité résultants permettent de mettre en évidence des modules, des groupes faiblement connectés entre eux et dont chacun comprend des sous-ensembles d'un même type, qui eux sont fortement connectés *(Wagner et al. 2007)*. Ces modules peuvent être des structures physiques (*e.g.* organite, cellule, tissus, organe, organisme, population) ou des fonctions biochimiques (*e.g.* voies métaboliques), et leur hétérogénéité reflète la complexité et la spécialisation des systèmes biologiques.

En-dehors des aspects évoqués ci-dessus, diverses interrogations liées à la reconstruction et à la compartimentation cellulaire peuvent être soulevées. Par exemple, quelle proportion du protéome est couverte par ***i*Prub22**, et comment cette couverture évolue-t-elle en fonction des organites ? Comment les réactions liées à ces gènes se répartissent-elles entre les différents compartiments ? De plus, existe-t-il, une spécificité des produits géniques, et donc des réactions, vis à vis de certains organites ? Enfin, les données disponibles mettent-elles en évidence une prédominance de certains compartiments par rapport à d'autres ?

Ainsi, afin de quantifier la pertinence des éléments précédents, nous proposons d'analyser la compartimentation des réseaux métaboliques préexistants, à savoir *i*AL1006 et Prubens. Ensuite, nous concentrerons notre attention sur les informations disponibles au sein d'***i*Prub22**.

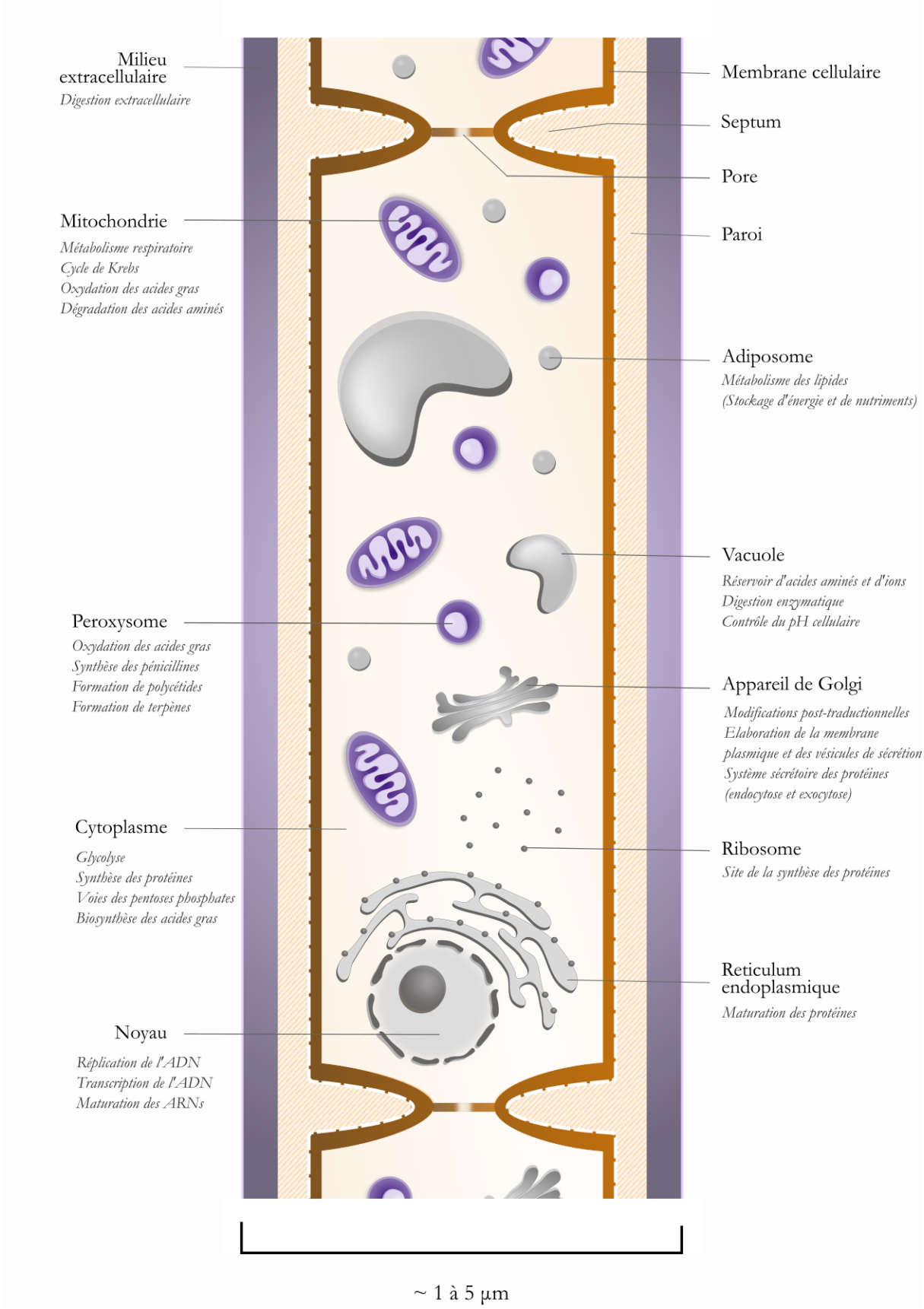


Figure 4-1 : La compartimentation des voies métaboliques, des organites aux fonctions spécifiques. Illustration sur une cellule théorique de champignons filamenteux. *Adapté à partir des informations de (Nielsen 1997), (Raven et al. 2011) et (Richard et al. 2011).*

1.2. Les GSMNs antérieurs de *Penicillium rubens* présentent une compartimentation intracellulaire, une caractéristique perdue chez *i*Prub22

1.2.1. La compartimentation intracellullaire de *i*AL1006 et Prubens

Dans un **GSMN**, l'information sur la compartimentation est une étiquette associée aux métabolites. Cette qualification permet alors de distinguer deux catégories de réactions :

- **Les réactions S :** où les réactants et les produits appartiennent au même compartiment.
- **Les réactions T** : où les réactants et les produits appartiennent à des compartiments différents.

L'assignation de compartiments aux métabolites des modèles antérieurs résulte de processus semi-automatisé pour *i*AL1006 et automatique pour Prubens. Les informations relatives à la compartimentation de ces deux **GSMNs** sont présentées dans le **Tableau 4-1**. À noter que, dans la suite de cette section, nous considérerons les limites du système, connues sous le nom de « *boudaries* », comme un compartiment à part entière. Ainsi, l'expression réactions de transport couvre également les processus d'échange inclus dans les divers modèles.

L'ajout de la couche de compartimentation à la reconstruction *i*AL1006 provient du processus semi-automatique inclus dans la RAVEN Toolbox. L'algorithme utilisé par les auteurs avait pour objectif d'assurer la cohérence entre la détection de peptides signaux, les propriétés physico-chimiques des protéines et le nombre de réactions de transport nécessaires, afin de garantir la fonctionnalité et la connectivité du réseau. En outre, *i*AL1006, bénéficie d'une intense curation manuelle comme nous l'avons mentionné antérieurement. Ainsi, cette reconstruction propose une compartimentation en cinq éléments distincts : **(1)** le cytosol, **(2)** le milieu extracellulaire, **(3)** la mitochondrie, **(4)** le peroxysome et **(5)** les limites du système. Les métabolites localisés dans le cytoplasme conservent leurs identifiants d'origine, tandis que les autres sont caractérisés par la présence d'un suffixe correspondant à la première lettre du compartiment concerné (*e.g. :* l'alpha-D-glucose existe sous les formes `GLC`, `GLCe` et `GLCb` ce qui implique que ce composé est à la fois présent dans le cytoplasme, dans le milieu extracellulaire et qu'il est défini pour les réactions d'échange du modèle). Cette nomenclature permet d'isoler les réactions de transport qui représentent 29 % de l'ensemble des réactions.

La reconstruction Prubens, quant à elle, s'appuie sur l'annotation automatique liée à l'outil PahwayTools et à la collection de bases de données qui lui est rattachée. L'ontologie des composants cellulaires qui y est utilisée (*i.e. « Cell component Ontology »* – CCO *(Zhang et al. 2005)*) comprend plus de 160 termes qui décrivent les composants et les compartiments cellulaires, ainsi que les relations entre ces termes. L'information de compartimentation, qui est toujours portée en suffixe des métabolites, est suffisamment explicite pour être comprise à la simple lecture. Pour reprendre l'exemple du glucose, ce composé coexiste sous quatre formes au sein de Prubens : `ALPHA-GLUCOSE_CCO-BOUNDARY`, `ALPHA-GLUCOSE_CCO-CYTOSOL`, `ALPHA-GLUCOSE_CCO-EXTRACELLULAR`, et `ALPHA-GLUCOSE_CCO-PERI-BAC`. Ce dernier

terme correspond à la « région située entre la membrane interne (cytoplasmique) et la membrane externe des bactéries Gram-négatives ». Comme tout ensemble de vocabulaire contrôlé, la CCO a pour visée de faciliter la sélection et la visualisation des données, basée ici sur la localisation subcellulaire des enzymes. Toutefois, cette ontologie, rattachée à un type de bases de données spécifiques, semble peu usitée de nos jours. Au sein de Prubens, nous dénombrons 14 étiquettes de compartiments différents. Cependant, la proportion de réactions de transport de 8 % est nettement inférieure à celle d'*i*AL1006. À titre de comparaison, ***i*Prub22** où seul le cytosol, le milieu extracellulaire et les phénomènes d'export sont modélisés, comprend 13 % de réactions qui sont associées à un phénomène de transport. De plus, il est à noter que la modélisation des peroxysomes de Prubens semble sous représentée par rapport aux données d'*i*AL1006, et que 5,6 % des composés possèdent une localisation subcellulaire discutable en termes de réalité biologique.

Tableau 4-1 : Description de la compartimentation au sein des GSMNs *i*AL1006 et Prubens. Les cellules colorées dans le réseau de 2018 mettent en évidence une portion de 5,6 % de composés dont l'annotation automatique suscite des interrogations quant à la cohérence biologique.

Nombre de réactions			
***i*AL1006**		***Prubens***	
Nombre de réactions S	1162	Nombre de réactions S	2 374
Nombre de réactions T	470	Nombre de réactions T	200
Total	1 632	**Total**	2 574

Nombre de métabolites par compartiment			
***i*AL1006**		***Prubens***	
Cytosol	1 762	Cytosol	2 529
Mitochondrie	242	Membrane interne mitochondriale	9
		Lumière mitochondriale	200
Peroxysome	105	Membrane peroxysomale	6
Espace extracellulaire	160	Espace extracellulaire	83
« Boundary »	160	« Boundary »	13
		Extérieur*	11
		Intérieur*	17
		Extérieur/Intérieur*	2
		Lumière du réticulum endoplasmique rugueux	17
		Paroi cellulaire (Bactérie Gram négatives)	6
		Périplasme (Bactérie Gram négatives)	163
		Lumière de l'appareil de Golgi (Bactérie Gram négatives)	1
		Stroma du chloroplaste	1
Total	2 429	**Total**	3 058

* Espace cellulaire générique utilisé pour représenter les espaces de part et d'autre des membranes lors des évènements de transport.

1.2.2. L'annotation fonctionnelle du protéome pour guider la filtration des données

Dans le cadre de notre reconstruction, afin d'intégrer la couche d'annotations de localisation subcellulaire nous avions envisagé de : **(1)** tirer profit des informations contenues dans les réseaux antérieurs (**Tableau 4-1**), **(2)** d'extraire les annotations fonctionnelles relatives au protéome, **(3)** de réaliser un consensus des données puis **(4)** de corriger les incohérences. Concernant l'extraction des données issues de l'annotation fonctionnelle du protéome de *P. rubens*, trois outils différents et complémentaires ont été utilisés (cf. Chapitre 2, section 3.2 *Annotation fonctionnelle et compartimentation subcellulaire*, page 99) :

- **SignalP** *(Petersen et al. 2011)* (v4.1) prédit l'existence de signaux peptides et des sites de clivage au sein des séquences protéiques en s'appuyant sur les méthodes de réseaux neuronaux (*i.e. « neural network based method »*). Le peptide signal de sécrétion est un signal de triage de protéines ubiquitaires autorisant notamment la translocation à travers la membrane du réticulum endoplasmique.
- **TMHMM** *(Krogh et al. 2001)* (v2.0) sert à la prédiction de la structure des protéines, et plus précisément, détecte, à partir de profils de modèles de Markov caché, l'existence ou non d'hélices transmembranaires à hauteur de 97-98 %. En moyenne, 20 à 30 % des gènes codent pour des protéines transmembranaires.
- **DeepLoc** *(Almagro Armenteros et al. 2017)* (v1.0) effectue une prédiction basée sur du « *deep learning* » de la localisation subcellulaire des protéines eucaryotes. L'information contenue dans les séquences permet une assignation des produits géniques à travers 10 compartiments avec une précision de 78 %.

Avec l'annotation fonctionnelle, les informations dont nous disposons se rapportent à la compartimentation des produits géniques, en l'occurrence celle des enzymes. La plupart d'entre elles agissent de manière non liée à une structure cellulaire spécifique et évoluent librement dans le cytosol. Ainsi, lorsque l'assignation d'un métabolite est inconnue ou discutable, il est admis qu'il appartient au cytosol. Concernant les produits géniques qui font partie intégrante des membranes, les identifier permet de cibler les réactions impliquées dans les processus de transport.

L'analyse descriptive que nous proposons s'effectue graduellement. Dans un premier temps, nous cherchons à caractériser la localisation subcellulaire des séquences du protéome de *P. rubens* à l'aide de l'outil DeepLoc. Cet outil nous permet d'évaluer la couverture du protéome de *P. rubens* en fonction de l'assignation par compartiment. Les résultats révèlent que 31 % des protéines sont localisées dans le cytoplasme, 24 % dans le noyau, 14 % dans les mitochondries, 9 % sont associées aux membranes cellulaires, 6 % se trouvent dans l'espace extracellulaire et 12 % se répartissent entre le réticulum endoplasmique, les peroxysomes, les vacuoles et les appareils de Golgi. Nous constatons également que 431 séquences sont associées aux plastides, un organite absent chez les champignons filamenteux, et, selon les éléments énoncés précédemment, seront donc associés au cytosol. Le **Tableau 4-2** présente le nombre de séquences protéiques assignées à chacun des compartiments et la **Figure 4-2** permet de visualiser les proportions de gènes inclus ou non dans ***i*Prub22**.

Tableau 4-2 : Localisation subcellulaire des séquences protéiques de *Penicillium rubens* Wisconsin 54-1255 selon les résultats de DeepLoc.

Compartiment	Description [1]	Nombre de séquences protéiques assignées par compartiment [2]	
Cytoplasme	*Cytoplasme (cytosol et cytosquelette)*	3 937	**2 070**
Nucleus	*Enveloppe, membrane interne et externe, matrice, lamina, chromosome, nucléoles*	2 959	**910**
Mitochondrie	*Enveloppe, membrane interne et externe, matrice espace intermembranaire*	1 731	**619**
Membrane cellulaire	*Apical, apicolatéral, basal, basolatéral, latéral, membrane cellulaire, projection cellulaire*	1 179	**576**
Extracellulaire	*Extracellulaire*	790	**375**
Réticulum endoplasmique (RE)	*Membrane et lumière du RE, microsome, RE rugueux, RE lisse, RE sarcoplasmique*	633	**327**
Peroxysome	*Matrice et membrane des peroxysomes*	572	**438**
▲ Plastide	*Membrane plastidienne, stroma et thylakoïde*	431	**194**
▲ Lysosome/Vacuole	*Vacuoles contractiles, lytiques et de stockage des protéines, lumière et membranes des vacuoles et des lysosomes*	202	**100**
Appareil de Golgi	*Membrane et lumière de l'appareil de Golgi*	122	**63**
	TOTAL	12 556	**5 672***

[1] Descriptions issues de *(Almagro Armenteros et al. 2017)*.

[2] Les séquences présentes dans ***i*Prub22** sont matérialisées en gras.

▲ Les champignons sont dépourvus de plastides et de lysosomes. Les premiers sont spécifiques aux plantes et algues et les seconds aux cellules animales.

* Le modèle ***i*Prub22** comporte un total de 5 703 séquences génomiques et 5 672 d'entre elles proviennent des données téléchargées sur Ensembl. Les 31 séquences de différences récupérées sur UniProt sont localisées pour 12 d'entre elles dans le cytoplasme, 4 dans le réticulum endoplasmique, 3 dans le milieu extracellulaire, 7 dans la mitochondrie, 3 dans le noyau et 2 dans le peroxysome.

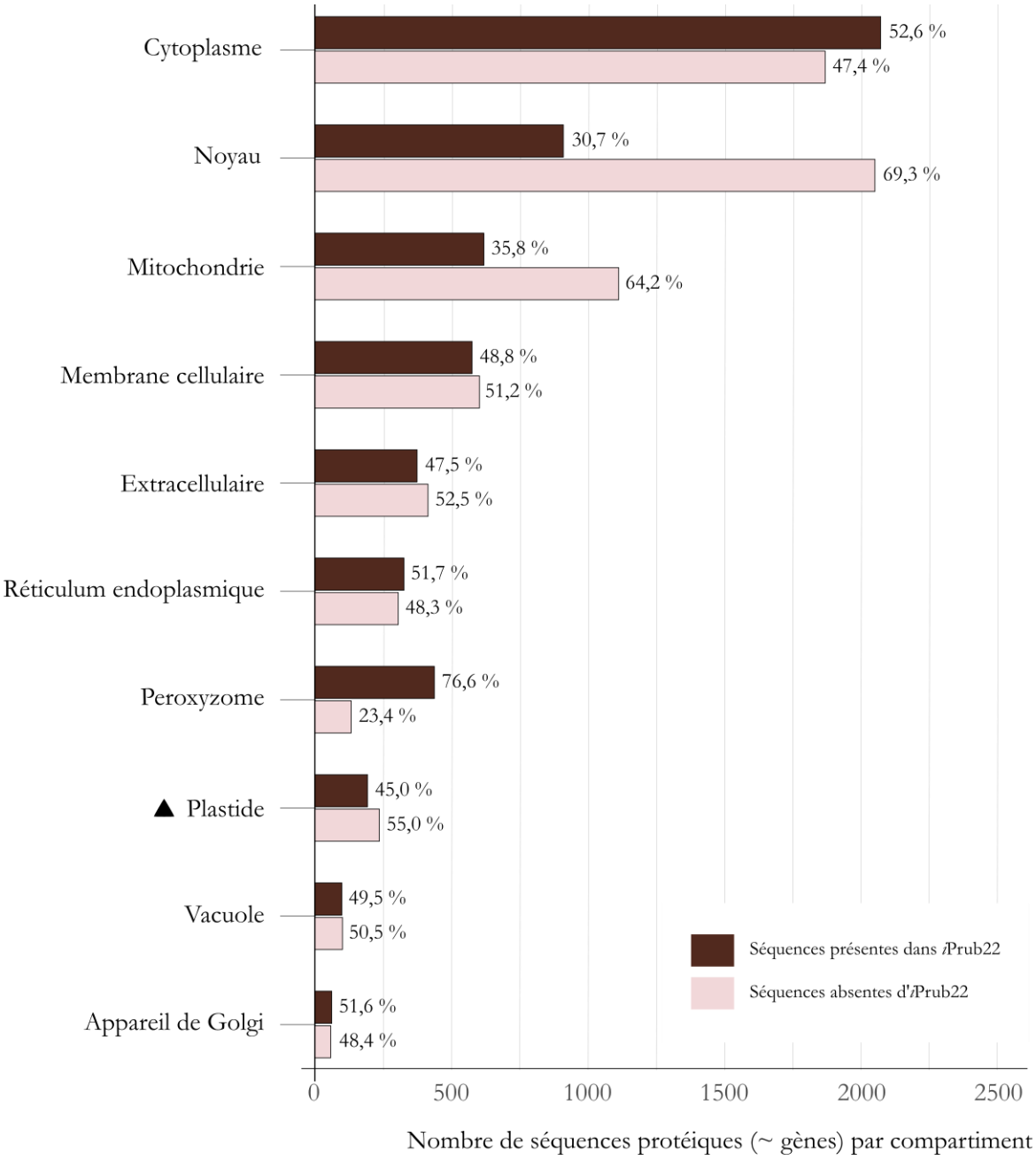


Figure *4-2 : Histogramme empilé représentant la répartition des produits géniques à travers les différents compartiments cellulaires en fonction des gènes retrouvés ou non au sein d'i*Prub22. *Cette représentation permet de visualiser la part de couverture métabolique propre à chaque organite. ▲ Les champignons sont dépourvus de plastides, des organistes qui sont spécifiques aux plantes et algues.*

Ensuite, en quantifiant, pour chaque organite ou compartiment cellulaire, la proportion de séquences protéiques présentes dans ***i*Prub22** par rapport à celles absentes, nous observons que, hormis le noyau et les peroxysomes, la couverture métabolique des organites est relativement équilibrée. En première instance, si nous nous référons à la **Figure 4-1**, ces résultats nous semblent cohérents avec les attentes biologiques. D'une part, la machinerie nucléaire est essentiellement liée aux phénomènes de transcription et de régulation, des éléments faiblement représentés dans la modélisation du métabolisme. D'autre part, les peroxysomes, qui proportionnellement sont les organites les mieux caractérisés, sont le lieu prépondérant des voies de biosynthèse des produits naturels et de leurs précurseurs. Ce dernier point, vient ainsi corroborer la curation du métabolisme spécialisé que nous avons effectué.

La seconde étape consiste à affiner les annotations des produits géniques. Nous nous attendons à ce que les résultats de TMHMM couplés éventuellement à ceux de SignalP mettent en relief les séquences associées à du transport cellulaire, et viennent appuyer et confirmer les résultats de DeepLoc. Ainsi, en analysant préférentiellement les réactions associées à ces gènes, nous espérons pouvoir corriger plus aisément les potentielles lacunes résultant d'une compartimentation automatique. Le protéome de *P. rubens* est composé de 1*2 556* séquences et il existe une prédiction de la présence d'un domaine transmembranaire ou de la présence d'un signal-peptide pour *2 283* (18 %) et 921 (7 %) d'entre elles. Respectivement, 992 (43 %) et 452 (49 %) de ces séquences sont retrouvées au sein d'***i*Prub22**. Associés à la localisation putative des séquences ces résultats sont présentés en **Figure 4-3**. Une fois encore, les résultats semblent en adéquation avec les attentes biologiques. Les produits géniques présentant des domaines transmembranaires se répartissent majoritairement et logiquement entre la membrane cellulaire, le réticulum endoplasmique, les vacuoles (*i.e.* organites de stockage) et les appareils de Golgi (*i.e.* lieu d'adressage des protéines). Ceux annotés par SignalP et possédant *de facto* un signal-peptide sont retrouvés dans le milieu extracellulaire (*i.e.* des protéines synthétisées au sein de la cellule et exportées permettant l'absorption extracellulaire effectuée par les champignons filamenteux). Enfin, il est raisonnable de penser que l'ensemble des séquences non annotées concernent des gènes codant pour des réactions internes (*i.e.* sans implication dans les processus de transport), où la transformation des substrats en produits se fait dans le même compartiment.

Une fois que nous avons établi une image globale concernant l'annotation des gènes, qu'en est-il pour les réactions associées à ces gènes ? En ayant extrait pour chaque gène, sa ou ses réactions associées, nous présentons en **Figure 4-4** la répartition théorique des réactions d'***i*Prub22** à travers les différents compartiments étudiés. Selon cette représentation, nous constatons que suivant le compartiment considéré le nombre médian de réactions associées par gène varie entre 1,5 pour les vacuoles et 6 pour les peroxysomes. Même si nous savons que dans les systèmes biologiques, les relations un gène–une réaction sont peu communes, ces valeurs nous semblent relativement élevées et pourraient être le témoin d'une faible spécificité des associations GPR d'***i*Prub22**.

La question de l'exactitude et, plus précisément, de la spécificité des associations GPR a été une préoccupation récurrente tout au long de nos travaux. À l'échelle d'un **GSMN**, et à notre connaissance, il semble impossible de quantifier précisément ces deux paramètres. De surcroît, dans l'hypothèse où les associations GPR présenteraient une très faible spécificité, envisager de les corriger dépasse largement les objectifs de cette thèse. Fondamentalement, c'est la présence de faux positifs (*i.e.* gènes associés à tort à une réaction) qui peuvent être potentiellement dommageables pour la structure du réseau. Cela est particulièrement problématique lorsque des réactions existent uniquement en raison de tels faux positifs. Toutefois, cet impact doit être nuancé lorsque des faux positifs coexistent avec des vrais positifs dans un même ensemble GPR. En effet, si nous nous concentrons exclusivement sur l'aspect de la reconstruction de **GSMNs**, ces éléments provenant principalement de recherches d'homologie peuvent aider à comprendre certains concepts évolutifs, et ainsi orienter les axes de recherche en relation. Ainsi, le temps consacré à la correction d'une reconstruction ou d'un modèle, ainsi que le type de curation entrepris, dépendra de l'objectif initial ayant conduit à la genèse du **GSMN**.

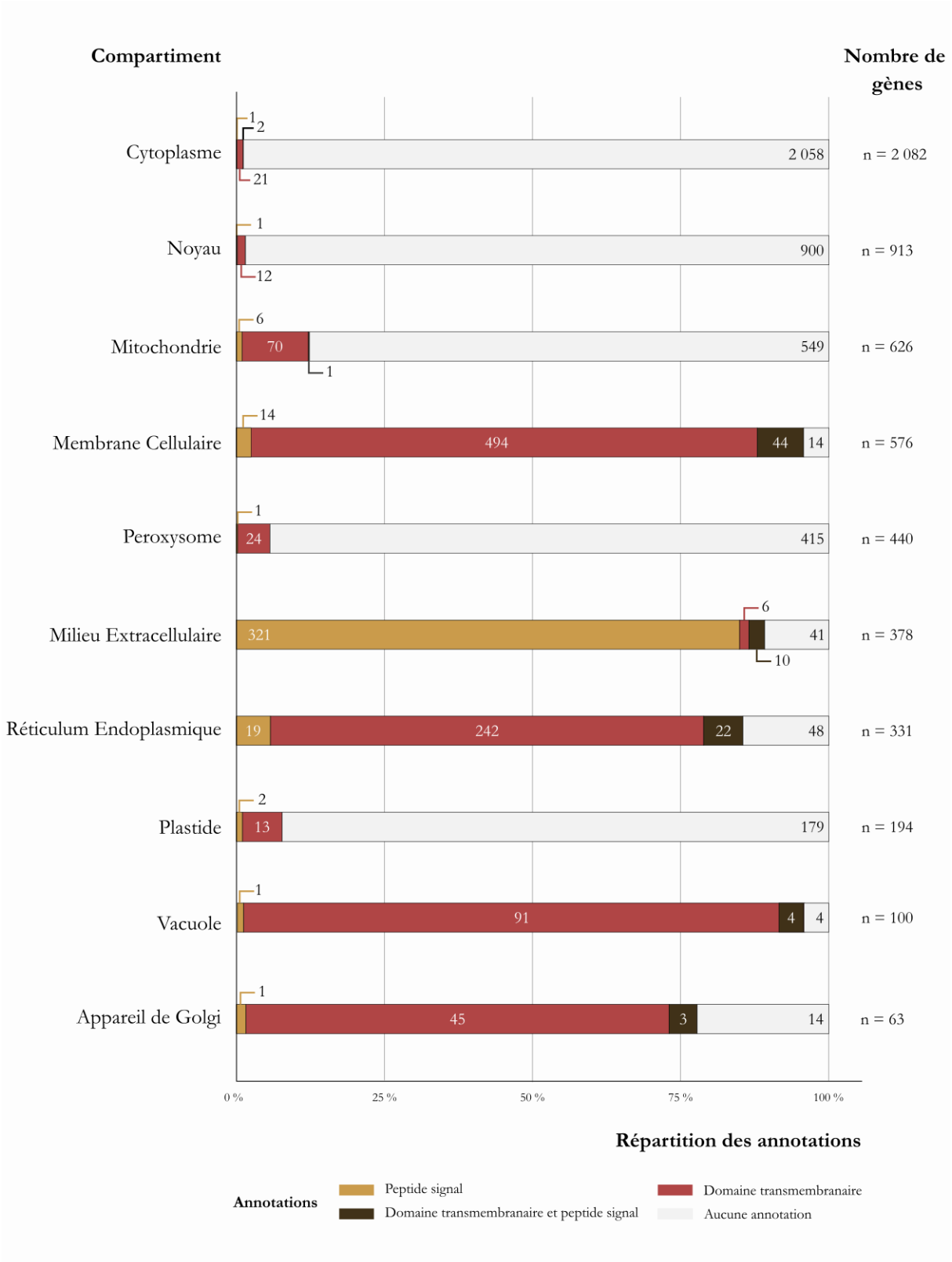


Figure 4-3 : Diagramme à barres empilées en pourcentage illustrant les annotations des produits géniques des gènes présents dans iPrub22 en relation avec leur localisation cellulaire. *L'axe des abscisses affiche les pourcentages de chacune des catégories d'annotation pour chaque type de compartiment. Le nombre de séquences concernées par ces annotations est inscrit à l'intérieur de chaque barre. Résultats déterminés avec DeepLoc (v1.0).*

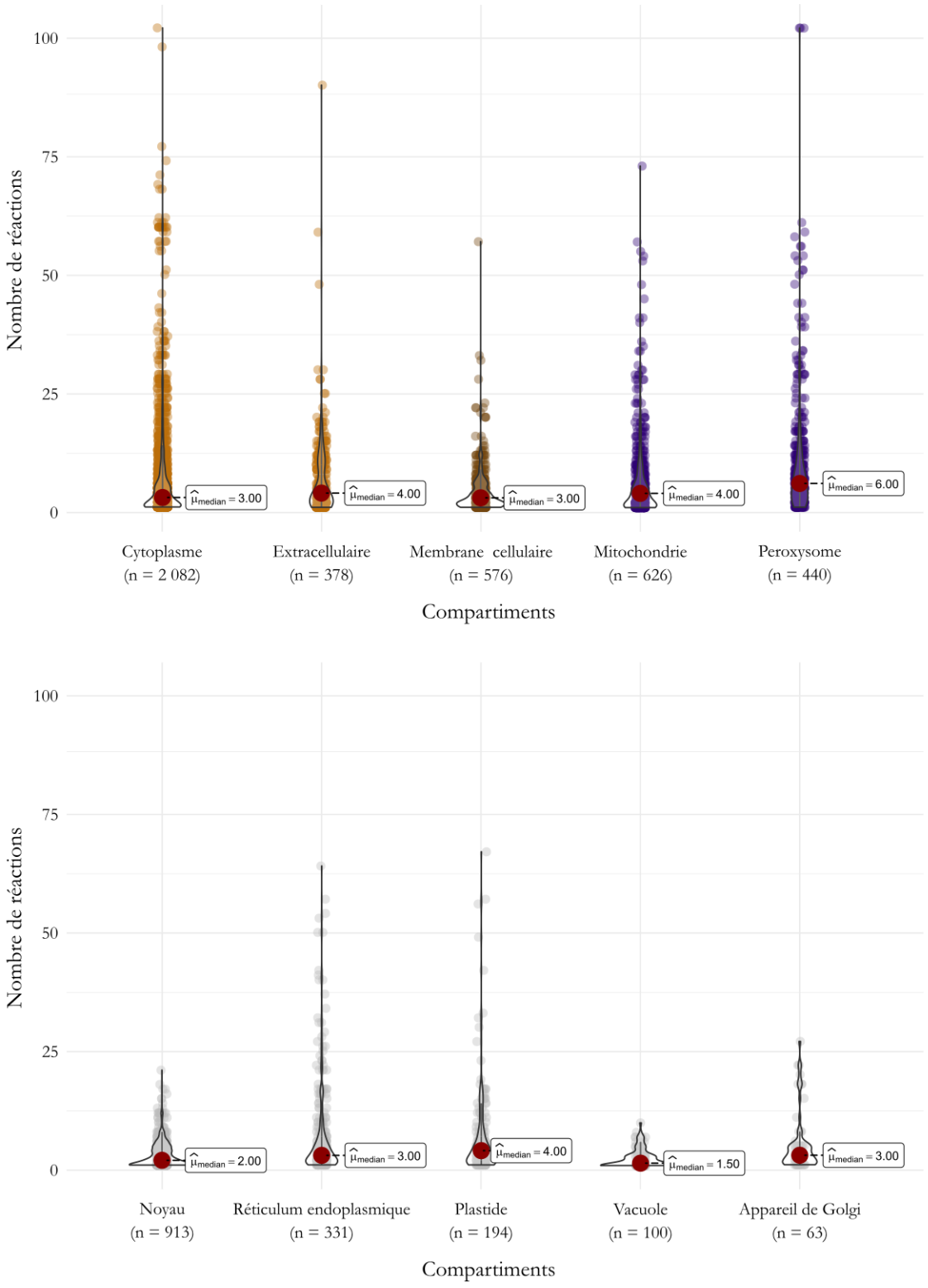


Figure 4-4 : Distribution du nombre de réactions associées par gène selon les différents compartiments cellulaires. *Chaque point sur ce graphique en violon symbolise un gène, et le nombre de réactions dans lesquelles il est impliqué, se lit en ordonnée. Les compartiments cytoplasme et milieu extracellulaire, affichés en orange (●), sont les deux compartiments modélisés au sein d'**iPrub22**. L'annotation membrane cellulaire, matérialisée en marron (●), a permis d'affiner les associations GPR lors de la curation des réactions de transport entre les deux compartiments mentionnés précédemment. En violet (●), sont représentés les organites mitochondries et peroxysomes, compartiments les plus pertinents à modéliser chez Penicillium rubens Wisconsin 54-1255, et sur lesquels nous nous sommes concentrés préférentiellement. Enfin, les données des compartiments dont les gènes sont colorés en gris (●) n'ont pas été analysées plus en détails.*

Le dernier point consiste désormais à visualiser l'ensemble des informations présentées jusqu'alors (*i.e.* recoupement des résultats des analyses SignalP, TMHMM et DeepLoc) avec cette fois-ci une vision centrée sur les réactions. La **Figure 4-6** représente la répartition théorique des réactions en fonction des compartiments et donne un aperçu des modifications à entreprendre. Cette visualisation permet de filtrer les données en vue d'effectuer et d'adapter ensuite les différents traitements nécessaires à l'encodage de la compartimentation intracellulaire (cf. section *1.2.4 Procédure envisagée pour la modélisation de la compartimentation intracellulaire et causes de son abandon*, page 369). Les configurations où tous les gènes d'une même association auraient la même localisation subcellulaire sont les plus simples à traiter, et 45 % des réactions d'***i*Prub22** correspondent à cette situation. Les 55 % restants correspondent à des réactions ayant lieu dans des compartiments multiples qui devront dans les cas les plus simples être dupliquées. La **Figure 4-5** synthétise et simplifie les relations entre gènes et réactions pouvant exister (*i.e.* ensemble non bijectif).

Cas 1. Un gène annote une seule réaction ou une réaction est annotée par un seul gène. Renseigne sur la spécificité et l'unicité des entités concernées. Ces situations sont les plus simples à traiter.

Cas 2. Un gène annote deux réactions. Informe soit sur la multi-fonctionnalité de l'enzyme, soit met en relief une annotation en numéro EC peu spécifique.

Cas 3. Deux gènes annotent la même réaction et leurs produits géniques sont localisés dans le même compartiment. Fournit, soit un éclaircissement potentiel sur la redondance et la duplication de gènes (*i.e.* modélisation des isoenzymes), soit souligne l'existence de complexes enzymatiques.

Cas 4. Deux gènes annotent la même réaction, mais leurs produits géniques sont localisés dans des compartiments différents. Dans le cadre de notre tentative de modélisation intracellulaire, ce type de configuration induit des modifications profondes dans l'encodage des informations au sein du **SBML**.

Cas 5. Les situations précédentes font état des exemples les plus simples retrouvés dans le réseau. Elles peuvent être généralisées pour éclairer des cas plus complexes et enrichir la compréhension des associations GPR.

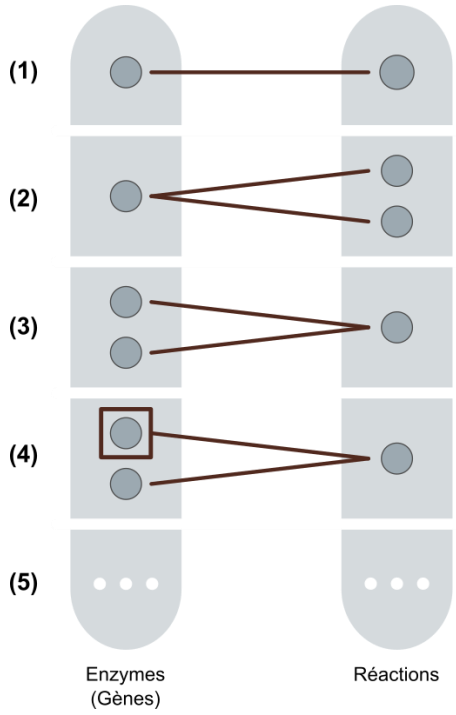


***Figure 4-5 : Représentation simplifiée d'un échantillon des diverses associations GPR présentes dans le* GSMN.** *Ces associations offrent des informations sur le degré de spécificité des gènes et/ou des réactions. Le cadre marron symbolise un gène dont le produit génique possède une localisation subcellulaire différente des autres.*

À titre d'exemple, considérons **le Cas 4** et admettons que la localisation subcellulaire de leur produit génique soit assignée au cytosol et à la mitochondrie. Si nous désirons modéliser cette possibilité, nous devons dupliquer cette réaction présente initialement dans le cytosol en générant un nouvel identifiant (*i.e.* ancrage) et s'ils ne sont pas déjà présents, dupliquer l'ensemble des réactants et produits de cette réaction pour les assigner à la mitochondrie (*i.e.* $Rxn\_S\_1 = A_c + B_c \rightarrow C_c + D_c$ et $Rxn\_S\_2 = A_m + B_m \rightarrow C_m + D_m$).

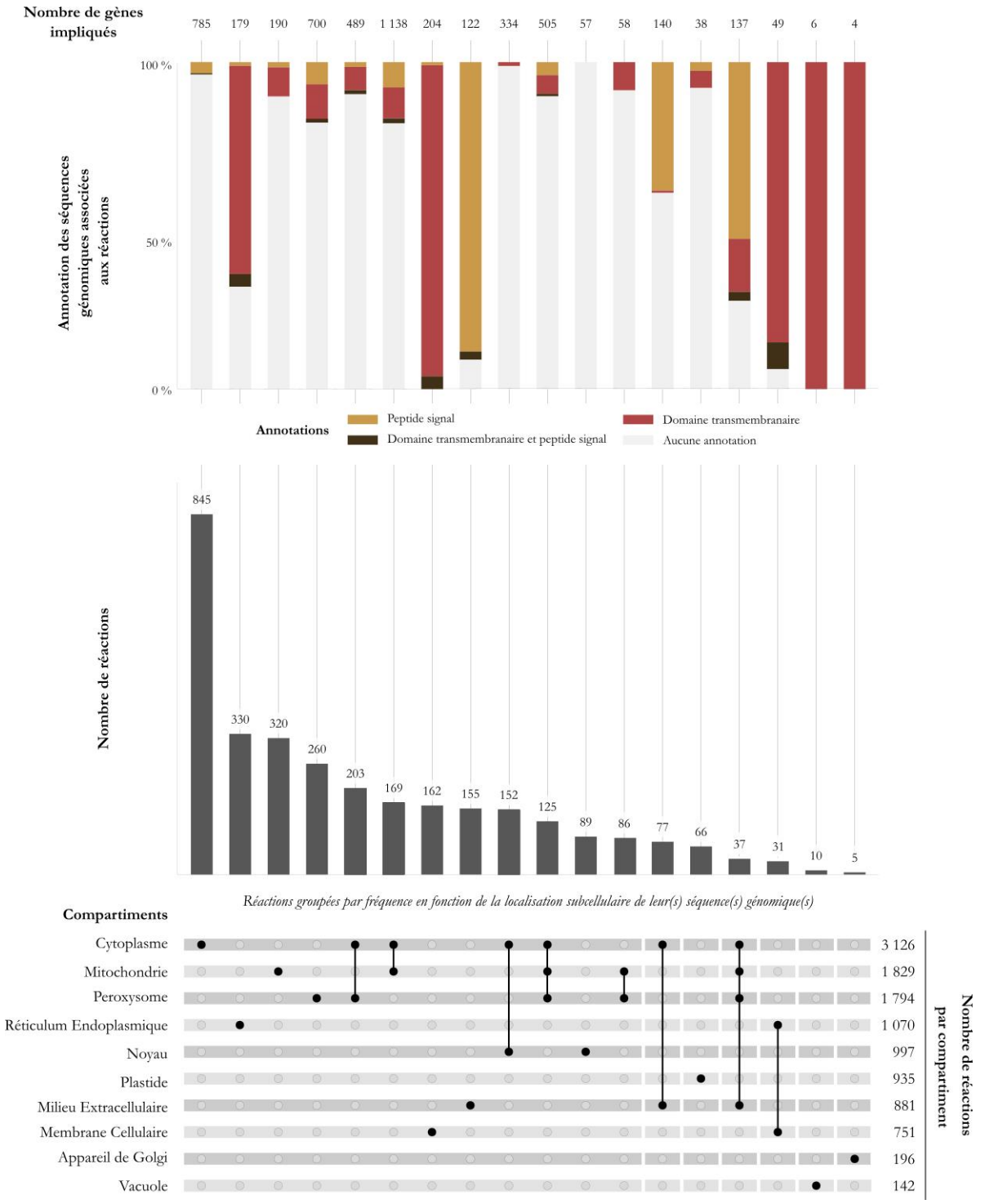


Figure 4-6 : Visualisation du nombre de réactions par compartiment et des annotations fonctionnelles des gènes qui leur sont associés. *Sur les 5 919 réactions d'**iPrub22**, 5 024 d'entre elles peuvent êtres assignées à un compartiment en fonction de l'annotation de leur(s) gène(s). En raison du fort nombre d'intersections possibles avec 10 variables, cette visualisation représente un échantillon de 3 122 réactions (53 % des réactions) et 5 135 gènes (90 % des gènes). Les intersections sont présentées selon le nombre décroissant de réactions. Nous avons choisi de représenter les intersections où les réactions ne s'effectueraient que dans un seul compartiment (i.e. tous les produits géniques des gènes des associations GPR sont localisés dans le même compartiment) ou des réactions dupliquées, tripliquées ou multiples dans des compartiments déjà modélisés dans **iPrub22** ou d'intérêt biologique.*

1.2.3. Restriction à l'étude de la modélisation des mitochondries et des peroxysomes

L'annotation « compartiment » étant portée par le gène, comment la transposer aux métabolites et gérer les réactions ? Nous rappelons que nous disposons de 5 024 réactions qui sont associées à au moins une séquence génomique, et que la modélisation des organites entraînera de ce fait une multiplication du nombre de métabolites et de réactions. À titre d'exemple, nous avons estimé que la représentation de l'ensemble des compartiments étudiés jusqu'alors augmenterait le nombre de 5 024 réactions à 10 927, et ce, sans comptabiliser l'intégralité des connexions qui devront être effectuées pour la fonctionnalité du réseau métabolique. Ainsi, sachant que la gestion de la compartimentation et l'intégration de ces informations est fastidieuse et qu'elle va nécessiter une modification profonde de la structure du réseau, le parti pris est de se focaliser dans un premier temps exclusivement sur l'assignation des informations spécifiques à la mitochondrie et au peroxysome. Cette approche nous permettra ainsi d'évaluer la faisabilité de la compartimentation intracellulaire pour notre réseau.

Notre choix s'est porté sur ces deux organites en raison de leur classicisme et de leur rôle fondamental au sein de l'organisme. Chez les champignons filamenteux, la machinerie de biosynthèse est essentiellement compartimentée entre le cytosol et les peroxysomes *(Meijer et al. 2010)*. Ces deux espaces sont à la fois les lieux de biosynthèse et de dégradation de métabolites tels que l'oxydation des acides gras, le cycle du glyoxylate ou la formation des polycétides et des terpènes *(Bartoszewska et al. 2011; Martín et al. 2012)*. Ils interviennent également dans le métabolisme primaire de plusieurs sources de carbone inhabituelles utilisées pour la croissance de l'organisme. Dans le cadre de la production de pénicilline, les souches hautement productrices de cet antibiotique présentent un nombre plus élevé de peroxysomes que les souches sauvages *(van den Berg et al. 2007)*. Enfin, les peroxysomes tiennent leur nom du peroxyde d'hydrogène qui est le sous-produit de leur enzyme oxydante. La séparation physique de ces réactions, correspondant à l'addition d'hydrogène à de l'oxygène, est donc essentielle pour le maintien du métabolisme pour ne pas interférer avec le grand nombre de réactions ayant lieu au sein du cytoplasme *(Raven et al. 2011)*. Les mitochondries quant à elles sont les usines énergétiques, centre de la respiration cellulaire et de la synthèse d'ATP en métabolisant les sucres *(Raven et al. 2011)*. Possédant leur propre matériel génétique, ces organites ont fait l'objet de très nombreuses études depuis leur caractérisation *(Ernster et Schatz 1981)*.

Désormais, si nous considérons la modélisation de quatre « compartiments » (*i.e.* milieu extracellulaire, cytosol, mitochondrie et peroxysome), il existe théoriquement 15 configurations possibles à envisager en fonction de la répartition des métabolites. La **Figure 4-7** illustre ces possibilités et met en lumière les réactions de transport qui sous-tendent chacun de ces éléments. Nous noterons toutefois que certains scénarios sont peu, voire improbables, d'un point de vue biologique (*e.g.* la présence d'un métabolite dans le milieu extracellulaire et dans l'un ou l'autre des organites sans qu'il soit présent dans le cytosol). Lors de l'assignation automatique de la compartimentation intracellulaire, ces catégories devraient ainsi correspondre à des ensembles vides et pourront servir alors d'indices sur la cohérence des informations modifiées.

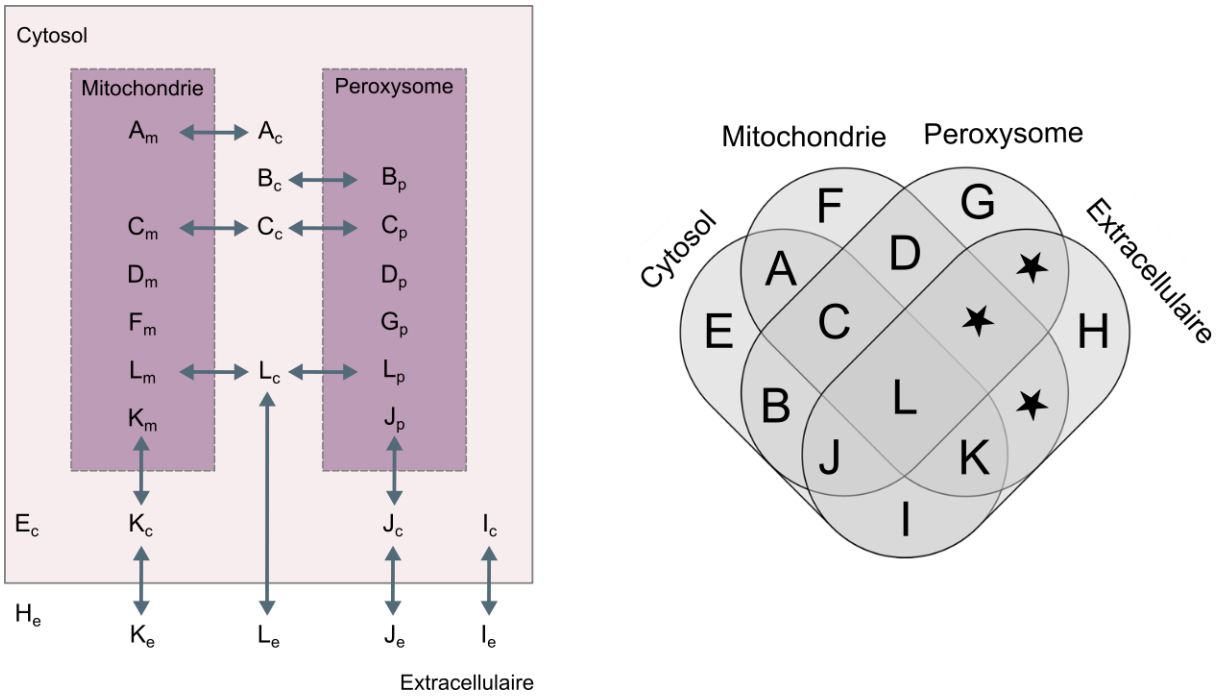


Figure 4-7 : Ensembles et cas de figure théorique de la gestion des métabolites lors de la modélisation de deux organites supplémentaires. *Chaque lettre capitale dans le schéma de gauche représente un métabolite, son compartiment est représenté par sa lettre correspondante en indice (c : cytosol, m : mitochondrie, p : peroxysome, e : extracellulaire). A_m et A_c représente la même entité biologique A. En revanche pour modéliser son action dans deux compartiments, il est nécessaire d'avoir recours à deux entités distinctes pour son encodage dans le* **SBML**. *Les flèches symbolisent les réactions de transport, réversible ou non. Les étoiles (★) dans le diagramme de Veen symbolisent des catégories qui, en raison d'une absence de connexion physique entre eux (i.e. un métabolite provenant du milieu extracellulaire devra traverser le cytosol pour atteindre l'un des deux organites), devraient être vides.*

Pour connaître le volume de données à modifier, intéressons-nous à présent au nombre de réactions associées à chacun des gènes liés aux mitochondries ou aux peroxysomes. Ces informations devraient alors nous renseigner sur le degré de spécificité des voies métaboliques dans chacun de ces organites. La **Figure 4-8** présente le nombre de réactions associées aux gènes dont le produit est localisé dans la mitochondrie ou le peroxysome. Parmi les 626 gènes liés aux mitochondries, 24 % peuvent être considérés comme spécifiques à cet organite, étant associés à une seule réaction. Cependant, 17 % de ces gènes sont impliqués dans plus de 9 réactions. Pour les 440 gènes associés aux peroxysomes, 9 % ne sont liés qu'à une seule réaction, tandis que 32 % participent à plus de 9 réactions.

Les réactions que nous désignons sous le terme « spécifique » sont les plus faciles à traiter puisque la modification à opérer se résume sommairement à une substitution de compartiment. En revanche, à partir du moment où une réaction est associée à au moins deux gènes dont les produits géniques ne possèdent pas la même localisation subcellulaire de nouvelles interrogations, liées à terme à l'exactitude des associations GPR, peuvent émerger. Par exemple, ce phénomène est-il une représentation de robustesse pour ce type de réaction ou résulte-t-il d'incohérences au niveau de l'annotation entraînant de ce fait l'ajout de gènes faux positifs ? L'intégration automatisée de l'information relative à la compartimentation intracellulaire, telle que nous la concevons, repose entièrement et intrinsèquement sur la qualité de ces associations, mais au vu de la somme de données concernées, il n'est pas envisageable de toutes les vérifier manuellement.

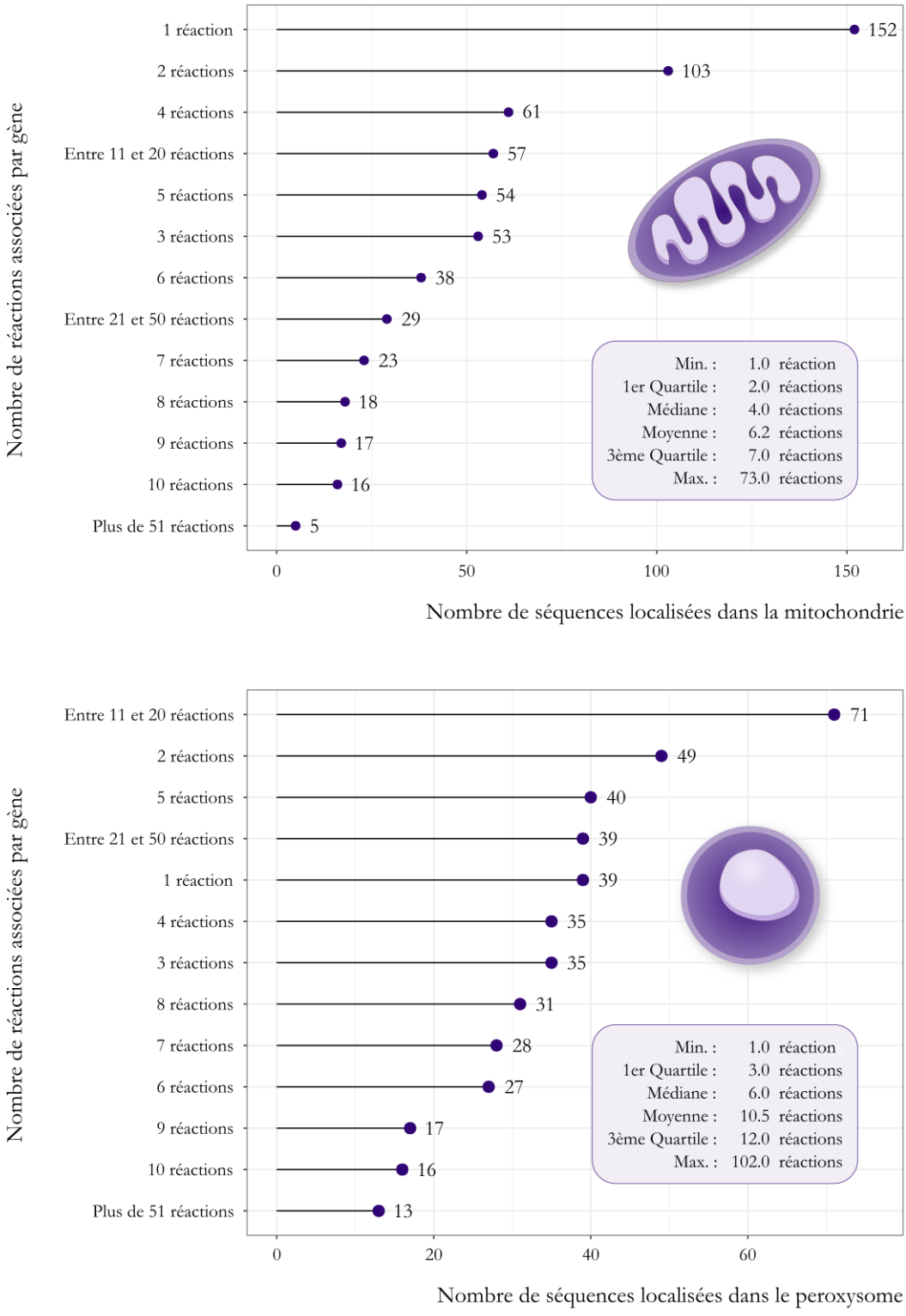


Figure 4-8 : Diagramme en bâtons illustrant le niveau de spécificité des gènes au sein des organites mitochondries et peroxysomes. *Les résultats issus de DeepLoc ont permis de recenser 626 et 440 séquences appartenant respectivement aux mitochondries et aux peroxysomes et impliquées au total dans 3 883 et 4 613 réactions différentes.*

1.2.4. Procédure envisagée pour la modélisation de la compartimentation intracellulaire et causes de son abandon

Nous venons de voir que dans un **GSMN** les informations relatives à la compartimentation sont portées par les métabolites, que les annotations fonctionnelles de localisation sont ajoutées aux gènes et qu'une multiplication des réactions va devoir être envisagée. Quelles sont donc les étapes à suivre pour relier toutes ces informations et, *in fine*, fournir un modèle compartimenté au sens biologique du terme ? La première étape consiste à discrétiser les réactions en réactions T et en réactions S puis à les traiter selon le contenu de leur association GPR.

Dans le cadre des réactions de transport, si au moins un gène est associé à la réaction étudiée (*i.e.* modélisation d'une protéine de transport) l'assignation DeepLoc permettra d'adapter la compartimentation des métabolites. À l'instar des actions effectuées pour déterminer les réactions de transport nécessaires entre les compartiments intracellulaire et extracellulaire (cf. Chapitre 3, section *2.2.2 Modélisation de différentes conditions de culture*, à partir de la page 265) coupler ces informations aux résultats de TMHMM et SignalP devrait éclairer les besoins du réseau en réactions de transport, une étape pouvant être complémentée, au besoin, par l'appui de données issues de la base de données *The Transporter Classification Database (Saier et al. 2021)* ou de la page dédiée aux transporteurs de *P. rubens* sur KEGG BRITE (https://www.kegg.jp/brite/pcs02000.keg).

Pour les réactions dont les substrats et produits sont localisés dans le même compartiment, les actions à entreprendre sont guidées par les résultats présentés en **Figure 4-6**. Pour les réactions ayant lieu dans un seul compartiment autre que le cytosol, l'action principale correspond à substituer la précédente localisation (*i.e.* cytosol) à la nouvelle. Cependant, près de la moitié des réactions aurait lieu dans au moins deux espaces différents induisant alors, dans le plus simple des cas, une duplication des informations. Nous avions envisagé pour chaque réaction à multiplier d'ajouter à la fin de son identifiant une étiquette de repérage comme réalisé dans Prubens. Enfin, pour nous assurer du maintien de la cohérence topologique de notre réseau, nous comptions sur l'appui des résultats de Mene check pour vérifier la productibilité des composés modifiés.

Techniquement, il n'y a aucun frein pour l'intégration de ces données. Le mode de fonctionnement d'AuReMe nous permet d'effectuer les changements nécessaires en *parsant* aisément le fichier `*.padmet` qui sert d'intermédiaire à la génération du fichier `*.sbml`. Le fichier `*.padmet` est une structure modulaire qui représente le réseau métabolique à travers des nœuds, des relations et des attributs, offrant ainsi une grande souplesse pour l'intégration de nouvelles données. Fondamentalement, les scripts permettant de modifier les informations nécessaires à la compartimentation intracellulaire ont été écrits, testés et semblent fonctionner correctement. Nous tenons pour exemple purement illustratif la **Figure 4-9**. Cependant, le temps indispensable pour sélectionner les données d'entrée, l'expertise requise sous-jacente (*i.e.* le transport des molécules, actif ou passif, doit être évalué composé par composé en fonction de ses propriétés physico-chimiques) et l'envergure des curations nécessaires engendrées par la compartimentation nous ont contraints à abandonner cette approche.

Néanmoins, à défaut de pouvoir fournir une reconstruction avec une compartimentation intracellulaire, toutes les informations relatives à la localisation des composés (*i.e.* les résultats de DeepLoc) sont contenues dans les balises `<notes>` des produits géniques au sein du fichier **SBML**. De surcroît, la topologie transmembranaire de chaque produit génique (*i.e.* les résultats de TMHMM) ainsi que la présence potentielle de peptides signaux (*i.e.* les résultats de SignalP lorsque la probabilité dépasse 0,45) sont également renseignées dans ces balises.

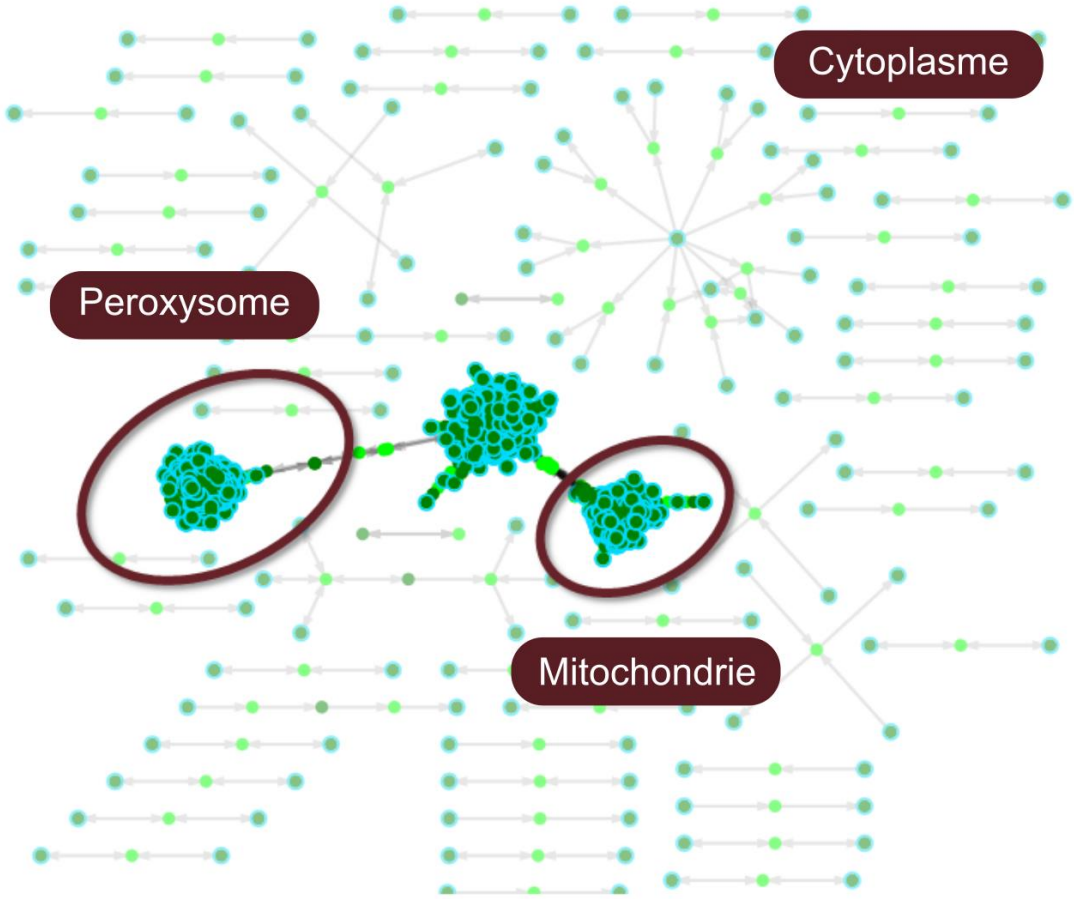


Figure 4-9 : Visualisation de la topologie de l'un de nos drafts après assignation des compartiments mitochondrie et peroxysome. *Nous observons une ségrégation des données en fonction des compartiments, mais également un faible taux de connexions entre chacun des organites et le cytoplasme. Ce constat pourrait indiquer une relative autonomie de la machine cellulaire de ces organites, mais selon toute vraisemblance ces éléments suggèrent surtout un manque de réactions de transport. Réalisé avec ModelExplorer (version 2.1).*

2 - Vers une intégration automatique des voies du métabolisme spécialisé dans les futures reconstructions ?

Comme mentionné dans le chapitre 3, section *3 - Focus sur le métabolisme spécialisé : descriptions de diverses voies de biosynthèse de métabolites spécialisés*, page 308, nous avons été confrontés à des défis significatifs concernant la définition du métabolisme spécialisé et de son (non) intégration dans diverses bases de données. Nous avons constaté, par exemple, une large disparité dans les composés étiquetés « spécialisés » selon les ressources utilisées. De plus, les étapes automatiques mises en œuvre lors du processus de reconstruction se sont révélées peu efficaces en regard de l'intégration du métabolisme spécialisé strict (*i.e.* tel que défini par

exemple par les résultats des outils SMASH). À titre indicatif, les voies de biosynthèse des pénicillines ont nécessité des corrections majeures, tandis que celles des roquefortines et de la méléagrine ont exigé une intégration manuelle intégrale lors de l'étape de *gap-filling*. Ces deux exemples sont particulièrement intrigants puisque ces éléments sont supposés être parmi les mieux documentés et référencés pour notre organisme. Si les voies de biosynthèse des métabolites « standards » ne peuvent être que pauvrement reconstruites automatiquement, il est légitime de s'interroger sur la qualité et la fiabilité des autres éléments détectés.

Ainsi, ce sont les diverses étapes de *gap-filling* subies par notre reconstruction, ainsi que la curation manuelle directe que nous avons réalisée qui garantissent l'intégration et la qualité du métabolisme spécialisé d'***i*Prub22**, du moins tel qu'il est disponible dans la base de données MetaCyc. Toutefois, cette approche ne permet pas d'apporter une réponse généralisable à l'intégration de telles données pour les reconstructions futures. Pourtant, des solutions techniques potentiellement simples à mettre en œuvre existent. Dépassant le cadre de cette thèse et faute de temps, cette question n'a pas été explorée en détail. Néanmoins, nous souhaitons aborder une piste de résolution basée sur le constat suivant : à notre connaissance, il n'existe pas encore de connexion directe entre les bases de données de reconstruction et celles dédiées à la caractérisation du métabolisme spécialisé.

Pour ce faire, nous pourrions envisager, en nous inspirant notamment des principes liés à la génération de la reconstruction du sous-réseau d'orthologie (cf. Chapitre 2, section *3.3 Orthologie - Reconstruction du draft, page 104 et Annexe 2 : Reconstruction d'iPrub22 – sous-réseau issu des recherches d'orthologie*, page 433) de restreindre ce principe à la seule propagation de données liées au métabolisme spécialisé afin d'automatiser leur intégration. Au lieu d'effectuer une reconstruction globale, nous proposons ainsi de revenir aux fondamentaux et d'effectuer une reconstruction voie par voie pour ce type de métabolisme. Dans ce cadre, deux fichiers seulement seraient initialement nécessaires, l'un contenant les voies d'intérêt et ses métadonnées associées (*i.e.* fichier **SBML** restreint), l'autre recensant les séquences génomiques associées aux réactions précédemment définies (*i.e.* un protéome ou un génome artificiel – fichier fasta). La base de données *Minimum Information about a Biosynthetic Gene cluster* (MIBiG) *(Terlouw et al. 2023)* offre un cadre standardisé pour fournir des métadonnées complètes et uniformes afin de décrire les BGCs, et 2 502 entrées y sont recensées en octobre 2024. L'extraction et la compilation des séquences génomiques, notamment celles des gènes biosynthétiques, associées à chaque cluster, pourrait permettre de générer rapidement les fichiers de séquences à interroger. Le fichier **SBML** qui servirait alors de *template* de voies du métabolisme spécialisé, devrait, quant à lui, répondre à l'ensemble des critères essentiels établis précédemment (*e.g.* définition des réactions et des métabolites contenus dans chaque voie, référencement vers diverses voies de données, *etc.*) et représenterait un cadre uniforme de recherche. Les recherches d'homologie à effectuer ensuite entre le protéome de l'organisme étudié et le protéome artificiel pourraient bénéficier de critères de filtration supplémentaires en fonction, par exemple, du nombre de gènes de BGCs identifiés, avant d'inférer l'intégralité de la voie à sa reconstruction.

Ainsi, envisager une passerelle (*i.e.* une interface de propagation des données) entre les données de MIBiG, ou issues plus spécifiquement des résultats des outils SMASH, et celles des bases de données de reconstruction permettrait de généraliser à grande échelle l'intégration du métabolisme spécialisé au sein des reconstructions, qu'elles soient anciennes ou nouvelles. La conception et la diffusion de tels fichiers permettraient à l'avenir de rechercher et d'intégrer automatiquement des données de qualité relatives aux voies de biosynthèse du métabolisme spécialisé, et ce, pour tout nouvel organisme.

3 - Le pouvoir prédictif de la comparaison de GSMNs

Nos travaux se sont concentrés sur l'étude d'un seul organisme. Cependant, rappelons qu'un organisme évolue dans un environnement complexe et interconnecté, influencé par des interactions multiples avec d'autres espèces. Pour mieux comprendre ces interactions et leur impact, il est nécessaire d'adopter une perspective plus large, en s'éloignant de l'analyse centrée sur un organisme unique. La comparaison des **GSMNs** entre espèces offre l'opportunité de révéler ces relations complexes et d'éclairer l'histoire évolutive des organismes. Nous présentons brièvement ci-après quelques exemples d'études ayant adopté une échelle d'analyse plus large, afin d'identifier des motifs évolutifs, écologiques et fonctionnels dans le métabolisme. Ces approches offrent un fort potentiel prédictif pour mieux appréhender les dynamiques métaboliques dans des contextes plus complexes et interspécifiques.

La comparaison des **GSMNs**, bien que complexe en raison des choix de reconstruction et de la nature des données, permet de mettre en évidence les similitudes et différences évolutives entre les organismes. Cependant, un défi majeur réside dans l'accessibilité des modèles, qui sont souvent disponibles uniquement par le biais des articles de recherche associés, leur conférant ainsi une certaine confidentialité et limitant leur réutilisation. Un aspect qui tend néanmoins à s'atténuer avec l'adoption de la mise à disposition systématique des nouveaux **GSMNs** sur des plateformes spécifiques telles que BioModels *(Malik-Sheriff et al. 2020)*, des systèmes de contrôle de version comme GitHub et GitLab *(Choudhury et al. 2020)*, ou des systèmes d'archivage tels que Zenodo *(Peters et al. 2017)*.

Ces approches de comparaison de réseaux ont permis, par exemple, de prédire la valeur adaptative des modes de vie pathogènes et non pathogènes chez *Pseudomonas (Mithani et al. 2011)* ou d'identifier les gènes essentiels du métabolisme central des actinomycètes *(Alam et al. 2011)*. En étudiant la capacité de production hétérologue théorique de 15 métabolites secondaires à partir de données issues de 38 **GSMNs** d'actinobactéries, les travaux de Zakrzewski *et al. (2012)* confirment que la production de métabolites spécialisés au sein d'une espèce n'est pas corrélée avec le nombre et la diversité des clusters de gènes contenus dans son génome. Dans le domaine de la santé humaine, Magnúsdóttir *et al. (2017)* ont montré que les interactions entre les espèces du microbiote intestinal (*i.e.* 773 **GSMNs** générés) dépendent à la fois de leur potentiel métabolique et des nutriments présents chez l'hôte humain.

L'élargissement des analyses des **GSMNs** aux communautés microbiennes, comme l'illustre la revue de Biggs *et al.* *(2015)*, ainsi que l'émergence de nouveaux outils tels que Metage2Metabo *(Belcour et al. 2020)*, permettent d'étudier la complémentarité métabolique et la coopération entre espèces. À une échelle plus large, la caractérisation des composés exogènes extraits de l'environnement par un organisme pour son développement constitue un indicateur de l'étendue de sa niche écologique *(Borenstein et al. 2008)*. De plus, le couplage de l'analyse topologique inter-espèces des **GSMNs**, afin de définir l'ensemble des *seeds* nécessaires à chaque **GSMN** avec des analyses phylogénétiques fournit des informations sur les interactions évolutives entre organismes et habitats (*i.e. « reverse ecology »*) *(Borenstein et al. 2008; Borenstein et Feldman 2009)*. Il convient toutefois de noter que la sensibilité de ce type de reconstruction phylogénétique, et plus généralement de la comparaison des **GSMNs**, dépend essentiellement de la taille et de la qualité intrinsèque des modèles utilisés *(Schulz et Almaas 2020)*. L'extension et la combinaison de ces approches ouvrent ainsi de nouvelles perspectives pour l'optimisation des étapes de curation supplémentaires des **GSMNs**. Le raffinement des modèles qui en résulte permettra alors une amélioration, par exemple, de l'orientation des projets d'ingénierie métabolique et cellulaire.

Enfin, la comparaison des **GSMNs** ouvre également des perspectives pour la compréhension de l'histoire évolutive des espèces. La reconstruction d'arbres phylogénétiques à partir de voies métaboliques *(Hong et al. 2004)* ou de réseaux *(Gamermann et al. 2014)* met en lumière la distance évolutive et les liens entre les profils métaboliques. En effet, le regroupement de réactions en fonction de leur présence ou absence à travers diverses espèces constitue un indicateur de leur degré de spécificité (*i.e.* métabolisme central ubiquitaire versus métabolisme hautement spécialisé) *(Schulz et Almaas 2020)*. Sur cette base, la reconstruction d'un réseau ancestral complet pour un groupe d'espèces peut également être envisagée par une estimation parcimonieuse du nombre de mutations survenues au cours de l'évolution, fournissant des informations sur les gains ou pertes de voies métaboliques au fil du temps *(Pitkänen et al. 2013)*. Cette approche, qui repose sur la phylogénie *via* des modèles probabilistes, est développée et généralisée au sein du pipeline CoReCo (*Comparative Reconstruction*). La preuve de concept a été réalisée initialement sur la reconstruction et la comparaison de 49 espèces de champignons filamenteux *(Pitkänen et al. 2014)*, puis sur la génération d'un **GSMN** de haute qualité pour *Trichoderma reesei* *(Castillo et al. 2016)*. En 2020, Schultz et Almaas *(Schulz et Almaas 2020)* ont présenté un arbre phylogénétique reconstruit à partir de la comparaison de réseaux métaboliques de 975 espèces. Bien qu'il existe quelques inexactitudes dans le placement de certains cas spécifiques, la forte congruence de cet arbre avec celui de « l'Arbre de la Vie » à plusieurs niveaux taxonomiques est prometteuse pour le développement et les résultats des analyses comparatives.

Nous pouvons supposer que, dans le futur, la mise en évidence systématique des spécificités entre les modèles aura des impacts bénéfiques sur la recherche et la caractérisation des produits naturels.

4 - Vers une nouvelle reconstruction adaptée d'*i*Prub22 pour une souche marine de *Penicillium* ?

Nous avons souligné que chaque **GSMN** représente les potentialités exprimées par son génome. Pour cibler les voies réellement conservées chez notre souche, des couches de données transcriptomiques et de régulation sont nécessaires. Nous avons annoncé avoir reconstruit un **GSMN** pour la souche Wisconsin 54-1255, mais est-il réellement spécifique à cette souche, à l'espèce *rubens* ou au genre *Penicillium* ? L'objectif des éléments présentés dans cette section est ainsi double, tout en cherchant à répondre à cette question (*i.e.* la profondeur, la finesse et la spécificité des **GSMNs**), nous mettons également en perspective notre protocole de reconstruction afin d'acter ou non sa répétabilité.

4.1. Les *Penicillium* des champignons terrestres mais aussi marins, un nouveau champ de prospection ?

Les champignons filamenteux sont reconnus pour produire une vaste diversité et une grande quantité de produits naturels *(Wiemann et Keller 2014)*. L'étude de leur chimiodiversité s'est longtemps limitée aux souches terrestres et de nouveaux champs de prospection se dessinent avec la mise en évidence de nombreuses lignées marines *(Richards et al. 2012)*. Selon toute vraisemblance, les champignons marins sont susceptibles d'être un réservoir pour la découverte de nouveaux produits naturels. En effet, en raison de la nature de l'environnement dans lequel ils évoluent, ils sont soumis à une gamme plus large de stress et de pression de sélection – température, pression osmotique, lumière, *etc.* – que les souches terrestres. Depuis le milieu des années 80, la découverte et la caractérisation des produits naturels marins sont suivies régulièrement par la publication de revues de la littérature. Ces dernières, qui se sont longtemps concentrées sur les métabolites produits par des algues, des éponges ou des mollusques *(Bhadury et al., 2006; Blunt et al. 2011; Imhoff, 2016)*, accordent aujourd'hui une attention croissante aux métabolites fongiques *(Blunt et al., 2017; Overy et al., 2019; Carroll et al., 2023)*.

Le laboratoire ISOMer dispose d'une souche marine de *Penicillium chrysogenum*, isolée à partir d'une coque collectée au Croisic (France) en Novembre 1994. Suite au séquençage de son génome, réalisé au moyen de la technologie PacBio2 (*i.e. long-read sequencing*), deux assemblages ont été mis à notre disposition : l'un réalisé avec Canu *(Koren et al. 2017)* et l'autre avec Flye (v2.6) *(Kolmogorov et al. 2019)*, accompagnés des prédictions de gènes associées.

4.2. Caractéristiques des assemblages, vers une annotation structurale et fonctionnelle de *Penicillium chrysogenum* MMS5 ?

Un assemblage est un ensemble de scaffolds, c'est-à-dire des séquences ordonnées et orientées, obtenues par la concaténation de contigs et pouvant contenir des gaps, déduits à partir des *reads* de séquençage. Dans l'optique de générer un premier draft de reconstruction pour la souche marine MMS5, nous avons mené des analyses sommaires et préliminaires de comparaison des assemblages afin de déterminer celui qui serait de meilleure qualité pour la poursuite des analyses. Tout d'abord, nous avons vérifié la présence de contaminants procaryotes à l'aide de Taxoblast (v1.21) *(Dittami et Corre 2017)*. Sur les 143 contigs de l'assemblage produit avec Canu, 116 ont été assignés à des séquences eucaryotes tandis que 27 n'ont pas pu être identifiés. En revanche, les 82 contigs de l'assemblage réalisé avec Flye ont tous été identifiés comme d'origine eucaryote. Des analyses complémentaires ont ensuite été effectuées à l'aide des outils BUSCO (*Benchmarking Universal Single-Copy Orthologs* – v4.0.5) *(Manni et al. 2021)* et QUAST (*QUality ASsessment Tool* – v5.0.2) *(Gurevich et al. 2013)*. Nous avons ensuite effectué un alignement des assemblages de MMS5 avec l'assemblage de Wisconsin 54-1255 à l'aide de MUMmer (~ *Maximal Unique Matches* – v3.23) *(Kurtz et al. 2004)*. Ces analyses révèlent que le génome de MMS5 s'aligne à 98 % sur celui de la souche Wisconsin 54-1255, confirmant ainsi une forte similarité. Enfin, une première recherche d'ARN a été menée avec Barrnap (v0.9).

Une synthèse des diverses métriques issues des outils mentionnés ci-dessus est présentée dans le **Tableau 4-3**, tandis que la **Figure 4-10** récapitule les résultats des analyses BUSCO, réalisées avec des assemblages plusieurs souches de *P. rubens*, qu'elles soient sauvages *(Peng et al. 2014; Gujar et al. 2018)* ou industrielles *(Van den Berg et al. 2008; Specht et al. 2014; Wang et al. 2014)* afin de mettre en perspective les résultats obtenus pour les assemblages de MMS5. Le **Tableau 4-4** se concentre, quant à lui, sur les informations issues de MUMmer, dont les résultats d'alignements sont présentés en **Figure 4-11**.

En l'absence de données RNAseq et face à des résultats contrastés (*i.e.* l'assemblage obtenu avec Canu est plus complet, tandis que celui de Flye est plus compact), il ne nous est pas possible de justifier l'utilisation préférentielle d'un assemblage par rapport à l'autre. Nous avons alors décidé de réaliser le draft de la reconstruction sur les deux jeux de données en parallèle, afin d'évaluer si des différences significatives pourraient émerger lors du processus de reconstruction.

Compte tenu de la grande similarité entre les génomes de *P. rubens* Wisconsin 54-1255 et *P. chrysogenum* MMS5, il est légitime de se demander s'il est nécessaire de réaliser une nouvelle reconstruction pour cet organisme. La détection des différences suppose un niveau de précision élevé dans l'annotation pour saisir les spécificités propres à la souche marine. De plus, la comparaison des reconstructions n'aura de valeur que si le protocole de reconstruction pour MMS5 est rigoureusement identique à celui utilisé pour Wisconsin 54-1255, notamment en ce qui concerne les versions des outils et bases de données. Ces dernières évoluent rapidement, souvent avec des mises à jour bi-annuelles, ce qui peut introduire des divergences significatives. Enfin, il convient de souligner le caractère exploratoire et préliminaire des éléments discutés dans cette section.

Tableau 4-3 : Synthèse des résultats et métriques d'intérêt concernant les deux assemblages mis à disposition

Caractéristiques générales				
	Données issues de l'assemblage Canu		***Données issues de l'assemblage Flye***	
	Contigs	*CDS*	*Contigs*	*CDS*
Nombre de séquences	143	*9 431*	82	*9 546*
Nombre de résidus	*30 053* 642	*13 614* 771	*31 790* 321	*13 494* 402
Taille minimale des séquences	*4 850*	201	*8 816*	201
Taille maximale des séquences	*133 0050*	*21 465*	*1 627* 753	*21 483*
Taille moyenne des séquences	*210 165*.33	*1 443*.62	*387 686*.84	*1 413*.62
N50	*394 927*	*1 704*	*683 356*	*1 677*

Taxoblast v1.21β	
Assemblage Canu	116 contigs sont définis comme appartenant au règne eucaryote 27 contigs n'ont pas reçu d'assignation
Assemblage Flye	Les 82 contigs sont définis comme appartenant au règne eucaryote

QUAST (v5.0.2)		
	Contigs issus de l'assemblage Canu	***Contigs issus de l'assemblage Flye***
Statistiques du génome		
Fraction du génome	89,774 %	93,745 %
Taux de duplication	1,011	1,017
Plus grand alignement (pb)	*733 888*	*1 204* 622
Longueur totale alignée (pb)	*29 202* 834	*30 659* 441
NG50	*325 967*	*683 356*
NA50	*273 043*	*367 004*
NGA50	*240 532*	*367 004*
LG50	26	16
LA50	36	30
LGA50	41	30

***NG50** : Longueur du contig pour lequel la somme des contigs égaux ou plus grands atteint 50 % de la longueur totale du génome de référence estimé. Mesure de la contiguïté prenant en compte la taille attendue du génome.*

***NA50** : Longueur du plus petit alignement pour lequel la somme des alignements plus grands ou égaux couvre 50 % de la longueur totale des alignements par rapport au génome de référence. Mesure de la contiguïté en tenant compte des alignements avec le génome de référence.*

***NGA50** : Longueur du plus petit contig aligné correctement pour lequel la somme des contigs alignés couvre 50 % de la longueur du génome de référence.*

***LG50** : Nombre minimum de contigs requis pour atteindre 50 % de la longueur totale du génome de référence*

***LA50** : Nombre minimum d'alignements nécessaires pour atteindre 50 % de la longueur totale des alignements par rapport au génome de référence.*

***LGA50** : Nombre minimum de contigs correctement alignés nécessaires pour atteindre 50 % de la longueur totale du génome de référence.*

Non alignés		
Nombre de contigs partiellement non alignés	53	41
Longueur partiellement non alignée	*813 439*	*1 099* 285

Tableau 4-3 (suite)

Assemblages erronés		
Nombre d'assemblages erronés	173	244
Nombre de relocalisations	67	84
Nombre de translocations	101	154
Nombre d'inversions	5	6
Nombre de contigs erronés	67	43
Longueur des contigs erronés	18 035 501	22 629 409
Mésappariements		
Nombre de mésappariements	35 426	44 796
Nombre d'indels	19 920	48 038
Longueur des indels	42 385	85 042
Statistiques sans référence		
Nombre de contigs	143	82
Nombre de contigs (≥ 10000 pb)	142	81
Nombre de contigs (≥ 25000 pb)	136	80
Nombre de contigs (≥ 50000 pb)	109	75
Plus grand contig	13 30 050	1 627 753
Longueur totale	30 053 642	31 790 321
N50	394 927	683 356
N75	206 568	341 017
L50	23	16
L75	50	33
Gènes prédits		
Nombre de gènes prédits (uniques)	17 083	18 920
Nombre de gènes prédits ≥ 0 pb	17 196 + 14 partiels	18 960 + 4 partiels
Nombre de gènes prédits ≥ 300 pb	14 699 + 10 partiels	15 988 + 4 partiels
Nombre de gènes prédits ≥ 1500 pb	3 666 + 0 partiels	3 604 + 1 partiel
Nombre de gènes prédits ≥ 3000 pb	754 + 0 partiels	641 + 0 partiels
Nombre de gènes rRNA prédits	55 + 2 partiels	35 + 0 partiels

BUSCO v4.0.5 (prédiction de gènes avec Augustus – intrinsèque à Busco)

	Assemblage Canu	***Assemblage Flye***	***Assemblage Wisconsin***	
			Prédiction	*En réalité*
Nombre de gènes prédits	4 219	4 223	4 220	12 556
Nombre de séquences protéiques prédites	8 542	8 446	8 550	-

Tableau 4-3 (suite et fin)

Barnnap (v0.9)		
	Assemblage Canu	***Assemblage Flye***
Nombre de séquences	72	37
5S	30	33
28S	13	1
5.8S	14	1
18S	15	2
Nombre de résidus	*75 659*	*10 073*
Longueur séquences (Min)	115	115
Longueur séquences (Max)	*3 675*	*3 675*
Longueur séquences (Moyenne)	*1 050.82*	272.24

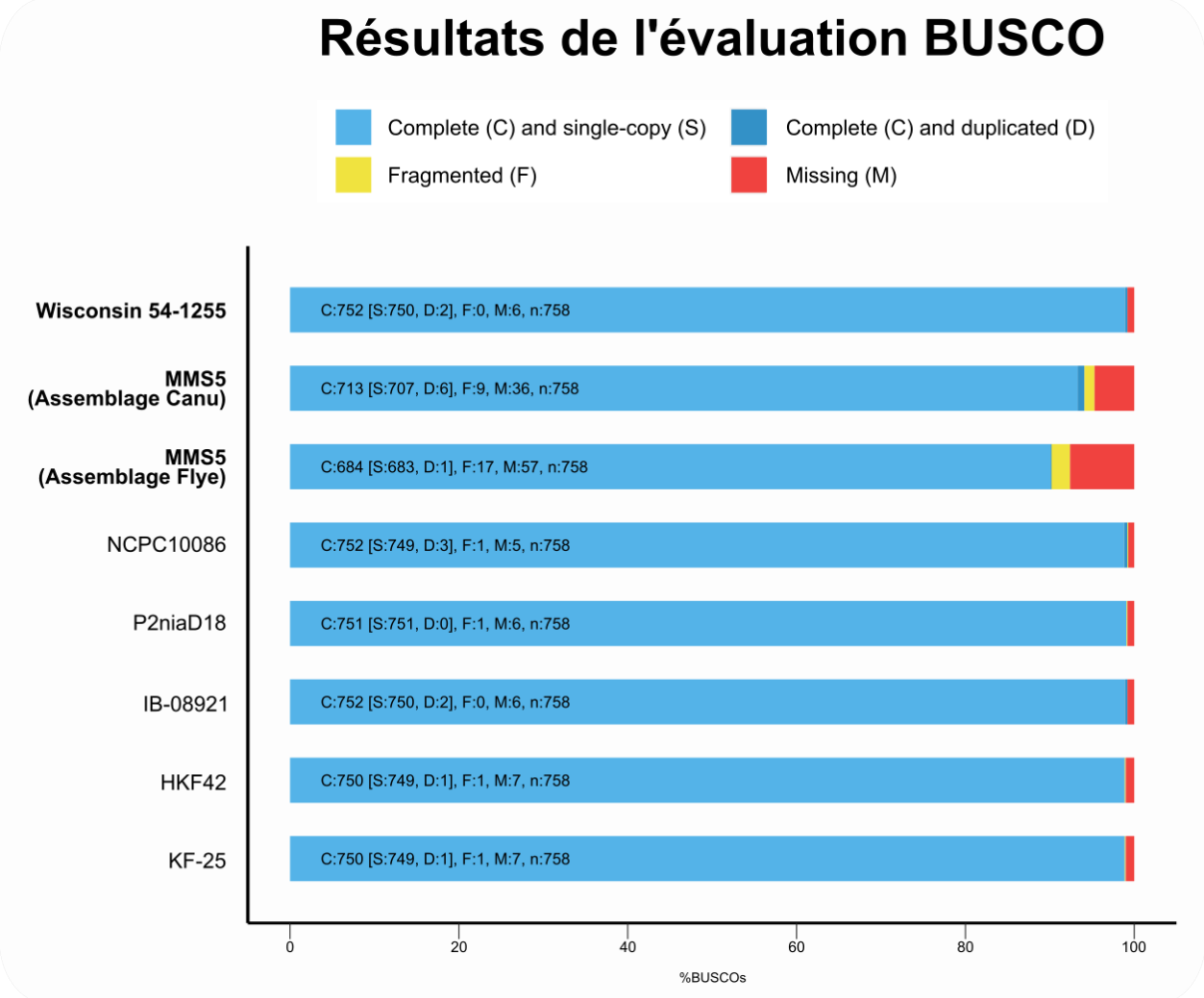


Figure 4-10 : Résultats Busco pour les deux assemblages de MMS5 réalisés avec Canu et Flye. *À titre de comparaison, les résultats issus des assemblages de souches sauvages de Penicillium chrysogenum/rubens HKF42 (Gujar et al. 2018) et KF-25 (Peng et al. 2014), et de souches industrielles NCPC10086 (Wang et al. 2014), P2niaD18 (Specht et al. 2014) et Wisconsin 54-1255 (Van den Berg et al. 2008) sont également présentés.*

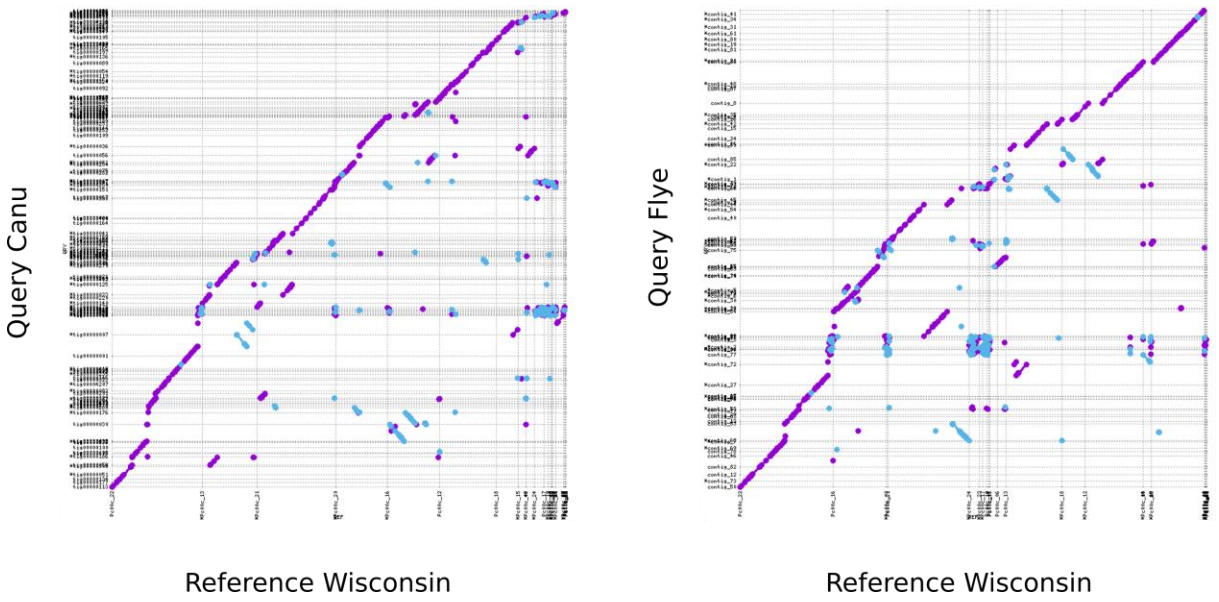


Figure 4-11 : Alignements des nucléotides des assemblages de MMS5 contre ceux de la référence terrestre Wisconsin 54-1255. *Ces alignements confirment une forte similarité, et donc une conservation des séquences, entre la souche marine MMS et la souche terrestre Wisconsin 54-1255. Alignements générés avec MUMmer (v3.23).*

Tableau 4-4 : Synthèse des caractéristiques des résultats des alignements réalisée avec MUMmer (v3.23)

	Référence Wisconsin 54-1255	Query MMS5 Assemblage Canu	Référence Wisconsin 54-1255	Query MMS5 Assemblage Flye
Séquences				
Nombre Total	49	143	49	82
Alignées	48 (97.96 %)	143 (100.00%)	48 (97.96%)	82 (100.00%)
Non Alignées	1 (2.04 %)	0 (0.00%)	1 (2.04%)	0 (0.00%)
Bases				
Nombre Total	*32 223* 735	*30 053* 642	*32 223* 735	*31 790* 321
Alignées	*30 538* 824 (94.77 %)	*29 454* 422 (98.01 %)	*31 457* 519 (97.62 %)	*31 099* 032 (97.83 %)
Non Alignées	*1 684* 911 (5.23 %)	599 220 (1.99 %)	766 216 (2.38 %)	691 289 (2.17 %)
Alignements				
1-to-1	575	575	601	601
Longueur totale	*29 239* 876	*29 224* 458	*30 744* 583	*3 0699 344*
M-to-M	*2 247*	*2 247*	*2 058*	*2 058*
Longueur totale	*31 446* 511	*31 429* 467	*32 934* 431	*32 887* 780

Tableau 4-4 (suite et fin)

Estimations des caractéristiques				
Breakpoints	*4 402*	*4 177*	*4 024*	*3 982*
Relocations	26	32	51	37
Translocations	237	114	187	139
Inversions	16	24	27	32
Insertions	*2 454*	685	*1 749*	935
Tandem	25	22	25	26

4.3. Vers un draft de reconstruction pour la souche MMS5 ?

Les résultats d'alignement présentés en **Tableau 4-4** et **Figure 4-11** témoignent d'une forte similarité entre les génomes des souches marine et terrestre. Il est donc raisonnable de supposer que les divergences entre ces deux souches résultent de l'évolution (*i.e.* acquisition ou perte de fonctions) et de leur adaptation à leur environnement de vie. Cependant, les connaissances actuelles sont-elles suffisantes pour détecter des différences significatives dans des génomes qui présentent moins de 2 % de différences ?

Les analyses exploratoires du génome, réalisées avec FungiSMASH (v6.0.0) *(Blin et al. 2021)* pour la détection des clusters de gènes sur la base de l'existence de domaines protéiques spécifiques, révèlent de légères différences entre ces deux souches. Pour rappel, l'outil détecte 42 clusters de gènes chez la souche Wisconsin 54-1255, dont 12 sont associés à au moins un produit naturel identifié (cf. **Tableau 3-14**, page 314). En ce qui concerne la souche MMS5, 38 clusters de gènes sont détectés et 15 d'entre eux sont associés à au moins un produit naturel identifié. Toujours selon FungiSMASH, il semblerait respectivement que la chrysogine et que l'andrastine A, le trans-resorcylide, le NG-391 et la fusarinine soient spécifiques, ou du moins identifiés, à la souche terrestre et à la souche marine.

Toutefois, tout au long de notre utilisation de cet outil, nous avons observé une variabilité notable, d'une part, entre les résultats générés par AntiSMASH et FungiSMASH, mais également entre les versions de l'outil. Par exemple, sur les données de la souche marine MMS5, un des clusters de gènes qui ressortait avec une version antérieure était celui de l'aspercryptine *(Chiang et al. 2016; Romsdahl et Wang 2019)*. Ce dernier n'est pas caractérisé de prime abord chez la souche Wisconsin 54-1255. Cependant, des recherches d'homologie (*i.e.* la correspondance entre la souche marine et terrestre) permettent d'associer un cluster détecté, mais non caractérisé chez Wisconsin 54-1255 à ce métabolite. Ainsi, soit cet ensemble de gènes a perdu sa fonctionnalité au cours de l'évolution, soit le produit de ce cluster de gènes est similaire à l'aspercryptine (*e.g.* andrastine ?) et a été corrigé depuis. Aujourd'hui, les résultats fournis avec la version 7.0 d'AntiSMASH/FungiSMASH *(Blin et al. 2023)* semblent être en mesure de détecter et d'identifier un plus grand nombre de BGCs qu'au début de nos travaux.

C'est typiquement ce genre d'éléments que nous cherchons à détecter avec la génération d'un draft de reconstruction pour la souche marine MMS5 : cibler les différences potentielles et identifier les similarités du métabolisme de ces deux souches. Pour ce faire, nous réalisons, comme pour ***i*Prub22** deux sous-réseaux : l'un issu de l'annotation fonctionnelle basé sur les prédictions des gènes de MMS5, l'autre issu de recherches d'homologie. La différence réside ici dans le choix des *templates* puisque seule une recherche d'homologie entre les données des souches MMS5 et Wisconsin 54-1255 a été effectuée. Par l'analyse des différences (*i.e.* voies du métabolisme soutenues par des séquences génomiques chez ***i*Prub22**, mais non inférées au draft de la reconstruction de MMS5) nous espérons mettre en exergue les spécificités de la souche terrestre. Sans réelle surprise, la quasi-intégralité des séquences génomiques putatives de MMS5 trouve une séquence homologue dans le protéome de Wisconsin 54-1255 (**Tableau 4-5**). Selon l'assemblage considéré, entre 153 et 178 réactions seraient exclusives à la souche terrestre. L'étape suivante consistera alors à analyser ces réactions : quelles sont leurs fonctions ? Se regroupent-elles au sein de voies métaboliques complètes ou partielles ? Nos analyses n'ont cependant pas été approfondies.

Tableau 4-5 : Résultats de la recherche d'orthologie réalisée avec OrthoFinder et de la génération du sous-réseau qui en résulte.

	MMS5 – Canu	MMS5 – Flye	Wisconsin 54-1255
Nombre de gènes	*9 431*	*9 546*	*12 556*
Nombre de gènes assignés à des orthogroupes	*9 370 (99,4 %)*	*9 437 (98,9 %)*	*10 822 (86,2%)*
Nombre de gènes non assignés à des orthogroupes	*61 (1,1 %)*	*109 (0,6 %)*	*1 734 (13,8 %)*
Nombre total de réactions inférées	*4 846*	*4 871*	-
Nombre de réactions inférées possédant un identifiant **MetaCyc**	*4104*	*4 122*	-

Afin de confirmer et de potentiellement compléter le sous-réseau d'orthologie, nous avons ensuite réalisé, comme pour ***i*Prub22**, l'annotation fonctionnelle des données. En détournant Trinotate de sa fonction première, nous avons profité de ses fonctionnalités internes afin d'extraire automatiquement les identifiants KO (*i.e.* ancrage pour notre recherche de numéro EC) à partir des identifiants swissprot correspondants. Cependant, comme détaillé dans le Chapitre 2 : *Glossaire des outils, ressources et concepts employés*, page 11 lors du déploiement de la base de données SQLITE médiée par Trinotate (*i.e.* étape réalisée en vue de regrouper les données de l'annotation fonctionnelle pour faciliter leur sélection), cette dernière s'initialise en téléchargeant les versions au temps T des bases de données qu'elle va utiliser. Entre le moment où nous avons généré les premiers drafts d'***i*Prub22** et les tests que nous avons effectués pour MMS5, la base de données Swissprot a été mise à jour et cette fonctionnalité est devenue caduque. La perte de cette information essentielle compromet ainsi une partie du protocole suivie pour la génération du sous-réseau issu de l'annotation fonctionnelle. Des alternatives devront donc être trouvées à l'avenir afin de générer le fichier complémentaire d'annotation fonctionnelle à fournir à Pathways Tools.

Nous profitons de ce constat pour faire un bref aparté sur la méthodologie que nous avons suivie pour générer ***i*Prub22**. La critique majeure que nous pouvons faire à notre protocole de reconstruction est de ne pas suivre un « pipeline standardisé », sous-entendu un *workflow* où les différents outils, et *de facto* versions, seraient stables car encapsulés dans un environnement isolé. Cet aspect est notamment marqué pour la génération du sous-réseau issu de l'annotation fonctionnelle et s'atténue par la suite avec l'utilisation d'AuReMe et MATLAB. Le recours nécessaire à une méthodologie FAIR, que ce soit dans l'acquisition, la gestion des données ou les outils employés à cette fin, n'est plus à démontrer. Ce que nous venons d'énoncer n'est qu'un exemple supplémentaire des enjeux liés à la pérennité des outils et à l'application de procédure durable et stable. En l'état, notre protocole de reconstruction dépend intégralement du maintien des ressources telles qu'elles étaient au début de cette thèse. Nous pourrions également ajouter à cet exemple, le changement de politique concernant l'accès aux ressources BioCyc, désormais payant, qui, comparé à KEGG ou BiGG, pourrait à l'avenir influencer les choix méthodologiques.

Cependant, comme nous disposions d'une ancienne base de données SQLITE propre, générée lors de la reconstruction d'***i*Prub22**, nous avons pu poursuivre notre procédure et finaliser un draft de reconstruction pour la souche marine en fusionnant les deux sous-réseaux obtenus. Les caractéristiques sommaires des drafts ainsi obtenus sont présentées dans le **Tableau 4-6**.

La génération des drafts a révélé que 35 % des informations provenaient exclusivement de la recherche d'homologie, tandis que l'annotation fonctionnelle n'a enrichi que légèrement ces résultats préliminaires (**Figure 4-12**). Néanmoins, comme attendu, les différences en termes de réactions et de métabolites entre les reconstructions de la souche marine et de la souche terrestre sont peu nombreuses. Ce constat suggère que les bases de données actuelles ne permettent pas de produire des modèles de réseaux métaboliques suffisamment spécifiques et pertinents à l'échelle d'une souche de *Penicillium*. À terme, nous pouvons supposer que des **GSMNs** suffisamment précis permettront de mieux appréhender les différences exprimées entre souches en fonction de leur environnement. En l'état, la finesse de ces modèles reste limitée par la qualité et la granularité des bases de données disponibles. En conclusion, nous notons à titre anecdotique, la présence de quelques produits naturels parmi les composés exclusifs de la souche marine tels que la demethylsordarine (*i.e.* terpénoïde), le coumaroylquinate, la sordaricine (*i.e.* terpénoïde) et le *compound class* `PENICILLIN` (*i.e.* entité regroupant toutes les instances des pénicillines, mais aussi les carbenicillines, l'ampicilline, *etc.*).

Tableau 4-6 : Chiffres clés des sous-reconstructions pour la souche marine.

Assemblage	Réseau	Nombre de Réactions	Nombre de Métabolites				Nombre de Gènes	Couverture métabolique
			Total	Seulement consommés	Seulement produits	Consommés et produits		
Flye (9 546 séquences)	Orthologie	*4 114*	*4 458*	*1 146*	*1 297*	*2 015*	4 132	43.29 %
	Annotations	*2 924*	*3 291*	916	931	*1 444*	*2 762*	28.93 %
	Draft	*4 511*	*4 804*	*1 239*	*1 382*	*2 183*	*4 527*	47.42 %
Canu (9 431 séquences)	Orthologie	*4 096*	*4 455*	*1 158*	*1 308*	*1 989*	*4 099*	43.46 %
	Annotations	*2 908*	*3 295*	932	957	*1 406*	*2 690*	28.52 %
	Draft	*4 494*	*4 807*	*1 254*	*1 411*	*2 142*	*4 447*	47.15 %

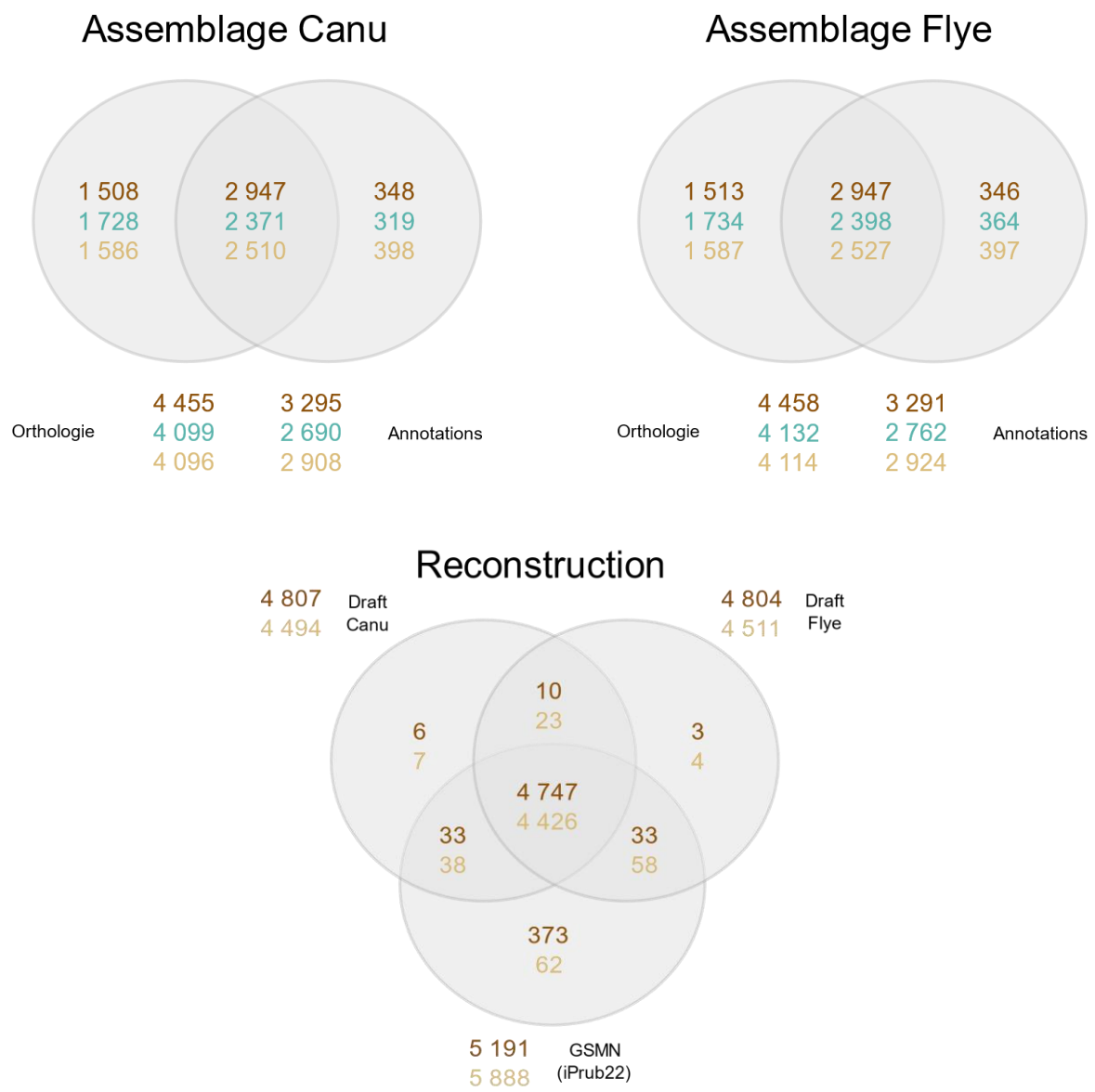


Figure 4-12 : Comparaisons du nombre de métabolites (■), de gènes (■) et de réactions (■) au sein des sous-réseaux et des drafts obtenus pour la souche MMS5. *Les deux premiers diagrammes de Venns montrent la complémentarité des approches apportées par les sous-réseaux issus de la recherche d'orthologie et de l'annotation fonctionnelle. Le troisième nous indique que 19 métabolites et 34 réactions semblent spécifiques à la souche marine.*

5 - Quel avenir pour *i*Prub22 : entre pérennité et obsolescence programmée ?

Aujourd'hui, soit cinq ans après le lancement des travaux présentés dans ce manuscrit, quelle perspective pouvons-nous envisager pour le modèle ***i*Prub22** ? Conçu pour offrir une modélisation métabolique à jour, ce modèle visait à intégrer le métabolisme spécialisé de *P. rubens*. Dans cette optique, une ré-annotation fonctionnelle a été entreprise, associée à une réconciliation rigoureuse des données disponibles. Cependant, l'intégration du métabolisme spécialisé demeure partielle, largement limitée par les ressources des bases de données métaboliques auxquelles il se rattache.

La « Connaissance Scientifique », et les données qui en découlent évoluent très rapidement. Avec la montée en puissance et l'engouement suscité par l'ère de l'intelligence artificielle et des approches d'apprentissage profond, nous pourrions bientôt assister à un nouveau « *bloom* » de données. Néanmoins, atteindre un niveau de généralisation et une stabilité des résultats demeurent des processus complexes aux frontières incertaines. Si les simulations d'***i*Prub22** sont reproductibles – un engagement que nous avons tenu – leur répétabilité (*i.e.* extrapolation vers d'autres organismes) en revanche reste plus incertaine. Le facteur temporel dans l'acquisition et le traitement des données devient alors une variable fondamentale pour la modélisation des **GSMNs**. Qualifier un **GSMN** de « *snapshot* » des connaissances s'avère ainsi pertinent, illustrant la dualité et la tension entre l'instantanéité d'un modèle statique, figé dans le temps, et la dynamique rapide de l'évolution des données, des outils et des infrastructures scientifiques. Si nous exagérons le trait : ce caractère instantané pourrait rendre le **GSMN** obsolète dès sa création.

Par ailleurs, certains éléments du processus de reconstruction d'***i*Prub22** se révèlent aujourd'hui fragilisés. Par exemple, la base de données MetaCyc, initialement choisie pour sa richesse en annotations et son caractère libre d'accès, tant pour la consultation des données que pour ses outils associés (*e.g.* SmartTables), nécessite désormais une licence payante, freinant l'accès aux ressources et compromettant l'objectif de Science Ouverte. Cette instabilité touche le cœur même d'***i*Prub22** puisque désormais la souche Wisconsin 54-1255 n'est plus la souche de référence de l'espèce *P. rubens* et a été remplacée par la souche IBT 27055 (*i.e.* numéro d'accession RefSeq GCF_028828025.1 – assemblage publié en février 2023). Se faisant, les données utilisées dans notre modèle possèdent aujourd'hui le statut « *suppressed* » sur le site du NCBI.

Au regard de ces éléments et évolutions, une question s'impose : un chercheur désireux aujourd'hui d'explorer le métabolisme spécialisé de *P. rubens*, serait-il mieux avisé de poursuivre avec ***i*Prub22** ou d'entreprendre une nouvelle reconstruction/réconciliation, à l'instar de ce que nous avons nous-même réalisé ? Afin d'atténuer cet écueil, nous avons veillé à intégrer autant que possible les standards actuels en matière de transparence, d'interopérabilité et de diffusion, avec pour objectif de faciliter la réutilisation d'***i*Prub22** et son adaptation aux évolutions futures.

Conclusion Générale

L'objectif initial de cette thèse était de développer une approche basée sur la biologie des systèmes pour rationaliser l'accès aux produits naturels fongiques. Plus spécifiquement, nous souhaitions déterminer si un modèle de réseau métabolique fongique pouvait être utilisé pour analyser les relations entre la disponibilité des nutriments et la production de métabolites spécialisés. Cette démarche visait à mieux comprendre, et idéalement à prédire, les conditions optimales de culture favorisant la biosynthèse de composés d'intérêt. Néanmoins, pour atteindre cet objectif, le recours à un modèle actualisé intégrant le métabolisme spécialisé de l'organisme s'est révélé indispensable. En vue d'une utilisation prolongée et facilitée, ce modèle devait également être adapté aux standards de diffusion pour accroître son accessibilité. La reconstruction du modèle ***i*Prub22** a couvert une part prépondérante de nos recherches, et, tout au long de nos travaux, nous avons été confrontés à divers obstacles qui trouvent une origine commune dans la même question fondamentale : quelle est la définition précise des concepts considérés (*e.g.* modèle, qualité, métabolisme spécialisé, standard, *etc.*) ?

Le **GSMN *i*Prub22** développé au cours de ces travaux est enrichi d'un nombre notable de réactions et de métabolites interconnectés, surpassant ainsi ses prédécesseurs. Cette amélioration devrait ainsi offrir une cartographie plus précise des interactions métaboliques afin d'explorer le métabolisme de *P. rubens* Wisconsin 54-1255, tant basal que spécialisé, et ce, sous divers scénarios environnementaux. La compartimentation intracellulaire, bien que cruciale pour représenter fidèlement la physiologie fongique, impose des défis substantiels dans la gestion et la curation des données. Toutefois, malgré cela, notre modèle est fonctionnel et produit des réponses de simulations de croissance en accord avec les attentes biologiques. En outre, nos simulations révèlent que la production de métabolites spécialisés varie en fonction des conditions d'import, suggérant ainsi que, même de légères modifications du milieu peuvent induire des changements quantitatifs ou qualitatifs des métabolites produits. Ensemble, ces avancées ouvrent la voie à une approche **OSMAC *in silico*** adaptable et extensible.

Notre **GSMN *i*Prub22** a été conçu non seulement pour répondre aux besoins des travaux présentés ici, mais également pour constituer une plateforme de ressources pour quiconque souhaitant approfondir les connaissances sur cet organisme. Ainsi, les principes de traçabilité et de transparence ont été placés au cœur de nos travaux, afin de garantir une reconstruction, et par extension des modèles, conformes aux normes actuelles, afin de sécuriser leur pérennité tout en facilitant leur utilisation et leur expansion future.

La reconstruction d'un **GSMN** et sa transformation en modèle fonctionnel ont pour objectif de fournir une représentation idéalisée, et de ce fait simplifiée, d'un phénomène biologique. À ce titre, il est nécessaire de trouver un compromis équilibré entre la complexité du modèle, sa facilité d'utilisation et le niveau de simplification tolérable. Tout au long de ces travaux, nous avons montré qu'***i*Prub22** est certes perfectible à bien des égards, mais qu'il fournit également les ressources nécessaires pour établir une base solide dans l'exploration des capacités métaboliques de *P. rubens* Wisconsin 54-1255, tant en termes de conformité aux standards de modélisation, qu'en termes de contenu.

Pour illustrer ces dires, nous souhaitons conclure ce manuscrit par une citation, qui aura été le fil conducteur et révélateur de nos travaux, attribuée au statisticien britannique George E. P. Box qui déclarait, "*Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful*" *(Box 1976)*.

Références

A

Abatenh E, Gizaw B, Tsegaye Z, Wassie M, *et al.* The Role of Microorganisms in Bioremediation- A Review. Open J Environ Biol. 10 nov 2017;2(1):038-46.

Acevedo A, Conejeros R, Aroca G. Ethanol production improvement driven by genome-scale metabolic modeling and sensitivity analysis in Scheffersomyces stipitis. PLOS ONE. 28 juin 2017;12(6):e0180074.

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, *et al.* Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 2019;66(1):4-119.

Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. Maranas CD, éditeur. PLoS Comput Biol. 21 mars 2013;9(3):e1002980.

Aguilar-Pontes MV, Brandl J, McDonnell E, Strasser K, Nguyen TTM, Riley R, *et al.* The gold-standard genome of *Aspergillus niger* NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi. Stud Mycol. 1 sept 2018;91:61-78.

Aichem M, Czauderna T, Zhu Y, Zhao J, Klapperstück M, Klein K, *et al.* Visual exploration of large metabolic models. Bioinformatics. 1 déc 2021;37(23):4460-8.

Aichem M, Klein K, Czauderna T, Garkov D, Zhao J, Li J, *et al.* Towards a hybrid user interface for the visual exploration of large biomolecular networks using virtual reality. J Integr Bioinforma. 2022;19(4):20220034.

Aimo L, Liechti R, Hyka-Nouspikel N, *et al.* The SwissLipids knowledgebase for lipid biology. Bioinformatics. 2015;31(17):2860-2866. doi:10.1093/bioinformatics/btv285

Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, *et al.* Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. Nielsen J, éditeur. PLOS Comput Biol. 23 mai 2018;14(5):e1006146.

Alam MT, Medema MH, Takano E, Breitling R. Comparative genome-scale metabolic modeling of actinomycetes: The topology of essential core metabolism. FEBS Lett. 21 juill 2011;585(14):2389-94.

Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, *et al.* Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. Npj Digit Med. 25 nov 2019;2(1):1-11.

Ali H, Ries MI, Lankhorst PP, van der Hoeven RAM, Schouten OL, Noga M, *et al.* A Non-Canonical NRPS Is Involved in the Synthesis of Fungisporin and Related Hydrophobic Cyclic Tetrapeptides in *Penicillium chrysogenum*. PLoS ONE. 2 juin 2014;9(6):e98212.

Ali H, Ries MI, Nijland JG, Lankhorst PP, Hankemeier T, Bovenberg RAL, *et al.* A Branched Biosynthetic Pathway Is Involved in Production of Roquefortine and Related Compounds in *Penicillium chrysogenum*. PLOS ONE. 12 juin 2013;8(6):e65328.

Allam AM, Elzainy TA. Degradation of xanthine by *Penicillium chrysogenum*. J Gen Microbiol. juin 1969;56(3):293-300.

Allard PM, Gaudry A, Quirós-Guerrero LM, Rutz A, Dounoue-Kubo M, Walker TWN, *et al.* Open and reusable annotated mass spectrometry dataset of a chemodiverse collection of 1,600 plant extracts. GigaScience. 1 janv 2023;12:giac124.

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 1 nov 2017;33(21):3387-95.

Amara A., Frainay, C., Jourdan, F., Naake, T., Neumann, S., Novoa-Del-Toro, E. M., Salek, R. M., Salzer, L., Scharfenberg, S., & Witting, M.. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. Frontiers in molecular biosciences, 2022, 9, 841373. https://doi.org/10.3389/fmolb.2022.841373

Aminian-Dehkordi J, Mousavi SM, Jafari A, Mijakovic I, Marashi SA. Manually curated genome-scale reconstruction of the metabolic network of Bacillus megaterium DSM319. Sci Rep. 10 déc 2019;9(1):18762.

Anand S, Mukherjee K, Padmanabhan P. An insight to flux-balance analysis for biochemical networks. Biotechnol Genet Eng Rev. 9 déc 2020;0(0):1-24.

Ankeny RA, Leonelli S. What's so special about model organisms? Stud Hist Philos Sci Part A. 1 juin 2011;42(2):313-23.

Antonakoudis A, Barbosa R, Kotidis P, Kontoravdi C. The era of big data: Genome-scale modelling meets machine learning. Comput Struct Biotechnol J. 1 janv 2020;18:3287-300.

Arora D, Gupta P, Jaglan S, Roullier C, Grovel O, Bertrand S. Expanding the chemical diversity through microorganisms co-culture: Current status and outlook. Biotechnol Adv. 1 mai 2020;40:107521.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* Gene Ontology: tool for the unification of biology. Nat Genet. mai 2000;25(1):25-9.

Ashtekar N, Anand G, Thulasiram HV, Rajeshkumar KC. Genus *Penicillium*: Advances and application in the modern era. In: Singh J, Gehlot P, éditeurs. New and Future Developments in Microbial Biotechnology and Bioengineering. Elsevier; 2021. p. 201-13. Disponible sur: https://www.sciencedirect.com/science/article/pii/B9780128210055000144

Ausländer S, Ausländer D, Fussenegger M. Synthetic Biology—The Synthesis of Biology. Angew Chem Int Ed. 2017;56(23):6396-419.

Alzahrani EO, El-Dessoky MM, Dogra P. Global dynamics of a cell quota-based model of light-dependent algae growth in a chemostat. Commun Nonlinear Sci Numer Simul. 1 nov 2020;90:105295.

B

Badri A, Raman K, Jayaraman G. Uncovering Novel Pathways for Enhancing Hyaluronan Synthesis in Recombinant *Lactococcus lactis*: Genome-Scale Metabolic Modeling and Experimental Validation. Processes. juin 2019;7(6):343.

Bailey JE. Toward a science of metabolic engineering. Science. 21 juin 1991;252(5013):1668-75.

Baker M. 1,500 scientists lift the lid on reproducibility. Nat News. 26 mai 2016;533(7604):452.

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. févr 2004;5(2):101-13.

Baron NC, Rigobelo EC, Zied DC, Baron NC, Rigobelo EC, Zied DC. Filamentous fungi in biological control: current status and future perspectives. Chil J Agric Res. juin 2019;79(2):307-15.

Barredo JL, Cantoral JM, Alvarez E, Díez B, Martín JF. Cloning, sequence analysis and transcriptional study of the isopenicillin N synthase of *Penicillium chrysogenum* AS-P-78. Mol Gen Genet MGG. 1 mars 1989;216(1):91-8.

Barreiro C, Martín JF, García-Estrada C. Proteomics Shows New Faces for the Old Penicillin Producer *Penicillium chrysogenum*. J Biomed Biotechnol. 2012;2012:1-15.

Bartoszewska M, Opaliński Ł, Veenhuis M, van der Klei IJ. The significance of peroxisomes in secondary metabolite biosynthesis in filamentous fungi. Biotechnol Lett. 10 juin 2011;33(10):1921.

Barzee TJ, Cao L, Pan Z, Zhang R. Fungi for future foods. J Future Foods. 1 sept 2021;1(1):25-37.

Beck AE, Hunt KA, Carlson RP. Measuring Cellular Biomass Composition for Computational Biology Applications. Processes. mai 2018;6(5):38.

Bekaert M. Reconstruction of *Danio rerio* Metabolic Model Accounting for Subcellular Compartmentalisation. PLOS ONE. 14 nov 2012;7(11):e49903.

Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, *et al.* Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res. 2021;Epub. doi:10.1093/nar/gkab1016.

Belcour A, Frioux C, Aite M, Bretaudeau A, Hildebrand F, Siegel A. Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species. Zambrano MM, éditeur. eLife. 29 déc 2020;9:e61968.

Belcour A., Got, J., Aite, M., Delage, L., Collén, J., Frioux, C., *et al.* Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe. Genome research, 2023, 33(6), 972–987. https://doi.org/10.1101/gr.277056.122

Bergmann FT, Adams R, Moodie S, Cooper J, Glont M, Golebiewski M, *et al.* COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. BMC Bioinformatics. 14 déc 2014;15(1):369.

Berlinck RGS, Monteiro AF, Bertonha AF, Bernardi DI, Gubiani JR, Slivinski J, *et al.* Approaches for the isolation and identification of hydrophilic, light-sensitive, volatile and minor natural products. Nat Prod Rep. 17 juill 2019;36(7):981-1004.

Bernasconi A, Masseroli M. Biological and Medical Ontologies: Systems Biology Ontology (SBO). In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, éditeurs. Encyclopedia of Bioinformatics and Computational Biology. Oxford: Academic Press; 2019b. p. 858-66. Disponible sur: https://www.sciencedirect.com/science/article/pii/B9780128096338203993

Bernstein DB, Sulheim S, Almaas E, Segrè D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. Genome Biol. 18 févr 2021;22(1):64.

Bertile F, Matallana-Surget S, Tholey A, Cristobal S, Armengaud J. Diversifying the concept of model organisms in the age of omics. Commun Biol. 19 oct 2023;6(1):1-4.

Bertrand S, Roullier C, Guitton Y. Successes and Pitfalls in Automated Dereplication Strategy Using Mass Spectrometry Data: a CASMI Experience. Curr Metabolomics. 1 avr 2017;5(1):25-34.

Bertrand S, Schumpp O, Bohni N, Monod M, Gindro K, Wolfender JL. *De novo* Production of Metabolites by Fungal Co-culture of *Trichophyton rubrum* and *Bionectria ochroleuca*. J Nat Prod. 28 juin 2013;76(6):1157-65.

Bhadury P, Mohammad BT, Wright PC. The current status of natural products from marine fungi and their potential as anti-infective agents. J Ind Microbiol Biotechnol. 21 janv 2006;33(5):325.

Biggs MB, Medlock GL, Kolling GL, Papin JA. Metabolic network modeling of microbial communities. WIREs Syst Biol Med. 2015;7(5):317-34.

Bind S., Bind, S., Sharma, A. K., & Chaturvedi, P. Epigenetic Modification: A Key Tool for Secondary Metabolite Production in Microorganisms. Frontiers in microbiology, 2022, 13, 784109. https://doi.org/10.3389/fmicb.2022.784109

Blin K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic acids research, 2023, 51(W1), W46–W50. https://doi.org/10.1093/nar/gkad344

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 2 juill 2021;49(W1):W29-35.

Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2 juill 2019;47(W1):W81-7.

Blunt JW, Copp BR, Munro MHG, Northcote PT, Prinsep MR. Marine natural products. Nat Prod Rep. 25 janv 2011;28(2):196-268.

Blunt JW, Copp BR, Keyzers RA, Munro MHG, Prinsep MR. Marine natural products. Nat Prod Rep. 17 mars 2017;34(3):235-94.

Bode HB, Bethe B, Höfs R, Zeeck A. Big Effects from Small Changes: Possible Ways to Explore Nature's Chemical Diversity. ChemBioChem. 2002;3(7):619-27.

Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet. juill 2018;26(7):931-6.

Bogart E, Myers CR. Multiscale Metabolic Modeling of C4 Plants: Connecting Nonlinear Genome-Scale Models to Leaf-Scale Metabolism in Developing Maize Leaves. PLOS ONE. 18 mars 2016;11(3):e0151722.

Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet. févr 2014;15(2):107-20.

Borenstein E, Feldman MW. Topological Signatures of Species Interactions in Metabolic Networks. J Comput Biol. févr 2009;16(2):191-200.

Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci. 23 sept 2008;105(38):14482-7.

Bornstein BJ, Keating SM, Jouraku A, Hucka M. LibSBML: an API Library for SBML. Bioinformatics. 15 mars 2008;24(6):880-1.

Borer B., Magnúsdóttir, S. The *media* composition as a crucial element in high-throughput metabolic network reconstruction. Interface focus, 2023, 13(2), 20220070. https://doi.org/10.1098/rsfs.2022.0070

Box GEP. Science and Statistics. J Am Stat Assoc. 1976;71(356):791-9.

Brandl J, Andersen MR. Aspergilli: Models for systems biology in filamentous fungi. Curr Opin Syst Biol. 1 déc 2017;6:67-73.

Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. 2005;33(10):3390-400.

Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, *et al.* A Tissue-Mapped Axolotl *De novo* Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 17 janv 2017;18(3):762-76.

Burgard AP, Pharkya P, Maranas CD. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng. 2003;84(6):647-57.

Burns M. The development of penicillin in the netherlands 1940-1950: the pivotal role of nv nederlandsche gist- en spiritusfabriek, delft.

C

Caesar LK, Butun FA, Robey MT, Ayon NJ, Gupta R, Dainko D, *et al.* Correlative metabologenomics of 110 fungi reveals metabolite–gene cluster pairs. Nat Chem Biol. juill 2023;19(7):846-54.

Camborda S, Weder JN, Töpfer N. CobraMod: a pathway-centric curation tool for constraint-based metabolic models. Bioinformatics. 28 avr 2022;38(9):2654-6.

Camila Dos Santos Dias A, Couzinet-Mossion A, Ruiz N, Le Bellec M, Gentil E, Wielgosz-Collin G, *et al.* Sugar Induced Modification in Glycolipid Production in *Acremonium* sp. Revealed by LC-MS Lipidomic Approach. Curr Biotechnol. 1 août 2017;6(3):227-37.

Carey MA, Dräger A, Beber ME, Papin JA, Yurkovich JT. Community standards to facilitate development and address challenges in metabolic modeling. Mol Syst Biol. août 2020;16(8):e9235.

Carroll AR, Copp BR, Davis RA, Keyzers RA, Prinsep MR. Marine natural products. Nat Prod Rep. 22 févr 2023;40(2):275-325.

Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 8 janv 2020;48(D1):D445-53.

Castillo S, Barth D, Arvas M, Pakula TM, Pitkänen E, Blomberg P, *et al.* Whole-genome metabolic model of Trichoderma reesei built by comparative reconstruction. Biotechnol Biofuels. 2016;9:252.

Cavalier-Smith T. What are Fungi?. In: McLaughlin, D.J., McLaughlin, E.G., Lemke, P.A. (eds) Systematics and Evolution. The Mycota, 2001, vol 7A. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-10376-0_1

Chain E, Florey HW, Gardner AD, Heatley NG, Jennings MA, Orr-Ewing J, *et al.* Penicillin as a chemotherapeutic agent. Lancet. 1940;236:226-8.

Challis GL. Genome Mining for Novel Natural Product Discovery. J Med Chem. 1 mai 2008;51(9):2618-28.

Chambergo FS, Valencia EY. Fungal biodiversity to biotechnology. Appl Microbiol Biotechnol. 1 mars 2016;100(6):2567-77.

Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 20 sept 2021;49(16):9077-96.

Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021;49. doi:10.1093/nar/gkaa1025. PubMed: 33211880.

Chassagne F, Cabanac G, Hubert G, David B, Marti G. The landscape of natural product diversity and their pharmacological relevance from a focus on the Dictionary of Natural Products®. Phytochem Rev. 1 juin 2019;18(3):601-22.

Chaudhary VB, Aguilar-Trigueros CA, Mansour I, Rillig MC. Fungal Dispersal Across Spatial Scales. Annu Rev Ecol Evol Syst. 2 nov 2022;53(Volume 53, 2022):69-85.

Chen Y, Banerjee D, Mukhopadhyay A, Petzold CJ. Systems and synthetic biology tools for advanced bioproduction hosts. Curr Opin Biotechnol. 1 août 2020;64:101-9.

Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, Ramos-Aboites HE, *et al.* Evolutionary dynamics of natural product biosynthesis in bacteria. Nat Prod Rep. 29 avr 2020;37(4):566-99.

Chiang CY, Ohashi M, Tang Y. Deciphering chemical logic of fungal natural product biosynthesis through heterologous expression and genome mining. Nat Prod Rep. 25 janv 2023;40(1):89-127.

Chiang YM, Ahuja M, Oakley CE, Entwistle R, Asokan A, Zutz C, *et al.* Development of Genetic Dereplication Strains in *Aspergillus nidulans* Results in the Discovery of Aspercryptin. Angew Chem. 2016;128(5):1694-7.

Cho JS, Kim GB, Eun H, Moon CW, Lee SY. Designing Microbial Cell Factories for the Production of Chemicals. JACS Au. 22 août 2022;2(8):1781-99.

Choi HS, Lee SY, Kim TY, Woo HM. *In silico* Identification of Gene Amplification Targets for Improvement of Lycopene Production. Appl Environ Microbiol. 15 mai 2010;76(10):3097-105.

Choudhury P, Crowston K, Dahlander L, Minervini MS, Raghuram S. GitLab: work where you want, when you want. J Organ Des. 16 nov 2020;9(1):23.

Chroumpi T, Mäkelä MR, de Vries RP. Engineering of primary carbon metabolism in filamentous fungi. Biotechnol Adv. 11 mai 2020;107551.

Chu R, Li S, Zhu L, Yin Z, Hu D, Liu C, *et al.* A review on co-cultivation of microalgae with filamentous fungi: Efficient harvesting, wastewater treatment and biofuel production. Renew Sustain Energy Rev. 1 avr 2021;139:110689.

Chung CH, Lin DW, Eames A, Chandrasekaran S. Next-Generation Genome-Scale Metabolic Modeling through Integration of Regulatory Mechanisms. Metabolites. sept 2021;11(9):606.

Chung WY, Zhu Y, Mahamad Maifiah MH, Shivashekaregowda NKH, Wong EH, Abdul Rahim N. Novel antimicrobial development using genome-scale metabolic model of Gram-negative pathogens: a review. J Antibiot (Tokyo). 8 sept 2020;1-10.

Community standards to facilitate development and address challenges in metabolic modeling. Mol Syst Biol. août 2020;16(8):e9235.

Conroy MJ, Andrews RM, Andrews S, *et al.* LIPID MAPS: update to databases and tools for the lipidomics community. Nucleic Acids Res. 2023;51(D1)–D579. doi:10.1093/nar/gkad896. PubMed ID: 37855672.

Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinforma Oxf Engl. 15 sept 2017;33(18):2938-40.

Cordes H, Thiel C, Baier V, Blank LM, Kuepfer L. Integration of genome-scale metabolic networks into whole-body PBPK models shows phenotype-specific cases of drug-induced metabolic perturbation. Npj Syst Biol Appl. 26 févr 2018;4(1):1-11.

Cottret L, Frainay C, Chazalviel M, Cabanettes F, Gloaguen Y, Camenen E, *et al.* MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res. 2 juill 2018;46(W1):W495-502.

Cottret L, Jourdan F. Graph methods for the investigation of metabolic networks in parasitology. Parasitology. août 2010;137(9):1393-407.

Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, *et al.* Controlled vocabularies and semantics in systems biology. Mol Syst Biol. janv 2011;7(1):543.

Cox RJ. Engineered and total biosynthesis of fungal specialized metabolites. Nat Rev Chem. janv 2024;8(1):61-78.

Coyle CM, Panaccione DG. An Ergot Alkaloid Biosynthesis Gene and Clustered Hypothetical Genes from *Aspergillus fumigatus*. Appl Environ Microbiol. juin 2005;71(6):3112-8.

D

D'Ari R, Casadesús J. Underground metabolism. BioEssays. 1998;20(2):181-6.

David B, Wolfender JL, Dias DA. The pharmaceutical industry and natural products: historical status and new trends. Phytochem Rev. avr 2015;14(2):299-315.

Davis RH. The age of model organisms. Nat Rev Genet. janv 2004;5(1):69-76.

Deantas-Jahn C, Mendoza SN, Licona-Cassani C, Orellana C, Saa PA. Metabolic modeling of *Halomonas campaniensis* improves polyhydroxybutyrate production under nitrogen limitation. Appl Microbiol Biotechnol. 25 avr 2024;108(1):310.

De la Cruz Quiroz R, Roussos S, Hernández D, Rodríguez R, Castillo F, Aguilar CN. Challenges and opportunities of the bio-pesticides production by solid-state fermentation: filamentous fungi as a model. Crit Rev Biotechnol. 3 juill 2015;35(3):326-33.

De Martino D, Capuani F, Mori M, De Martino A, Marinari E. Counting and Correcting Thermodynamically Infeasible Flux Cycles in Genome-Scale Metabolic Networks. Metabolites. déc 2013;3(4):946-66.

Demain A. L., & Fang, A. The natural functions of secondary metabolites. Advances in biochemical engineering/biotechnology, 2000, 69, 1–39. https://doi.org/10.1007/3-540-44964-7_1

Demain AL, Sanchez S. Microbial drug discovery: 80 years of progress. J Antibiot (Tokyo). janv 2009;62(1):5-16.

Desouki AA, Jarre F, Gelius-Dietrich G, Lercher MJ. CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. Bioinformatics. 1 juill 2015;31(13):2159-65.

Deveau A, Bonito G, Uehling J, Paoletti M, Becker M, Bindschedler S, *et al.* Bacterial–fungal interactions: ecology, mechanisms and challenges. FEMS Microbiol Rev. 1 mai 2018;42(3):335-52.

Dias DA, Urban S, Roessner U. A Historical Overview of Natural Products in Drug Discovery. Metabolites. juin 2012;2(2):303-36.

Dittami SM, Corre E. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp *Saccharina japonica* using Taxoblast. PeerJ. 17 nov 2017;5:e4073.

Dougherty BV, Moutinho Jr., T. J., & Papin, J. Accelerating the Drug Development Pipeline with Genome-Scale Metabolic Network Reconstructions. In: Systems Metabolic Engineering, 2017, Wiley-VCH, pp. 139-162. https://doi.org/10.1002/9783527696130.ch5

Dräger A, Palsson BØ. Improving collaboration by standardization efforts in systems biology. Front Bioeng Biotechnol. 2014;2:61. Published 2014 Dec 8. doi:10.3389/fbioe.2014.00061

Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE. Reconstruction and Validation of a Genome-Scale Metabolic Model for the Filamentous Fungus *Neurospora crassa* Using FARM. PLOS Comput Biol. 18 juill 2013;9(7):e1003126.

Droste P, Miebach S, Niedenführ S, Wiechert W, Nöh K. Visualizing multi-omics data in metabolic networks with the software Omix: a case study. Biosystems. août 2011;105(2):154-61.

Duarte NC, Palsson BØ, Fu P. Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae*. BMC Genomics. 8 sept 2004;5:63.

Dufossé L, Fouillaud M, Caro Y, Mapari SA, Sutthiwong N. Filamentous fungi are large-scale producers of pigments and colorants for the food industry. Curr Opin Biotechnol. 1 avr 2014;26:56-61.

Dusad V., Thiel, D., Barahona, M., Keun, H. C., & Oyarzún, D. A.. Opportunities at the Interface of Network Science and Metabolic Modeling. Frontiers in bioengineering and biotechnology, 2021, 8, 591049. https://doi.org/10.3389/fbioe.2020.591049

E

Ebrahim A, Almaas E, Bauer E, Bordbar A, Burgard AP, Chang RL, *et al.* Do genome-scale models need exact solvers or clearer standards? Mol Syst Biol. oct 2015;11(10):831.

Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013;7(1):74.

Edwards JS, Palsson BO. Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype. J Biol Chem. 18 juin 1999;274(25):17410-6.

Edwards JS, Ramakrishna R, Palsson BO. Characterizing the metabolic phenotype: A phenotype phase plane analysis. Biotechnol Bioeng. 2002;77(1):27-36.

Egbuta MA, Mwanza M, Babalola OO. Health Risks Associated with Exposure to Filamentous Fungi. Int J Environ Res Public Health. juill 2017;14(7):719.

Egli T. Microbial growth and physiology: a call for better craftsmanship. Front Microbiol. 2015;6:287.

Ekkers DM, Tusso S, Moreno-Gamez S, Rillo MC, Kuipers OP, van Doorn GS. Trade-Offs Predicted by Metabolic Network Structure Give Rise to Evolutionary Specialization and Phenotypic Diversification. Mol Biol Evol. 1 juin 2022;39(6):msac124.

El Hajj Assaf C, Zetina-Serrano C, Tahtah N, Khoury AE, Atoui A, Oswald IP, *et al.* Regulation of Secondary Metabolism in the *Penicillium* Genus. Int J Mol Sci. 12 déc 2020;21(24).

Elander RP. Industrial production of β-lactam antibiotics. Appl Microbiol Biotechnol. 1 juin 2003;61(5):385-92.

El-Hawary SS, Hassan MHA, Hudhud AO, Abdelmohsen UR, Mohammed R. Elicitation for activation of the actinomycete genome's cryptic secondary metabolite gene clusters. RSC Adv. 14 févr 2023;13(9):5778-95.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 14 nov 2019;20(1):238.

Erciyes K. Graph-Theoretical Analysis of Biological Networks: A Survey. Computation. oct 2023;11(10):188.

Ernster L, Schatz G. Mitochondria: a historical review. J Cell Biol. 22 févr 1981;91(3):227s-55s.

Eustáquio AS, Ziemert N. Identification of Natural Product Biosynthetic Gene Clusters from Bacterial Genomic Data. In: Methods in Pharmacology and Toxicology. Humana Press, 2018. https://doi.org/10.1007/7653_2018_32

F

Fabian SJ, Maust MD, Panaccione DG. Ergot Alkaloid Synthesis Capacity of *Penicillium camemberti*. Appl Environ Microbiol. 17 sept 2018;84(19):e01583-18.

Fang G, Bhardwaj N, Robilotto R, Gerstein MB. Getting Started in Gene Orthology and Functional Analysis. PLOS Comput Biol. 26 mars 2010;6(3):e1000703.

Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. Biochem Soc Trans. 20 août 2018;46(4):931-6.

Fatumo S, Plaimas K, Mallm JP, Schramm G, Adebiyi E, Oswald M, *et al.* Estimating novel potential drug targets of *Plasmodium falciparum* by analysing the metabolic network of knock-out strains *in silico*. Infect Genet Evol. 1 mai

Feher M, Schmidt JM. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. J Chem Inf Comput Sci. 1 janv 2003;43(1):218-27.

Feist AM, Palsson BO. The Biomass Objective Function. Curr Opin Microbiol. juin 2010;13(3):344-9.

Fell DA, Wagner A. The small world of metabolism. Nat Biotechnol. nov 2000;18(11):1121-2.

Fierro F, Barredo JL, Díez B, Gutierrez S, Fernández FJ, Martín JF. The penicillin gene cluster is amplified in tandem repeats linked by conserved hexanucleotide sequences. Proc Natl Acad Sci. 20 juin 1995;92(13):6200-4.

Fierro F, Vaca I, Castillo NI, García-Rico RO, Chávez R. *Penicillium chrysogenum*, a Vintage Model with a Cutting-Edge Profile in Biotechnology. Microorganisms. mars 2022;10(3):573.

Fischbach MA, Clardy J. One pathway, many products. Nat Chem Biol. 1 juill 2007;3(7):353-5.

Fisher CP, Plant NJ, Moore JB, Kierzek AM. QSSPN: dynamic simulation of molecular interaction networks describing gene regulation, signalling and whole-cell metabolism in human cells. Bioinformatics. 15 déc 2013;29(24):3181-90.

Fleming A. On the Antibacterial Action of Cultures of a *Penicillium*, with Special Reference to their Use in the Isolation of *B. influenzæ*. Br J Exp Pathol. juin 1929;10(3):226-36.

Förster J, Famili I, Fu P, Palsson BØ, Nielsen J. Genome-Scale Reconstruction of the *Saccharomyces cerevisiae* Metabolic Network. Genome Res. 2 janv 2003;13(2):244-53.

Frainay C, Schymanski E, Neumann S, Merlet B, Salek R, Jourdan F, *et al.* Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. Metabolites. 15 sept 2018;8(3):51.

Frisvad JC. Taxonomy, chemodiversity, and chemoconsistency of *Aspergillus*, *Penicillium*, and *Talaromyces* species. Front Microbiol. 2015;5:773. Published 2015 Jan 12. doi:10.3389/fmicb.2014.00773

Fritzemeier CJ, Hartleb D, Szappanos B, Papp B, Lercher MJ. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Comput Biol. 18 avr 2017;13(4):e1005494.

G

Gamermann D, Montagud A, Conejero JA, Urchueguía JF, de Córdoba PF. New Approach for Phylogenetic Tree Recovery Based on Genome-Scale Metabolic Networks. J Comput Biol. 10 mars 2014;21(7):508-19.

Gao J, Ellis LBM, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. Nucleic Acids Res. 2010;38–D491.

García-Estrada C, Martín JF, Cueto L, Barreiro C. Omics Approaches Applied to *Penicillium chrysogenum* and Penicillin Production: Revealing the Secrets of Improved Productivity. Genes. juin 2020;11(6):712.

Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, *et al.* Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat Microbiol. mai 2022;7(5):726-35.

Gaynes R. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. Emerg Infect Dis. mai 2017;23(5):849-53.

Gebauer J, Gentsch C, Mansfeld J, Schmeißer K, Waschina S, Brandes S, *et al.* A Genome-Scale Database and Reconstruction of *Caenorhabditis elegans* Metabolism. Cell Syst. 25 mai 2016;2(5):312-22.

Geers AU, Buijs Y, Strube ML, Gram L, Bentzon-Tilia M. The natural product biosynthesis potential of the microbiomes of Earth - Bioprospecting for novel anti-microbial agents in the meta-omics era. Comput Struct Biotechnol J. 2022;20:343-52.

Gerhards N, Neubauer L, Tudzynski P, Li SM. Biosynthetic Pathways of Ergot Alkaloids. Toxins. déc 2014;6(12):3281-95.

Gerlin L, Cottret L, Cesbron S, Taghouti G, Jacques MA, Genin S, *et al.* Genome-scale investigation of the metabolic determinants generating bacterial fastidious growth. mSystems. 2020 ;5. doi:10.1128/msystems.00698-19.

Ghosh S, Rusyn I, Dmytruk OV, Dmytruk KV, Onyeaka H, Gryzenhout M, *et al.* Filamentous fungi for sustainable remediation of pharmaceutical compounds, heavy metal and oil hydrocarbons. Front Bioeng Biotechnol. 2023;11:1106973.

Goldstein B, King N. The Future of Cell Biology: Emerging Model Organisms. Trends Cell Biol. 1 nov 2016;26(11):818-24.

Golubev A, Hanson AD, Gladyshev VN. Non-enzymatic molecular damage as a prototypic driver of aging. J Biol Chem. 14 avr 2017;292(15):6029-38.

Gouka RJ, van Hartingsveldt W, Bovenberg RAL, van den Hondel CAMJJ, van Gorcom RFM. Cloning of the nitrate - nitrite reductase gene cluster of *Penicillium chrysogenum* and use of the niaD gene as a homologous selection marker. J Biotechnol. 1 sept 1991;20(2):189-99.

Gould IM. Antibiotic resistance: the perfect storm. Int J Antimicrob Agents. 1 août 2009;34:S2-5.

Grand View Research. Bioremediation Market Size, Share & Trends Report, 2030. San Francisco, CA: Grand View Research, Inc.; 2023a. Disponible sur: https://www.grandviewresearch.com/industry-analysis/bioremediation-market-report

Grand View Research. Microbiology & Bacterial Culture For Industrial Testing Market Report 2030. San Francisco, CA: Grand View Research, Inc.; 2023b. Disponible sur: https://www.grandviewresearch.com/industry-analysis/microbiology-culture-market

Grand View Research. Agricultural Microbials Market Size & Share Report, 2030. San Francisco, CA: Grand View Research, Inc.; 2023c. Disponible sur: https://www.grandviewresearch.com/industry-analysis/agricultural-microbials-market-report

Grand View Research. Industrial Enzymes Market Size, Share, Growth Report, 2030. San Francisco, CA: Grand View Research, Inc.; 2023d. Disponible sur: https://www.grandviewresearch.com/industry-analysis/industrial-enzymes-market

Grand View Research. Antibiotics Market Size, Share, Growth & Trends Report 2030. San Francisco, CA: Grand View Research, Inc.; 2023e. Disponible sur: https://www.grandviewresearch.com/industry-analysis/antibiotic-market

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 1 janv 2014;42(D1):D699-704.

Grijseels S, Nielsen JC, Nielsen J, Larsen TO, Frisvad JC, Nielsen KF, *et al.* Physiological characterization of secondary metabolite producing *Penicillium* cell factories. Fungal Biol Biotechnol. 17 oct 2017;4(1):8.

Grimm LH, Kelly S, Krull R, Hempel DC. Morphology and productivity of filamentous fungi. Appl Microbiol Biotechnol. 1 déc 2005;69(4):375-84.

Gross H. Genomic mining--a concept for the discovery of new bioactive natural products. Curr Opin Drug Discov Devel. mars 2009;12(2):207-19.

Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol. 13 juin 2019;20(1):121.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. Nature. janv 2003;421(6918):63-6.

Gudmundsson S, Thiele I. Computationally efficient flux variability analysis. BMC Bioinformatics. 29 sept 2010;11(1):489.

Gujar VV, Fuke P, Khardenavis AA, Purohit HJ. Draft genome sequence of *Penicillium chrysogenum* strain HKF2, a fungus with potential for production of prebiotic synthesizing enzymes. 3 Biotech. 1 févr 2018;8(2):106.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 15 avr 2013;29(8):1072-5.

Guzmán-Chávez F, Salo O, Nygård Y, Lankhorst PP, Bovenberg RAL, Driessen AJM. Mechanism and regulation of sorbicillin biosynthesis by *Penicillium chrysogenum*. Microb Biotechnol. 2017;10(4):958-68.

Guzmán-Chávez F, Zwahlen RD, Bovenberg RAL, Driessen AJM. Engineering of the Filamentous Fungus *Penicillium chrysogenum* as Cell Factory for Natural Products. Front Microbiol. 2018;9:2768.

H

Haarmann T, Rolke Y, Giesbert S, Tudzynski P. Ergot: from witchcraft to biotechnology. Mol Plant Pathol. 2009;10(4):563-77.

Haas H. Fungal siderophore metabolism with a focus on *Aspergillus fumigatus*. Nat Prod Rep. 2014;31(10):1266-76.

Hädicke O, Klamt S. Computing complex metabolic intervention strategies using constrained minimal cut sets. Metab Eng. 1 mars 2011;13(2):204-13.

Hallgren J, Tsirigos KD, Pedersen MD, Armenteros JJA, Marcatili P, Nielsen H, *et al.* DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks, bioRxiv; 2022.. Disponible sur: https://www.biorxiv.org/content/10.1101/2022.04.08.487609v1

Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, *et al.* Metabolic Network Analysis and Metatranscriptomics Reveal Auxotrophies and Nutrient Sources of the Cosmopolitan Freshwater Microbial Lineage acI. mSystems. 29 août 2017;2(4):e00091-17.

Hari A, Lobo D. Fluxer: a web application to compute, analyze and visualize genome-scale metabolic flux networks. Nucleic Acids Res. 2 juill 2020;48(W1):W427-35.

Hari A, Zarrabi A, Lobo D. *mergem*: merging, comparing, and translating genome-scale metabolic models using universal identifiers, NAR Genomics and Bioinformatics, Volume 6, Issue 1, March 2024, lqae010, https://doi.org/10.1093/nargab/lqae010

Hashemi S, Razaghi-Moghadam Z, Nikoloski Z. Identification of flux trade-offs in metabolic networks. Sci Rep. 10 déc 2021;11(1):23776.

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016;44(D1)–D1219. doi:10.1093/nar/gkv1031.

Hattwell JPN, Hastings J, Casanueva O, Schirra HJ, Witting M. Using Genome-Scale Metabolic Networks for Analysis, Visualization, and Integration of Targeted Metabolomics Data. Methods Mol Biol. 2020;2104:361-386. doi:10.1007/978-1-0716-0239-3_18

Hawksworth DL, Crous PW, Redhead SA, Reynolds DR, Samson RA, Seifert KA, *et al.* The Amsterdam Declaration on Fungal Nomenclature. IMA Fungus. 1 juin 2011;2(1):105-11.

Heinken A, Thiele I. Systems biology of host–microbe metabolomics. WIREs Syst Biol Med. 2015;7(4):195-219.

Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, *et al.* Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. mars 2019;14(3):639-702.

Heller, S.R., McNaught, A., Pletnev, I. *et al.* InChI, the IUPAC International Chemical Identifier. J Cheminform 7, 23 (2015). https://doi.org/10.1186/s13321-015-0068-4

Henriksen CM, Christensen LH, Nielsen J, Villadsen J. Growth energetics and metabolic fluxes in continuous cultures of *Penicillium chrysogenum*. J Biotechnol. 28 févr 1996;45(2):149-64.

Henry CS, DeJongh M, Best AB, Frybarger PM, Linsay B, Stevens RL. High-throughput generation and optimization of genome-scale metabolic models. Nat Biotechnol. 2010.

Herbert D. The chemical composition of micro-organisms as a function of their environment. In: Symp Soc Gen Microbiol. 1961. p. 7.

Herrmann HA, Dyson BC, Vass L, Johnson GN, Schwartz JM. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. Npj Syst Biol Appl. 2 sept 2019;5(1):1-8.

Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, *et al.* A higher-level phylogenetic classification of the Fungi. Mycol Res. 1 mai 2007;111(5):509-47.

Hidalgo PI, Ullán RV, Albillos SM, Montero O, Fernández-Bodega MÁ, García-Estrada C, *et al.* Molecular characterization of the PR-toxin gene cluster in *Penicillium roqueforti* and *Penicillium chrysogenum*: Cross talk of secondary metabolite pathways. Fungal Genet Biol. 1 janv 2014;62:11-24.

Hoenigl M, Arastehfar A, Arendrup MC, Brüggemann R, Carvalho A, Chiller T, *et al.* Novel antifungals and treatment approaches to tackle resistance and improve outcomes of invasive fungal disease. Clin Microbiol Rev. 13 juin 2024;37(2):e0007423.

Holm DK, Petersen LM, Klitgaard A, Knudsen PB, Jarczynska ZD, Nielsen KF, *et al.* Molecular and Chemical Characterization of the Biosynthesis of the 6-MSA-Derived Meroterpenoid Yanuthone D in *Aspergillus niger*. Chem Biol. 24 avr 2014;21(4):519-29.

Hong SH, Kim TY, Lee SY. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl Microbiol Biotechnol. 1 août 2004;65(2):203-10.

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703-14.

Houbraken J, Frisvad JC, Samson RA. Fleming's penicillin producing strain is not *Penicillium chrysogenum* but *Penicillium rubens*. IMA Fungus. 1 juin 2011;2(1):87-95.

Houbraken J, Frisvad JC, Seifert KA, Overy DP, Tuthill DM, Valdez JG, *et al.* New penicillin-producing *Penicillium* species and an overview of section Chrysogena. Persoonia - Mol Phylogeny Evol Fungi. 31 déc 2012;29(1):78-100.

Houbraken J, Kocsubé S, Visagie CM, Yilmaz N, Wang XC, Meijer M, *et al.* Classification of *Aspergillus*, *Penicillium*, *Talaromyces* and related genera (Eurotiales): An overview of families, genera, subgenera, sections, series and species. Stud Mycol. 1 mars 2020;95:5-169.

Houbraken J, Samson RA. Phylogeny of *Penicillium* and the segregation of *Trichocomaceae* into three families. Stud Mycol. 1 sept 2011;70(1):1-51.

Hucka M, Bergmann F, Dräger A, Hoops S, Keating S, *et al.* The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core. Journal of Integrative Bioinformatics. 2018;15(1): 20170080. https://doi.org/10.1515/jib-2017-0080 Huxtable RJ, Schwarz SKW. The Isolation of Morphine—First Principles in Science and Ethics. Mol Interv. 10 janv 2001;1(4):189.

Hyde KD, Xu J, Rapior S, Jeewon R, Lumyong S, Niego AGT, *et al.* The amazing potential of fungi: 50 ways we can exploit fungi industrially. Fungal Divers. 1 juill 2019;97(1):1-136.

I

Iacovelli R, Bovenberg RAL, Driessen AJM. Nonribosomal peptide synthetases and their biotechnological potential in *Penicillium rubens*. J Ind Microbiol Biotechnol. 1 déc 2021;48(9-10):kuab045.

Imhoff JF. Natural Products from Marine Fungi—Still an Underrepresented Resource. Mar Drugs. janv 2016;14(1):19.

Immanuel SRC, Banerjee D, Rajankar MP, Raghunathan A. Integrated constraints based analysis of an engineered violacein pathway in *Escherichia coli*. Biosystems. 1 sept 2018;171:10-9

Isidro IA, Ferreira AR, Clemente JJ, Cunha AE, Oliveira R. Analysis of culture *media* screening data by projection to latent pathways: The case of Pichia pastoris X-33. J Biotechnol. 10 janv 2016;217:82-9.

J

Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, *et al.* FAIR Principles: Interpretations and Implementation Considerations. Data Intell. 1 janv 2020;2(1-2):10-29.

Jami MS, Barreiro C, García-Estrada C, Martín JF. Proteome Analysis of the Penicillin Producer *Penicillium chrysogenum*: Characterization of Protein Changes During the Industrial Strain Improvement. Mol Cell Proteomics. 1 juin 2010;9(6):1182-98.

Jang S, Kim M, Hwang J, Jung GY. Tools and systems for evolutionary engineering of biomolecules and microorganisms. J Ind Microbiol Biotechnol. 1 oct 2019;46(9):1313-26.

Jeffryes JG, Lerma-Ortiz C, Liu F, Golubev A, Niehaus TD, Elbadawi-Sidhu M, *et al.* Chemical-damage MINE: A database of curated and predicted spontaneous metabolic reactions. Metabolic Engineering, 2022, 69, 302-312. https://doi.org/10.1016/j.ymben.2021.11.009

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. oct 2000;407(6804):651-4.

Jerums M, Yang X. Optimization of cell culture media. BioProcess Int. 2005;3(6):38-44.

Jin S, Zeng X, Xia F, Huang W, Liu X. Application of deep learning methods in biological networks. Briefings in bioinformatics, 2021, 22(2), 1902–1917. https://doi.org/10.1093/bib/bbaa043

Joyce AR, Palsson BØ. Predicting Gene Essentiality Using Genome-Scale *in silico* Models. In: Osterman, A.L., Gerdes, S.Y. (eds) Microbial Gene Essentiality: Protocols and Bioinformatics. Methods in Molecular Biology™, 2008, vol 416. Humana Press. https://doi.org/10.1007/978-1-59745-321-9_30

K

Kalra R, Conlan XA, Goel M. Fungi as a Potential Source of Pigments: Harnessing Filamentous Fungi. Front. Chem. 2020, 8:369. doi: 10.3389/fchem.2020.00369

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 8 janv 2021;49(D1):D545-51.

Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci Publ Protein Soc. 2020;29(1):28-35.

Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. Protein Sci. 2022;31(1):47-53.

Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, *et al.* The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 19 juill 2019a;20(4):1085-93.

Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, *et al.* Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 1 janv 2021;22(1):109-26.

Karp PD, Midford PE, Billington R, *et al.* Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 2021;22(1):109-126. doi:10.1093/bib/bbz104

Katz J, Rognstad R. Futile Cycles in the Metabolism of Glucose. In: Horecker BL, Stadtman ER, eds. Current Topics in Cellular Regulation. Vol 10. Academic Press; 1976, 237-289. doi:10.1016/B978-0-12-152810-2.50013-9.

Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Curr Opin Biotechnol. 1 oct 2003;14(5):491-6.

Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 3 juill 2017;45(W1):W55-63.

Keating SM, Waltemath D, König M, Zhang F, Dräger A, Chaouiya C, *et al.* SBML Level 3: an extensible format for the exchange and reuse of biological models. Mol Syst Biol. 1 août 2020;16(8):e9110.

Kędzia K, Ptak W, Sroka J, Kierzek AM. Simulation of multicellular populations with Petri nets and genome scale intracellular networks. Sci Comput Program. 1 juin 2018;157:3-16.

Keller MA, Piedrafita G, Ralser M. The widespread role of non-enzymatic reactions in cellular metabolism. Curr Opin Biotechnol. 1 août 2015;34:153-61.

Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol. mars 2019;17(3):167-80.

Keller NP, Turner G, Bennett JW. Fungal secondary metabolism — from biochemistry to genomics. Nat Rev Microbiol. déc 2005;3(12):937-47.

Kiel JAKW, van der Klei IJ, van den Berg MA, Bovenberg RAL, Veenhuis M. Overproduction of a single protein, Pc-Pex11p, results in 2-fold enhanced penicillin production by *Penicillium chrysogenum*. Fungal Genet Biol. 1 févr 2005;42(2):154-64.

Kim GB, Choi SY, Cho IJ, Ahn DH, Lee SY. Metabolic engineering for sustainability and health. Trends Biotechnol. 1 mars 2023;41(3):425-51.

Kim HU, Blin K, Lee SY, Weber T. Recent development of computational resources for new antibiotics discovery. Curr Opin Microbiol. 1 oct 2017a;39:113-20.

Kim J, Reed JL. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Syst Biol. 28 avr 2010;4(1):53.

Kim M, Park BG, Kim J, Kim JY, Kim BG. Exploiting transcriptomic data for metabolic engineering: toward a systematic strain design. Curr Opin Biotechnol. 1 déc 2018a;54:26-32.

Kim M, Sang Yi J, Kim J, Kim JN, Kim MW, Kim BG. Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in *Streptomyces coelicolor* A3(2). Biotechnol J. 2014;9(9):1185-94.

Kim S, Chen J, Cheng T, *et al.* PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373–D1380. doi:10.1093/nar/gkac956

Kim TY, Kim HU, Lee SY. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. Metab Eng. 1 mars 2010;12(2):105-11.

Kim WJ, Kim HU, Lee SY. Current state and applications of microbial genome-scale metabolic models. Curr Opin Syst Biol. 1 avr 2017b;2:10-8.

Kim YM, Poline JB, Dumas G. Experimenting with reproducibility: a case study of robustness in bioinformatics, GigaScience, Volume 7, Issue 7, July 2018, giy077, https://doi.org/10.1093/gigascience/giy077

King ZA, Lu JS, Dräger A, Miller PC, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE. BiGG Models: A platform for integrating, standardizing, and sharing genome-scale models. Nucleic Acids Res. 2016;44(D1). doi:10.1093/nar/gkv1049.

Kitano H. Biological robustness. Nat Rev Genet. nov 2004;5(11):826-37.

Kittikunapong C, Ye S, Magadán-Corpas P, Pérez-Valero Á, Villar CJ, Lombó F, *et al.* Reconstruction of a Genome-Scale Metabolic Model of *Streptomyces albus* J1074: Improved Engineering Strategies in Natural Product Synthesis. Metabolites. 11 mai 2021;11(5):304.

Klamt S. Generalized concept of minimal cut sets in biochemical networks. Biosystems. 1 févr 2006;83(2):233-47.

Klamt S, Saez-Rodriguez J, Gilles ED. Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst Biol. 2007;1:2. Published 2007 Jan 8. doi:10.1186/1752-0509-1-2

Klein DA, Paschke MW. Filamentous Fungi: the Indeterminate Lifestyle and Microbial Ecology. Microb Ecol. 1 avr 2004;47(3):224-35.

Klipp E, Liebermeister W, Wierling C, Kowald A. Systems Biology: A Textbook. John Wiley & Sons; 2016.

Klitgord N, Segrè D The Importance of Compartmentalization in Metabolic Flux Models: Yeast as an Ecosystem of Organelles. In: Genome Informatics 2009. Imperial College Press; 2010:41-55. doi:10.1142/9781848165786_0005.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. mai 2019;37(5):540-6.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. Genome Res. 5 janv 2017;27(5):722-36.

Kozakiewicz Z, Frisvad JC, Hawksworth DL, Pitt JI, Samson RA, Stolk AC. Proposal for Nomina Specifica Conservanda and Rejicienda in Aspergillus and *Penicillium* (Fungi). Taxon. 1992;41(1):109-13.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol. 19 janv 2001;305(3):567-80.

Kruse K, Ebenhöh O. Comparing flux balance analysis to network expansion: producibility, sustainability and the scope of compounds. In: Saito K, Dixon RA, Willmitzer L, eds. Genome Informatics 2008. Imperial College Press; 2008:91-101. doi:10.1142/9781848163003_0008.

Kumelj T, Sulheim S, Wentzel A, Almaas E. Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of *Streptomyces coelicolor*. Biotechnol J. avr 2019;14(4):e1800180.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, *et al.* Versatile and open software for comparing large genomes. Genome Biol. 30 janv 2004;5(2):R12.

L

Laat WD, Preusting J, Koekman B. Fermentative production of valuable compounds on an industrial scale using chemically defined *media*. 2002. Disponible sur: https://patents.google.com/patent/US20020039758A1/en

Lachance J, Matteau D, Brodeur J, Lloyd CJ, Mih N, King ZA, *et al.* Genome-scale metabolic modeling reveals key features of a minimal gene set. Mol Syst Biol. juill 2021;17(7):e10099.

Lachance JC, Lloyd CJ, Monk JM, Yang L, Sastry AV, Seif Y, *et al.* BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. PLOS Comput Biol. 22 avr 2019;15(4):e1006971.

Lacroix V, Cottret L, Thébault P, Sagot M. An Introduction to Metabolic Networks and Their Structural Analysis. IEEE/ACM Trans Comput Biol Bioinform. oct 2008;5(4):594-617.

Le VT, Bertrand S, Robiou du Pont T, Fleury F, Caroff N, Bourgeade-Delmas S, *et al.* Untargeted Metabolomics Approach for the Discovery of Environment-Related Pyran-2-Ones Chemodiversity in a Marine-Sourced *Penicillium restrictum*. Mar Drugs. juill 2021;19(7):378.

Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. Brief Bioinform. 1 juin 2006;7(2):140-50.

Lee JW, Na D, Park JM, Lee J, Choi S, Lee SY. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nat Chem Biol. juin 2012;8(6):536-46.

Lein J. The Panlabs penicillin strain improvement program. Overproduction Microb Metab. 1986;105-39.

Leonelli S, Ankeny RA. What makes a model organism? Endeavour. 1 déc 2013;37(4):209-12.

Lerma-Ortiz C, Jeffryes JG, Cooper AJL, Niehaus TD, Thamm AMK, Frelin O, *et al.* 'Nothing of chemistry disappears in biology': the Top 30 damage-prone endogenous metabolites. Biochem Soc Trans. 9 juin 2016;44(3):961-71.

Levy R, Borenstein E. Reverse ecology: from systems to environments and back. In: Soyer O, ed. Evolutionary Systems Biology. Vol 751. Advances in Experimental Medicine and Biology. Springer; 2012: 183-195. doi:10.1007/978-1-4614-3567-9_15.

Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. Nat Rev Microbiol. avr 2012;10(4):291-305.

Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, *et al.* Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol. déc 2010;28(12):1279-85.

Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC Bioinformatics. 16 juin 2009;10(1):185.

Li X, Yilmaz LS, Walhout AJM. Compartmentalization of metabolism between cell types in multicellular organisms: a computational perspective. Curr Opin Syst Biol. mars 2022;29:100407.

Li YW, Qian JY, Huang JC, Guo DS, Nie ZK, Ye C, *et al.* Improving Gibberellin GA3 Production with the Construction of a Genome-Scale Metabolic Model of *Fusarium fujikuroi*. J Agric Food Chem. 6 déc 2023;71(48):18890-7.

Licona-Cassani C, Marcellin E, Quek LE, Jacob S, Nielsen LK. Reconstruction of the *Saccharopolyspora erythraea* genome-scale model and its use for enhancing erythromycin production. Antonie Van Leeuwenhoek. oct 2012;102(3):493-502.

Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, *et al.* MEMOTE for standardized genome-scale metabolic model testing. Nat Biotechnol. mars 2020;38(3):272-6.

Ligon BL. Penicillin: its discovery and early development. Semin Pediatr Infect Dis. 1 janv 2004;15(1):52-7.

Lin X, Alspaugh JA, Liu H, Harris S. Fungal Morphogenesis. Cold Spring Harb Perspect Med. 2 janv 2015;5(2):a019679.

Liu D, Garrigues S, de Vries RP. Heterologous protein production in filamentous fungi. Appl Microbiol Biotechnol. 1 août 2023;107(16):5019-33.

Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM. Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. BMC Syst Biol. 18 sept 2014;8(1):110.

Llaneras F, Picó J. Stoichiometric modelling of cell metabolism. J Biosci Bioeng. 1 janv 2008;105(1):1-11.

Long MR, Ong WK, Reed JL. Computational methods in metabolic engineering for strain design. Curr Opin Biotechnol. 1 août 2015;34:135-41.

Lübeck M, Lübeck PS. Fungal Cell Factories for Efficient and Sustainable Production of Proteins and Peptides. Microorganisms. avr 2022;10(4):753.

M

Ma HW, Zhao XM, Yuan YJ, Zeng AP. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. Bioinformatics. 12 août 2004;20(12):1870-6.

Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 6 sept 2018;46(15):7542-53.

Machado D, Herrgård MJ. Co-evolution of strain design methods based on flux balance and elementary mode analysis. Metab Eng Commun. 1 déc 2015;2:85-92.

Machado D, Soons Z, Patil KR, Ferreira EC, Rocha I. Random sampling of elementary flux modes in large-scale metabolic networks. Bioinforma Oxf Engl. 15 sept 2012;28(18):i515-21.

Mackie A, Keseler IM, Nolan L, Karp PD, Paulsen IT. Dead End Metabolites - Defining the Known Unknowns of the *E. coli* Metabolic Network. Parkinson J, éditeur. PLoS ONE. 23 sept 2013;8(9):e75210.

Madhavan A, Arun K, Sindhu R, Alphonsa Jose A, Pugazhendhi A, Binod P, *et al.* Engineering interventions in industrial filamentous fungal cell factories for biomass valorization. Bioresour Technol. 1 janv 2022;344:126209.

Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol. janv 2017;35(1):81-9.

Maia P, Rocha M, Rocha I. *In silico* Constraint-Based Strain Optimization Methods: the Quest for Optimal Cell Factories. Microbiol Mol Biol Rev. 1 mars 2016;80(1):45-67.

Majeed A, Rauf I. Graph Theory: A Comprehensive Survey about Graph Theory Applications in Computer Science and Social Networks. Inventions. mars 2020;5(1):10.

Maldonado EM, Fisher CP, Mazzatti DJ, Barber AL, Tindall MJ, Plant NJ, *et al.* Multi-scale, whole-system models of liver metabolic adaptation to fat and sugar in non-alcoholic fatty liver disease. Npj Syst Biol Appl. 20 août 2018;4(1):1-10.

Maldonado EM, Leoncikas V, Fisher CP, Moore JB, Plant NJ, Kierzek AM. Integration of Genome Scale Metabolic Networks and Gene Regulation of Metabolic Enzymes With Physiologically Based Pharmacokinetics. CPT Pharmacomet Syst Pharmacol. 2017;6(11):732-46.

Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, *et al.* BioModels—15 years of sharing computational models in life science. Nucleic Acids Res. 8 janv 2020;48(D1):D407-15.

Mandal S, Krishnan R. Fungi: The budding source for biomaterials. Microb Biosyst. 1 juin 2021;6(1):55-65.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 1 oct 2021;38(10):4647-54.

Marinos G, Kaleta C, Waschina S. Defining the nutritional input for genome-scale metabolic models: A roadmap. PLOS ONE. 14 août 2020;15(8):e0236890.

Martín JF. Insight into the Genome of Diverse *Penicillium chrysogenum* Strains: Specific Genes, Cluster Duplications and DNA Fragment Translocations. Int J Mol Sci. janv 2020;21(11):3936.

Martín JF, Álvarez-Álvarez R, Liras P. Clavine Alkaloids Gene Clusters of *Penicillium* and Related Fungi: Evolutionary Combination of Prenyltransferases, Monooxygenases and Dioxygenases. Genes. déc 2017;8(12):342.

Martín JF, Liras P. Secondary Metabolites in Cheese Fungi. In: Mérillon, JM., Ramawat, K. (eds) Fungal Metabolites. Reference Series in Phytochemistry. Springer, Cham. 2017, https://doi.org/10.1007/978-3-319-25001-4_37

Martín JF, Ullán RV, García-Estrada C. Role of peroxisomes in the biosynthesis and secretion of β-lactams and other secondary metabolites. J Ind Microbiol Biotechnol. mars 2012;39(3):367-82.

Martyushenko N, Almaas E. ModelExplorer - software for visual inspection and inconsistency correction of genome-scale metabolic reconstructions. BMC Bioinformatics. 28 janv 2019;20(1):56.

Martyushenko N, Almaas E. ErrorTracer: an algorithm for identifying the origins of inconsistencies in genome-scale metabolic models. Bioinforma Oxf Engl. 1 mars 2020;36(5):1644-6.

Maskey RP, Grün-Wollny I, Laatsch H. Isolation, Structure Elucidation and Biological Activity of 8-O-Methylaverufin and 1, 8-O-Dimethylaverantin as New Antifungal Agents from *Penicillium chrysogenum*. J Antibiot (Tokyo). 25 mai 2003;56(5):459-63.

Matsuda Y, Wakimoto T, Mori T, Awakawa T, Abe I. Complete Biosynthetic Pathway of Anditomin: Nature's Sophisticated Synthetic Route to a Complex Fungal Meroterpenoid. J Am Chem Soc. 29 oct 2014;136(43):15326-36.

Matthews BJ, Vosshall LB. How to turn an organism into a model organism in 10 'easy' steps. Dickinson MH, Vosshall LB, Dow JAT, éditeurs. J Exp Biol. 1 févr 2020;223(Suppl_1):jeb218198.

McCloskey D, Palsson BØ, Feist AM. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol Syst Biol. 1 janv 2013;9(1):661.

Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nat Chem Biol. sept 2015;11(9):639-48.

Meijer WH, Gidijala L, Fekken S, Kiel JAKW, van den Berg MA, Lascaris R, *et al.* Peroxisomes are required for efficient penicillin biosynthesis in *Penicillium chrysogenum*. Appl Environ Microbiol. sept 2010;76(17):5702-9.

Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biol. 7 août 2019;20(1):158.

Meng J, Wang X, Xu D, Fu X, Zhang X, Lai D, *et al.* Sorbicillinoids from Fungi and Their Bioactivities. Molecules. 1 juin 2016;21(6):715.

Meng X, Fang Y, Ding M, Zhang Y, Jia K, Li Z, *et al.* Developing fungal heterologous expression platforms to explore and improve the production of natural products from fungal biodiversity. Biotechnol Adv. 1 janv 2022;54:107866.

Meyer V. Genetic engineering of filamentous fungi — Progress, obstacles and future trends. Biotechnol Adv. 1 mars 2008;26(2):177-85.

Meyer V, Basenko EY, Benz JP, Braus GH, Caddick MX, Csukai M, *et al.* Growing a circular economy with fungal biotechnology: a white paper. Fungal Biol Biotechnol. 2 avr 2020;7(1):5.

Mintz-Oron S, Aharoni A, Ruppin E, Shlomi T. Network-based prediction of metabolic enzymes' subcellular localization. Bioinformatics. 15 juin 2009;25(12):i247-1252.

Mintz-Oron S, Meir S, Malitsky S, Ruppin E, Aharoni A, Shlomi T. Reconstruction of *Arabidopsis* metabolic network models accounting for subcellular compartmentalization and tissue-specificity. Proc Natl Acad Sci. 3 janv 2012;109(1):339-44.

Mithani A, Hein J, Preston GM. Comparative Analysis of Metabolic Networks Provides Insight into the Evolution of Plant Pathogenic and Nonpathogenic Lifestyles in *Pseudomonas*. Mol Biol Evol. 1 janv 2011;28(1):483-99.

Moler C. Design of an interactive matrix calculator. In: Proceedings of the 1980 National Computer Conference; May 19–22, 1980; Anaheim, CA, USA. New York, NY: ACM; 1980:363–368. doi:10.1145/1500518.1500576.

Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, *et al.* Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci. 10 déc 2013;110(50):20338-43.

Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. Nat Biotechnol. mai 2014;32(5):447-52.

Moretti S, Tran VDT, Mehl F, Ibberson M, Pagni M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. Nucleic Acids Res. 8 janv 2021;49(D1):D570-4.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. juill 2007;35(Web Server issue):W182-185.

Moyer Andrew J., Coghill Robert D. Penicillin - The Laboratory Scale Production Of Penicillin In Submerged Cultures By *Penicillium Notatum* Westling (NRRL 832). J Bacteriol. 1 janv 1946;51(1):79-93.

Müller AC, Bockmayr A. Flux modules in metabolic networks. J Math Biol. 1 nov 2014;69(5):1151-79.

Müller B, Grossniklaus U. Model organisms — A historical perspective. J Proteomics. 10 oct 2010;73(11):2054-63.

N

Nadendla S, Jackson R, Munro J, Quaglia F, Mészáros B, Olley D, *et al.* ECO: the Evidence and Conclusion Ontology, an update for 2022. Nucleic Acids Res. 7 janv 2022;50(D1):D1515-21.

Nagy LG, Varga T, Csernetics Á, Virágh M. Fungi took a unique evolutionary route to multicellularity: Seven key challenges for fungal multicellular life. Fungal Biol Rev. 2020;34(4):151-169. doi:10.1016/j.fbr.2020.07.002.

Naithani S, Gupta P, Preece J, Garg P, Fraser V, Padgitt-Cobb LK, *et al.* Involving community in genes and pathway curation. Database J Biol Databases Curation. 1 janv 2019;2019.

Nanda P, Patra P, Das M, Ghosh A. Reconstruction and analysis of genome-scale metabolic model of weak Crabtree positive yeast *Lachancea kluyveri*. Sci Rep. 1 oct 2020;10(1):16314.

Naranjo-Ortiz MA, Gabaldón T. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. Biol Rev. 2019;94(6):2101-37.

Naranjo-Ortiz MA, Gabaldón T. Fungal evolution: cellular, genomic and metabolic complexity. Biol Rev. 17 avr 2020;brv.12605.

Nègre D, Aite M, Belcour A, Frioux C, Brillet-Guéguen L, Liu X, *et al.* Genome–Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae *Saccharina japonica* and *Cladosiphon okamuranus*. Antioxidants. nov 2019;8(11):564.

Nègre D, Larhlimi A, Bertrand S. Reconciliation and evolution of *Penicillium rubens* genome-scale metabolic networks–What about specialised metabolism? PLOS ONE. 30 août 2023;18(8):e0289757.

Newbert RW, Barton B, Greaves P, Harper J, Turner G. Analysis of a commercially improved *Penicillium chrysogenum* strain series: involvement of recombinogenic regions in amplification and deletion of the penicillin biosynthesis gene cluster. J Ind Microbiol Biotechnol. 1 juill 1997;19(1):18-27.

Newman DJ, Cragg GM. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. J Nat Prod. 23 mars 2012;75(3):311-35.

Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 27 mars 2020;83(3):770-803.

Nielsen H. Predicting secretory proteins with SignalP. In: Kihara D, ed. Protein Function Prediction. Vol 1611. Humana Press; 2017:59-73. doi:10.1007/978-1-4939-7015-5_6.

Nielsen J. Physiological Engineering Aspects of *Penicillium chrysogenum*. WORLD SCIENTIFIC; 1997. doi:10.1142/3195.

Nielsen J, Grijseels S, Prigent S, Ji B, Dainat J, Nielsen KF, *et al.* Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. Nat Microbiol. 2017;2(6):17044.

Nijland JG, Ebbendorf B, Woszczynska M, Boer R, Bovenberg RAL, Driessen AJM. Nonlinear Biosynthetic Gene Cluster Dose Effect on Penicillin Production by *Penicillium chrysogenum*. Appl Environ Microbiol. nov 2010;76(21):7109-15.

Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, *et al.* BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Res. 8 janv 2020;48(D1):D402-6.

Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B. Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics. 2006;7:296. Published 2006 Jun 13. doi:10.1186/1471-2105-7-296

Nouri H, Fouladiha H, Moghimi H, Marashi SA. A reconciliation of genome-scale metabolic network model of *Zymomonas mobilis* ZM4. Sci Rep. 8 mai 2020;10(1):7782.

Novère NL, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol. déc 2005;23(12):1509-15.

Nützmann HW, Scazzocchio C, Osbourn A. Metabolic Gene Clusters in Eukaryotes. Annu Rev Genet. 23 nov 2018;52(Volume 52, 2018):159-83.

O

Oarga A, Bannerman B, Júlvez J. Growth Dependent Computation of Chokepoints in Metabolic Networks. In: Abate A, Petrov T, Wolf V, éditeurs. Computational Methods in Systems Biology. Cham: Springer International Publishing; 2020. p. 102-19. (Lecture Notes in Computer Science).

Oberhardt MA, Puchałka J, Santos VAPM dos, Papin JA. Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis. PLOS Computational Biology. 31 mars 2011;7(3):e1001116.

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics. 7 oct 2011;3(1):33.

O'Brien EJ, Monk JM, Palsson BO. Using Genome-scale Models to Predict Biological Capabilities. Cell. 21 mai 2015;161(5):971-87.

de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK. AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in *Arabidopsis*. Plant Physiol. févr 2010;152(2):579-89.

Olivier B, Bergmann F. SBML Level 3 Package: Flux Balance Constraints version 2. Journal of Integrative Bioinformatics. 2018;15(1): 20170082. https://doi.org/10.1515/jib-2017-0082

Omokhagbor Adams G, Tawari Fufeyin P, Eruke Okoro S, Ehinomen I. Bioremediation, Biostimulation and Bioaugmention: A Review. Int J Environ Bioremediation Biodegrad. 29 oct 2020;3(1):28-39.

Orth JD, Palsson BØ. Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng. 2010;107(3):403-12.

Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. mars 2010;28(3):245-8.

Overy DP, Rämä T, Oosterhuis R, Walker AK, Pang KL. The Neglected Marine Fungi, Sensu stricto, and Their Isolation for Natural Products' Discovery. Mar Drugs. janv 2019;17(1):42.

P

Pan R, Bai X, Chen J, Zhang H, Wang H. Exploring Structural Diversity of Microbe Secondary Metabolites Using OSMAC Strategy: A Literature Review. Front Microbiol. 2019;10:294. Published 2019 Feb 26. doi:10.3389/fmicb.2019.00294

Pan S, Reed JL. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. Curr Opin Biotechnol. 1 juin 2018;51:103-8.

Papoutsakis ET. Equations and calculations for fermentations of butyric acid bacteria. Biotechnol Bioeng. 1984;67(6):813-26.

Park JM, Park HM, Kim WJ, Kim HU, Kim TY, Lee SY. Flux variability scanning based on enforced objective flux for identifying gene amplification targets. BMC Syst Biol. 21 août 2012;6:106.

Park SY, Binkley RM, Kim WJ, Lee MH, Lee SY. Metabolic engineering of *Escherichia coli* for high-level astaxanthin production with high productivity. Metab Eng. sept 2018;49:105-15.

Patel M, Kumar R, Kishor K, Mlsna T, Pittman CUJr, Mohan D. Pharmaceuticals of Emerging Concern in Aquatic Systems: Chemistry, Occurrence, Effects, and Removal Methods. Chem Rev. 27 mars 2019;119(6):3510-673.

Patil KR, Rocha I, Förster J, Nielsen J. Evolutionary programming as a platform for *in silico* metabolic engineering. BMC Bioinformatics. 23 déc 2005;6(1):308.

Peleg AY, Hogan DA, Mylonakis E. Medically important bacterial–fungal interactions. Nat Rev Microbiol. mai 2010;8(5):340-9.

Peng Q, Yuan Y, Gao M, Chen X, Liu B, Liu P, *et al.* Genomic characteristics and comparative genomics analysis of *Penicillium chrysogenum* KF-25. BMC Genomics. 2014;15(1):144.

Peng R. The reproducibility crisis in science: A statistical counterattack. Significance. 2015;12(3):30-2.

Pérez-Fernández BA, Fernandez-de-Cossio-Diaz J, Boggiano T, León K, Mulet R. In-silico *media* optimization for continuous cultures using genome scale metabolic networks: The case of CHO-K1. Biotechnol Bioeng. 2021;118(5):1884-97.

Perumal D, Lim CS, Sakharkar MK. A Comparative Study of Metabolic Network Topology between a Pathogenic and a Non-Pathogenic Bacterium for Potential Drug Target Identification. Summit Transl Bioinforma. 1 mars 2009;2009:100-4.

Peters I, Kraker P, Lex E, Gumpenberger C, Gorraiz JI. Zenodo in the Spotlight of Traditional and New Metrics. Front. Res. Metr. Anal. 2017 2:13. doi: 10.3389/frma.2017.00013

Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. oct 2011;8(10):785-6.

Pettit RK. Small-molecule elicitation of microbial secondary metabolites. Microb Biotechnol. juill 2011;4(4):471-8.

Pham HT, Lee KH, Jeong E, Woo S, Yu J, Kim WY, *et al.* Species Prioritization Based on Spectral Dissimilarity: A Case Study of Polyporoid Fungal Species. J Nat Prod. 26 févr 2021;84(2):298-309.

Pham N, van Heck RGA, van Dam JCJ, Schaap PJ, Saccenti E, Suarez-Diez M. Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling. Metabolites. févr 2019;9(2):28.

Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. Genome Res. nov 2004;14(11):2367-76.

Pharkya P, Maranas CD. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng. 1 janv 2006;8(1):1-13.

Pienaar E, Matern WM, Linderman JJ, Bader JS, Kirschner DE. Multiscale Model of *Mycobacterium tuberculosis* Infection Maps Metabolite and Gene Perturbations to Granuloma Sterilization Predictions. Ehrt S, éditeur. Infect Immun. mai 2016;84(5):1650-69.

Pienta R, Abello J, Kahng M, Chau DH. Scalable graph exploration and visualization: Sensemaking challenges and opportunities. In 2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015. Institute of Electrical and Electronics Engineers Inc. 2015. p. 271-278. 7072812. (2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015). doi: 10.1109/35021BIGCOMP.2015.7072812

Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, *et al.* Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites. avr 2019;9(4):76.

Pitkänen E, Arvas M, Rousu J. Reconstructing Gapless Ancestral Metabolic Networks. In: Fred A, Filipe J, Gamboa H, éditeurs. Biomedical Engineering Systems and Technologies. Berlin, Heidelberg: Springer; 2013. p. 126-40. (Communications in Computer and Information Science).

Pitkänen E, Jouhten P, Hou J, *et al.* Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. PLoS Comput Biol. 2014;10(2):e1003465. Published 2014 Feb 6. doi:10.1371/journal.pcbi.1003465

Pitkänen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol. 29 oct 2009;3(1):103.

Pitt JI. Review of The Genus *Penicillium* and Its Teleomorphic States Eupenicillium and Talaromyces, by JI Pitt. Mycologia. 1981;73(3):582-584. doi:10.2307/3759616.

Pohl C, Polli F, Schütze T, Viggiano A, Mózsik L, Jung S, *et al.* A *Penicillium rubens* platform strain for secondary metabolite production. Sci Rep. 6 mai 2020;10(1):7630.

Ponce-de-León M, Montero F, Peretó J. Solving gap metabolites and blocked reactions in genome-scale models: application to the metabolic network of *Blattabacterium cuenoti*. BMC Syst Biol. 31 oct 2013;7(1):114.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2 juill 2018;46(W1):W200-4.

Powers-Fletcher MV, Kendall BA, Griffin AT, Hanson KE. Filamentous fungi. In: Advances in Fungal Biotechnology for Industry, Agriculture, and Medicine. 2016:311-341. doi:10.1128/9781555819040.ch14.

Praußе MTE, Schäuble S, Guthke R, Schuster S. Computing the various pathways of penicillin synthesis and their molar yields. Biotechnol Bioeng. 2016;113(1):173-81.

Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health. 3 oct 2015;109(7):309-18.

Price ND, Famili I, Beard DA, Palsson BØ. Extreme Pathways and Kirchhoff's Second Law. Biophys J. 1 nov 2002;83(5):2879-82.

Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol. nov 2004;2(11):886-97.

Priebe S, Kreisel C, Horn F, Guthke R, Linde J. FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species. Bioinformatics. 1 févr 2015;31(3):445-6.

Prigent S, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, *et al.* Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. Kaleta C, éditeur. PLOS Comput Biol. 27 janv 2017;13(1):e1005276.

Prigent S, Nielsen JC, Frisvad JC, Nielsen J. Reconstruction of 24 *Penicillium* genome-scale metabolic models shows diversity based on their secondary metabolism. Biotechnol Bioeng. 2018;115(10):2604-12.

Q

Quiros-Guerrero LM, Nothias LF, Gaudry A, *et al.* Inventa: A computational tool to discover structural novelty in natural extracts libraries. Front Mol Biosci. 2022;9:1028334. Published 2022 Nov 11. doi:10.3389/fmolb.2022.1028334

R

Raajaraam L, Raman K. A Computational Framework to Identify Metabolic Engineering Strategies for the Co-Production of Metabolites. Front Bioeng Biotechnol. 2021;9:779405.

Rahman SA, Schomburg D. Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. Bioinformatics. 15 juill 2006;22(14):1767-74.

Rangel A, Gómez Ramírez JM, González Barrios AF. From industrial by-products to value-added compounds: the design of efficient microbial cell factories by coupling systems metabolic engineering and bioprocesses. Biofuels Bioprod Biorefining. 2020;14(6):1228-38.

Ramon C, Gollub MG, Stelling J. Integrating omics data into genome-scale metabolic network models: principles and challenges. Essays Biochem. 12 oct 2018;62(4):563-74.

Ranganathan S, Suthers PF, Maranas CD. OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions. PLOS Comput Biol. 15 avr 2010;6(4):e1000744.

Rangel AET, Gómez Ramírez JM, González Barrios AF. From industrial by-products to value-added compounds: the design of efficient microbial cell factories by coupling systems metabolic engineering and bioprocesses. Biofuels Bioprod Biorefining. 2020;14(6):1228-38.

Raper KB, Alexander DF, Coghill RD. Penicillin: II. Natural Variation and Penicillin Production in *Penicillium* notatum and Allied Species. J Bacteriol. déc 1944;48(6):639-59.

Raven PH, Johnson GB, Mason KA, Losos JB, Singer SS. Biologie. 2e Revue et augmentée. De Boeck Sup, éditeur. Bruxelles: De Boeck Supérieur; 2011.

Ravikrishnan A, Raman K. Critical assessment of genome-scale metabolic networks: the need for a unified standard. Brief Bioinform. 1 nov 2015;16(6):1057-68.

Rawls K, Dougherty BV, Papin J. Metabolic Network Reconstructions to Predict Drug Targets and Off-Target Effects. Methods Mol Biol Clifton NJ. 2020;2088:315-30.

Razmilic V, Castro JF, Andrews B, Asenjo JA. Analysis of metabolic networks of *Streptomyces leeuwenhoekii* C34 by means of a genome scale model: Prediction of modifications that enhance the production of specialized metabolites. Biotechnol Bioeng. juill 2018;115(7):1815-28.

Reddy SM, Reddy RamS, Narendra Babu G. Basic Industrial Biotechnology. 1st Edition. New Delhi: New Age International Publishers; 2012. (Fermentation process).

Reimers AC. Hierarchical decomposition of metabolic networks using k-modules. Biochem Soc Trans. 27 nov 2015;43(6):1146-50.

Rezola A, Pey J, Tobalina L, Rubio Á, Beasley JE, Planes FJ. Advances in network-based metabolic pathway analysis and gene expression data integration. Brief Bioinform. 1 mars 2015;16(2):265-79.

Richard D, Chevalet P, Soubaya T. Mémo visuel de biologie: L'essentiel en fiches. France: Dunod; 2011.

Richards TA, Jones MDM, Leonard G, Bass D. Marine Fungi: Their Ecology and Molecular Diversity. Annu Rev Mar Sci. 2012;4(1):495-522.

Robin J, Jakobsen M, Beyer M, Noorman H, Nielsen J. Physiological characterisation of *Penicillium chrysogenum* strains expressing the expandase gene from *Streptomyces clavuligerus* during batch cultivations. Growth and adipoyl-7-aminodeacetoxycephalosporanic acid production. Appl Microbiol Biotechnol. 1 oct 2001;57(3):357-62.

Robinson JL, Kocabaş P, Wang H, Cholley PE, Cook D, Nilsson A, *et al.* An atlas of human metabolism. Sci Signal. 24 mars 2020;13(624):eaaz1482.

Robinson SL, Panaccione DG. Diversification of Ergot Alkaloids in Natural and Modified Fungi. Toxins. janv 2015;7(1):201-18.

Roche CM, Loros JJ, McCluskey K, Glass NL. *Neurospora crassa*: Looking back and looking forward at a model microbe. Am J Bot. 2014;101(12):2022-35.

Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. Nat Chem. juin 2016;8(6):531-41.

Rodríguez-Sáiz M, Barredo JL, Moreno MA, Fernández-Cañón JM, Peñalva MA, Díez B. Reduced Function of a Phenylacetate-Oxidizing Cytochrome P450 Caused Strong Genetic Improvement in Early Phylogeny of Penicillin-Producing Strains. J Bacteriol. oct 2001;183(19):5465-71.

Rokas A, Mead ME, Steenwyk JL, Raja HA, Oberlies NH. Biosynthetic gene clusters and the evolution of fungal chemodiversity. Nat Prod Rep. 2020;37(7):10.1039.C9NP00045C.

Romano S, Jackson SA, Patry S, Dobson ADW. Extending the « One Strain Many Compounds » (OSMAC) Principle to Marine Microorganisms. Mar Drugs. 23 juill 2018;16(7).

Romsdahl J, Wang CCC. Recent advances in the genome mining of *Aspergillus* secondary metabolites (covering 2012–2018). MedChemComm. 19 juin 2019;10(6):840-66.

Roth MG, Westrick NM, Baldwin TT. Fungal biotechnology: From yesterday to tomorrow. Front Fungal Biol. 2023;4:1135263. Published 2023 Mar 27. doi:10.3389/ffunb.2023.1135263

Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, *et al.* The LOTUS initiative for open knowledge management in natural products research. Donoso DA, Akhmanova A, Tapley Hoyt C, éditeurs. eLife. 26 mai 2022;11:e70780.

S

Saa PA, Nielsen LK. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. Biotechnol Adv. 1 déc 2017;35(8):981-1003.

Saa PA, Zapararte S, Drovandi CC, Nielsen LK. LooplessFluxSampler: an efficient toolbox for sampling the loopless flux solution space of metabolic models. BMC Bioinformatics. 2 janv 2024;25(1):3.

Saier MH Jr, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, *et al.* The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 8 janv 2021;49(D1):D461-7.

Salo O, Guzmán-Chávez F, Ries MI, Lankhorst PP, Bovenberg RAL, Vreeken RJ, *et al.* Identification of a Polyketide Synthase Involved in Sorbicillin Biosynthesis by *Penicillium chrysogenum*. Appl Environ Microbiol. juill 2016;82(13):3971-8.

Salo OV, Ries M, Medema MH, Lankhorst PP, Vreeken RJ, Bovenberg RAL, *et al.* Genomic mutational analysis of the impact of the classical strain improvement program on β–lactam producing *Penicillium chrysogenum*. BMC Genomics. 14 nov 2015;16(1):937.

Sambamoorthy G, Raman K. Understanding the evolution of functional redundancy in metabolic networks. Bioinformatics. 1 sept 2018;34(17):i981-7.

Samol M. Genomic wake-up call: activating silent biosynthetic pathways for novel metabolites in *Penicillium chrysogenum*. [Thesis]. Groningen: University of Groningen; 2015.

Samol MM, Maire FB, Ries MI, Wolsink H, Salo O, Ali H, *et al.* Biosynthetic Pathways of Iron-Chelating Siderophores in Industrial Strains of. 2015;37.

Samson RA, Hadlok R, Stolk AC. A taxonomic study of the *Penicillium chrysogenum* series. Antonie Van Leeuwenhoek. 1977;43(2):169-75.

Sandberg TE, Salazar MJ, Weng LL, Palsson BO, Feist AM. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. Metab Eng. 1 déc 2019;56:1-16.

Savinell JM, Palsson BO. Optimal selection of metabolic fluxes for *in vivo* measurement. I. Development of mathematical methods. J Theor Biol. 21 mars 1992a;155(2):201-14.

Savinell JM, Palsson BO. Optimal selection of metabolic fluxes for *in vivo* measurement. II. Application to *Escherichia coli* and hybridoma cell metabolism. J Theor Biol. 21 mars 1992b;155(2):215-42.

Scheffler RJ, Colmer S, Tynan H, Demain AL, Gullo VP. Antimicrobials, drug discovery, and genome mining. Appl Microbiol Biotechnol. 1 févr 2013;97(3):969-78.

Schellenberger J, Lewis NE, Palsson BØ. Elimination of Thermodynamically Infeasible Loops in Steady-State Metabolic Models. Biophys J. 2 févr 2011;100(3):544-53.

Schueffler A, Anke T. Fungal natural products in research and development. Nat Prod Rep. 11 sept 2014;31(10):1425-48.

Schulz C, Almaas E. Genome-scale reconstructions to assess metabolic phylogeny and organism clustering. PLOS ONE. 29 déc 2020;15(12):e0240953.

Schulz C, Kumelj T, Karlsen E, Almaas E. Genome-scale metabolic modelling when changes in environmental conditions affect biomass composition. PLOS Comput Biol. 24 mai 2021;17(5):e1008528.

Schuster S, Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. J Biol Syst. juin 1994;02(02):165-82.

Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, *et al.* The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res. 8 janv 2021;49(D1):D575-88.

Seif Y, Kavvas E, Lachance JC, Yurkovich JT, Nuccio SP, Fang X, *et al.* Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. Nat Commun. 14 sept 2018;9(1):3771.

Seif Y, Choudhary KS, Hefner Y, Anand A, Yang L, Palsson BO. Metabolic and genetic basis for auxotrophies in Gram-negative species. Proc Natl Acad Sci. 17 mars 2020;117(11):6264-73.

Shahid MG, Ali S, Nadeem M. Biosynthesis of Ergot Alkaloids by *Penicillium citrinum* through Surface Culture Fermentation Process. Pak J Zool. 2015;47(2).

Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E. Metabolic Network Prediction of Drug Side Effects. Cell Syst. 23 mars 2016;2(3):209-13.

Shepelin D, Hansen ASL, Lennen R, Luo H, Herrgård MJ. Selecting the Best: Evolutionary Engineering of Chemical Production in Microbes. Genes. 11 mai 2018;9(5):249.

Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, *et al.* KNApSAcK: a comprehensive species-metabolite relationship database. In: Saito K, Dixon RA, Willmitzer L, eds. Plant Metabolomics. Berlin, Heidelberg: Springer; 2006:165-181. doi:10.1007/3-540-29782-0_13.

Sierra LAB, Mendes-Pereira T, García GJY, Werkhaizer CQ, Rezende JB de, Rodrigues TAB, *et al.* Current situation and future perspectives for the use of fungi in the biomaterial industry and proposal for a new classification of fungal-derived materials. PeerJ Mater Sci. 29 août 2023;5:e31.

Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BT. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC Syst Biol. 2010;4.

Singh A, Bajar S, Devi A, Pant D. An overview on the recent developments in fungal cellulase production and their industrial applications. Bioresour Technol Rep. 1 juin 2021;14:100652.

Singh S, Joshi P, An B, Chopade A. Choke point analysis of the metabolic pathways of *Acinetobacter baylyi*: A genomics approach to assess potential drug targets. J Bioinforma Seq Anal. 31 oct 2009;1(3):041-5.

Singh S, Malik BK, Sharma DK. Choke point analysis of metabolic pathways in *E. histolytica*: A computational approach for drug target identification. Bioinformation. 15 oct 2007;2(2):68-72.

Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 27 nov 2020;11(1):6058.

Sohn YJ, Kim HT, Jo SY, Song HM, Baritugo KA, Pyo J, *et al.* Recent Advances in Systems Metabolic Engineering Strategies for the Production of Biopolymers. Biotechnol Bioprocess Eng. 1 déc 2020;25(6):848-61.

Specht T, Dahlmann TA, Zadra I, Kurnsteiner H, Kuck U. Complete Sequencing and Chromosome-Scale Genome Assembly of the Industrial Progenitor Strain P2niaD18 from the Penicillin Producer *Penicillium chrysogenum*. Genome Announc. 24 juill 2014;2(4):e00577-14, 2/4/e00577-14.

Sperandio GB, Ferreira Filho EX. Fungal co-cultures in the lignocellulosic biorefinery context: A review. Int Biodeterior Biodegrad. 1 août 2019;142:109-23.

Srinivasan A, S V, Raman K, Srivastava S. Rational metabolic engineering for enhanced alpha-tocopherol production in *Helianthus annuus* cell culture. Biochem Eng J. 15 nov 2019;151:107256.

Stauffer JF, Backus MY. Spontaneous and Induced Variation in Selected Stocks of the *Penicillium chrysogenum* Series. Ann N Y Acad Sci. 1954;60(1):35-49.

Stephanopoulos G, Sinskey AJ. Metabolic engineering--methodologies and future prospects. Trends Biotechnol. sept 1993;11(9):392-6.

Stone S, Newman DJ, Colletti SL, Tan DS. Cheminformatic analysis of natural product-based drugs and chemical probes. Nat Prod Rep. 26 janv 2022;39(1):20-32.

Stratton CF, Newman DJ, Tan DS. Cheminformatic comparison of approved drugs from natural product versus synthetic origins. Bioorg Med Chem Lett. 1 nov 2015;25(21):4802-7.

Strong PJ, Self R, Allikian K, Szewczyk E, Speight R, O'Hara I, *et al.* Filamentous fungi for future functional food and feed. Curr Opin Biotechnol. 1 août 2022;76:102729.

Su G., Morris, J. H., Demchak, B., & Bader, G. D.. Biological network exploration with Cytoscape 3. Current protocols in bioinformatics, 2014, 47, 8.13.1–8.13.24. https://doi.org/10.1002/0471250953.bi0813s47

Sun Y., Zhang, T., Lu, B., Li, X., & Jiang, L. Application of cofactors in the regulation of microbial metabolism: A state of the art review. Frontiers in microbiology, 2023, 14, 1145784. https://doi.org/10.3389/fmicb.2023.1145784

Sulheim S, Fossheim FA, Wentzel A, Almaas E. Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters. BMC Bioinformatics. 23 févr 2021;22(1):81.

Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, *et al.* Recon 2.2: from reconstruction to model of human metabolism. Metabolomics. 7 juin 2016;12:109.

T

Tejera N., Crossman, L., Pearson, B., Stoakes, E., Nasher, F., Djeghout, B., Poolman, M., Wain, J., & Singh, D.. Genome-Scale Metabolic Model Driven Design of a Defined *Medium* for *Campylobacter jejuni* M1cam. Frontiers in microbiology, 2020, 11, 1072. https://doi.org/10.3389/fmicb.2020.01072

Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 6 janv 2023;51(D1):D603-10.

Terzer M, Maynard ND, Covert MW, Stelling J. Genome-scale metabolic networks. WIREs Syst Biol Med. 2009;1(3):285-97.

The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 8 janv 2021;49(D1):D325-34.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1)–D531. doi:10.1093/nar/gkac1052.

Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. janv 2010a;5(1):93-121.

Thiele I, Palsson BØ. Reconstruction annotation jamborees: a community approach to systems biology. Mol Syst Biol. 13 avr 2010b;6:361.

Thom C. Mycology Presents Penicillin. Mycologia. 1945;37(4):460-75.

Tiwari K, Kananathan S, Roberts MG, Meyer JP, Sharif Shohan MU, Xavier A, *et al.* Reproducibility in systems biology modelling. Mol Syst Biol. févr 2021;17(2):e9982.

Toya Y, Shiraki T, Shimizu H. SSDesign: Computational metabolic pathway design based on flux variability using elementary flux modes. Biotechnol Bioeng. avr 2015;112(4):759-68.

Trinh CT, Wlaschin A, Srienc F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Appl Microbiol Biotechnol. 1 janv 2009;81(5):813-26.

Tsang CC, Tang JYM, Lau SKP, Woo PCY. Taxonomy and evolution of Aspergillus, *Penicillium* and Talaromyces in the omics era – Past, present and future. Comput Struct Biotechnol J. 1 janv 2018;16:197-210.

Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, *et al.* International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017 [Internet]. Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, *et al.*, éditeurs. Glashütten: Koeltz Botanical Books; 2018. (Regnum Vegetabile 159). Disponible sur: https://doi.org/10.12705/Code.2018

V

Valderrama-Gomez MA, Kreitmayer D, Wolf S, Marin-Sanguino A, Kremling A. Application of theoretical methods to increase succinate production in engineered strains. Bioprocess Biosyst Eng. 1 avr 2017;40(4):479-97.

Van den Berg MA, Westerlaken I, Leeflang C, Kerkman R, Bovenberg RAL. Functional characterization of the penicillin biosynthetic gene cluster of *Penicillium chrysogenum* Wisconsin54-1255. Fungal Genet Biol. 1 sept 2007;44(9):830-44.

Van den Berg MA. Impact of the *Penicillium chrysogenum* genome on industrial production of metabolites. Appl Microbiol Biotechnol. 1 oct 2011;92(1):45-53.

Van den Berg MA, Albang R, Albermann K, Badger JH, Daran JM, Driessen AJM, *et al.* Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat Biotechnol. oct 2008;26(10):1161-8.

Van der Hoek SA, Rusnák M, Wang G, Stanchev LD, de Fátima Alves L, Jessop-Fabre MM, *et al.* Engineering precursor supply for the high-level production of ergothioneine in *Saccharomyces cerevisiae*. Metab Eng. 1 mars 2022;70:129-42.

Van der Lee TAJ, Medema MH. Computational strategies for genome-based natural product discovery and engineering in fungi. Fungal Genet Biol. 1 avr 2016;89:29-36.

Van der Oost, Patinios C. The genome editing revolution. Trends Biotechnol. 1 mars 2023;41(3):396-409.

Van Helden J, Wernisch L, Gilbert D, Wodak SJ. Graph-Based Analysis of Metabolic Networks. In: Mewes HW, Seidel H, Weiss B, éditeurs. Bioinformatics and Genome Analysis. Berlin, Heidelberg: Springer; 2002. p. 245-74. (Ernst Schering Research Foundation Workshop).

Van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 27 nov 2019;5(11):1824-33.

Van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, *et al.* The Natural Products Atlas 2.0: a database of microbially-derived natural products. Nucleic Acids Res. 7 janv 2022;50(D1):D1317-23.

Van Steijn L, Verbeek FJ, Spaink HP, Merks RMH. Predicting Metabolism from Gene Expression in an Improved Whole-Genome Metabolic Network Model of *Danio rerio*. Zebrafish. août 2019;16(4):348-62.

Viggiano A., Salo, O., Ali, H., Szymanski, W., Lankhorst, P. P., Nygård, Y., *et al.* Pathway for the Biosynthesis of the Pigment Chrysogine by *Penicillium chrysogenum*. Applied and environmental microbiology, 2018, 84(4), e02246-17. https://doi.org/10.1128/AEM.02246-17

Vijayakumar S, Conway M, Lió P, Angione C. Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. Brief Bioinform. 27 nov 2018;19(6):1218-35.

Visagie CM, Houbraken J, Frisvad JC, Hong SB, Klaassen CHW, Perrone G, *et al.* Identification and nomenclature of the genus *Penicillium*. Stud Mycol. 1 juin 2014;78(1):343-71.

Volk MJ, Tran VG, Tan SI, Mishra S, Fatma Z, Boob A, *et al.* Metabolic Engineering: Methodologies and Applications. Chem Rev. 10 mai 2023;123(9):5521-70.

W

Wagner A, Fell DA. The small world inside large metabolic networks. Proc R Soc Lond B Biol Sci. 7 sept 2001;268(1478):1803-10.

Wagner GP, Pavlicev M, Cheverud JM. The road to modularity. Nat Rev Genet. déc 2007;8(12):921-31.

Waller T.C., Berg, J. A., Lex, A., Chapman, B. E., & Rutter, J. Compartment and hub definitions tune metabolic networks for metabolomic interpretations. GigaScience, 2020, 9(1), giz137. https://doi.org/10.1093/gigascience/giz137

Walsh C, Tang Y. Natural product biosynthesis : chemical logic and enzymatic machinery. London: Royal Society of Chemistry; 2017.

Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R, *et al.* Minimum Information About a Simulation Experiment (MIASE). PLOS Comput Biol. 28 avr 2011;7(4):e1001122.

Wang F, Ren Z, Xu P, Zhao Y, Xiu J, Zhu Y, *et al.* Construction of vector pPIPKA and transformation of *Penicillium chrysogenum* industrial strai. Mycosystema. 1 janv 2004;23(1):66-72.

Wang FQ, Zhong J, Zhao Y, Xiao J, Liu J, Dai M, *et al.* Genome sequencing of high-penicillin producing industrial strain of *Penicillium chrysogenum*. BMC Genomics. 24 janv 2014;15(1):S11.

Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, *et al.* RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLOS Comput Biol. 18 oct 2018;14(10):e1006541.

Wang H, Robinson JL, Kocabas P, Gustafsson J, Anton M, Cholley PE, *et al.* Genome-scale metabolic network reconstruction of model animals as a platform for translational research. Proc Natl Acad Sci. 27 juill 2021;118(30):e2102344118.

Wang L, Dash S, Ng CY, Maranas CD. A review of computational tools for design and reconstruction of metabolic pathways. Synth Syst Biotechnol. 1 déc 2017;2(4):243-52.

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. août 2016;34(8):828-37.

Wang M, Wang H, Zheng H. A Mini Review of Node Centrality Metrics in Biological Networks. Int J Netw Dyn Intell. 22 déc 2022;99-110.

Wang Q, Arighi CN, King BL, Polson SW, Vincent J, Chen C, *et al.* Community annotation and bioinformatics workforce development in concert--Little Skate Genome Annotation Workshops and Jamborees. Database J Biol Databases Curation. 2012;2012:bar064.

Ward OP. Production of recombinant proteins by filamentous fungi. Biotechnol Adv. 1 sept 2012;30(5):1119-39.

Watson MR. A discrete model of bacterial metabolism. Bioinformatics. 1 avr 1986;2(1):23-7.

Weber T. *In silico* tools for the analysis of antibiotic biosynthetic pathways. Int J Med Microbiol. 1 mai 2014;304(3):230-5.

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31-36.

Wendisch VF, Brito LF, Lopez MG, Hennig G, *et al.*. The flexible feedstock concept in Industrial Biotechnology: Metabolic engineering of *Escherichia coli*, *Corynebacterium glutamicum*, *Pseudomonas*, *Bacillus*, and yeast strains for access to alternative carbon sources. J Biotechnol. 2016;234:139-157. doi:10.1016/j.jbiotec.2016.07.022.

Whittaker RH. New Concepts of Kingdoms of Organisms. Science. 10 janv 1969;163(3863):150-60.

Wickham H. Data Analysis. In:. ggplot2: Elegant Graphics for Data Analysis Use R!. Cham: Springer International Publishing; 2016. p. 189-201. Disponible sur: https://doi.org/10.1007/978-3-319-24277-4_9

Wiemann P, Keller NP. Strategies for mining fungal natural products. J Ind Microbiol Biotechnol. 1 févr 2014;41(2):301-13.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 15 mars 2016;3(1):160018.

***Wishart DS**, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1). doi:10.1093/nar/gkx1037.*

Wishart DS, Guo AC, Oler E, *et al.* HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. 2022;50(D1)–31. doi:10.1093/nar/gkab1062. PubMed ID: 34986597.

Wittig U, Rey M, Weidemann A, Kania R, Müller W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res. 2018;46(D1)–D660. doi:10.1093/nar/gkx1065.

Witting M, Hastings, J., Rodriguez, N., Joshi, C. J., Hattwell, J. P. N., Ebert, P. R., *et al.* Modeling Meets Metabolomics-The WormJam Consensus Model as Basis for Metabolic Studies in the Model Organism *Caenorhabditis elegans*. Frontiers in molecular biosciences, 2018, 5, 96. https://doi.org/10.3389/fmolb.2018.00096

Wolfender JL, Nuzillard JM, van der Hooft JJJ, Renault JH, Bertrand S. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography–High-Resolution Tandem Mass Spectrometry and NMR Profiling, *in silico* Databases, and Chemometrics. Anal Chem. 2 janv 2019;91(1):704-42.

Woodcraft C, Chooi YH, Roux I. The expanding CRISPR toolbox for natural product discovery and engineering in filamentous fungi. Nat Prod Rep. 25 janv 2023;40(1):158-73.

World Health Organization. World health statistics 2023: monitoring health for the SDGs, sustainable development goals. 1, éditeur. Geneva: World Health Organization; 2023. Disponible sur: https://www.who.int/publications/i/item/9789240074323

Wu H, von Kamp A, Leoncikas V, Mori W, Sahin N, Gevorgyan A, *et al.* MUFINS: multi-formalism interaction network simulator. NPJ Syst Biol Appl. 2016;2:16032.

X

Xavier JC, Patil KR, Rocha I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. Metab Eng. 1 janv 2017;39:200-8.

Xu N, Ye C, Chen X, Liu J, Liu L. Genome-scale metabolic modelling common cofactors metabolism in microorganisms. J Biotechnol. 10 juin 2017;251:1-13.

Xu P, Ranganathan S, Fowler ZL, Maranas CD, Koffas MAG. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. Metab Eng. 1 sept 2011;13(5):578-87.

Y

Yabe K, Nakajima H. Enzyme reactions and genes in aflatoxin biosynthesis. Appl Microbiol Biotechnol. 1 juin 2004;64(6):745-55.

Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, *et al.* Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 7 janv 2022;50(D1):D996-1003.

Yee DA, Niwa K, Perlatti B, Chen M, Li Y, Tang Y. Genome mining for unknown-unknown natural products. Nat Chem Biol. mai 2023;19(5):633-40.

Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB. Computational Analysis of Plasmodium falciparum Metabolism: Organizing Genomic Information to Facilitate Drug Discovery. Genome Res. 5 janv 2004;14(5):917-24.

Yilmaz LS, Walhout AJM. A *Caenorhabditis elegans* Genome-Scale Metabolic Network Model. Cell Syst. 25 mai 2016;2(5):297-311.

Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics. 15 juin 2010;26(12):i255-60.

Yurekten O, Payne T, Tejera N, *et al.* MetaboLights: open data repository for metabolomics. Nucleic Acids Res. 2023;51(D1)–D613. doi:10.1093/nar/gkad1045.

Z

Zakrzewski P, Medema MH, Gevorgyan A, Kierzek AM, Breitling R, Takano E. MultiMetEval: Comparative and Multi-Objective Analysis of Genome-Scale Metabolic Models. PLOS ONE. 14 déc 2012;7(12):e51511.

Zarins-Tutt JS, Barberi TT, Gao H, Mearns-Spragg A, Zhang L, Newman DJ, *et al.* Prospecting for new bacterial metabolites: a glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. Nat Prod Rep. 23 déc 2015;33(1):54-72.

Zhang C, Sánchez BJ, Li F, Eiden CWQ, Scott WT, Liebal UW, *et al.* Yeast9: a consensus genome-scale metabolic model for *S. cerevisiae* curated by the community. Mol Syst Biol. 12 août 2024;1-17.

Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, *et al.* MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research. Plant Physiol. 1 mai 2005;138(1):27-37.

Zhu F, Qin C, Tao L, Liu X, Shi Z, Ma X, *et al.* Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. Proc Natl Acad Sci. 2 août 2011;108(31):12943-8.

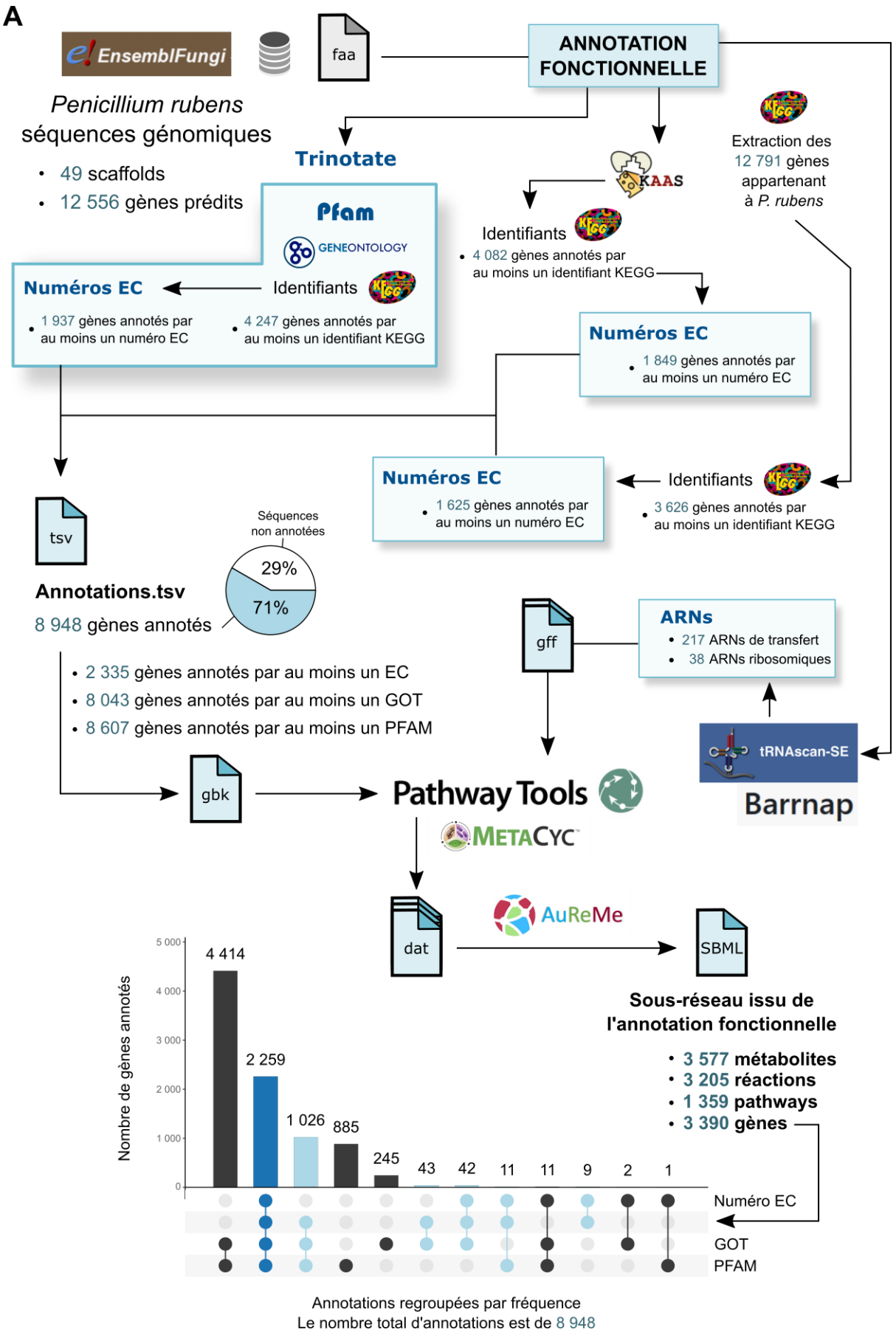
Zhu Z, Zhang J, Ji X, Fang Z, Wu Z, Chen J, *et al.* Evolutionary engineering of industrial microorganisms-strategies and applications. Appl Microbiol Biotechnol. juin 2018;102(11):4615-27.

Ziemons S, Koutsantas K, Becker K, Dahlmann T, Kück U. Penicillin production in industrial strain *Penicillium chrysogenum* P2niaD18 is not dependent on the copy number of biosynthesis genes. BMC Biotechnol. 16 févr 2017;17(1):16.

Annexes

ANNEXE 1 :

Reconstruction d'*i*Prub22 – sous-réseau issu de l'annotation fonctionnelle



Annexe 1A : *Résumé des étapes pour la reconstruction du sous-réseau issu de l'annotation fonctionnelle du génome.*

Pour reconstruire le sous-réseau issu de l'annotation fonctionnelle du génome deux types d'annotations sont essentielles : les numéros EC (*Enzyme Commission Numbers*) et les termes issus de la Gene Ontology (GOT). Les premiers fournissent une information sur la nature des enzymes et le type de réactions qu'elles catalysent, alors que les seconds regroupent, sous forme de graphe orienté, les informations géniques des systèmes biologiques. Outre ces annotations, nous avons également recherché à caractériser les domaines protéiques en annotant nos séquences avec des identifiants PFAM. De plus, nous avons enrichi l'annotation en incorporant des données relatives aux ARNs ribosomiques et de transfert.

Afin de collecter les annotations en numéros EC, nous avons initialement exploité les annotations existantes pour *P. rubens*, récupérant ainsi les données déposées sous l'identifiant T01091 dans KEGG GENOME. Cette source répertorie 12 791 gènes protéiques, que nous désignons comme gènes pcs (*i.e.* du nom du code de l'organisme dans KEGG GENOME). Ces données datant de 2009, nous avons décidé de réaliser une nouvelle annotation fonctionnelle afin de les mettre à jour. Pour ce faire, nous avons récupéré, entre autres, les identifiants KO qui servent de point d'ancrage pour l'annotation en numéros EC, en utilisant le serveur KAAS (*i.e.* réalisation d'une une analyse BBH contre KEGG GENES) et le pipeline Trinotate (*i.e.* recherche d'homologie de séquences avec BLAST contre UniProt).

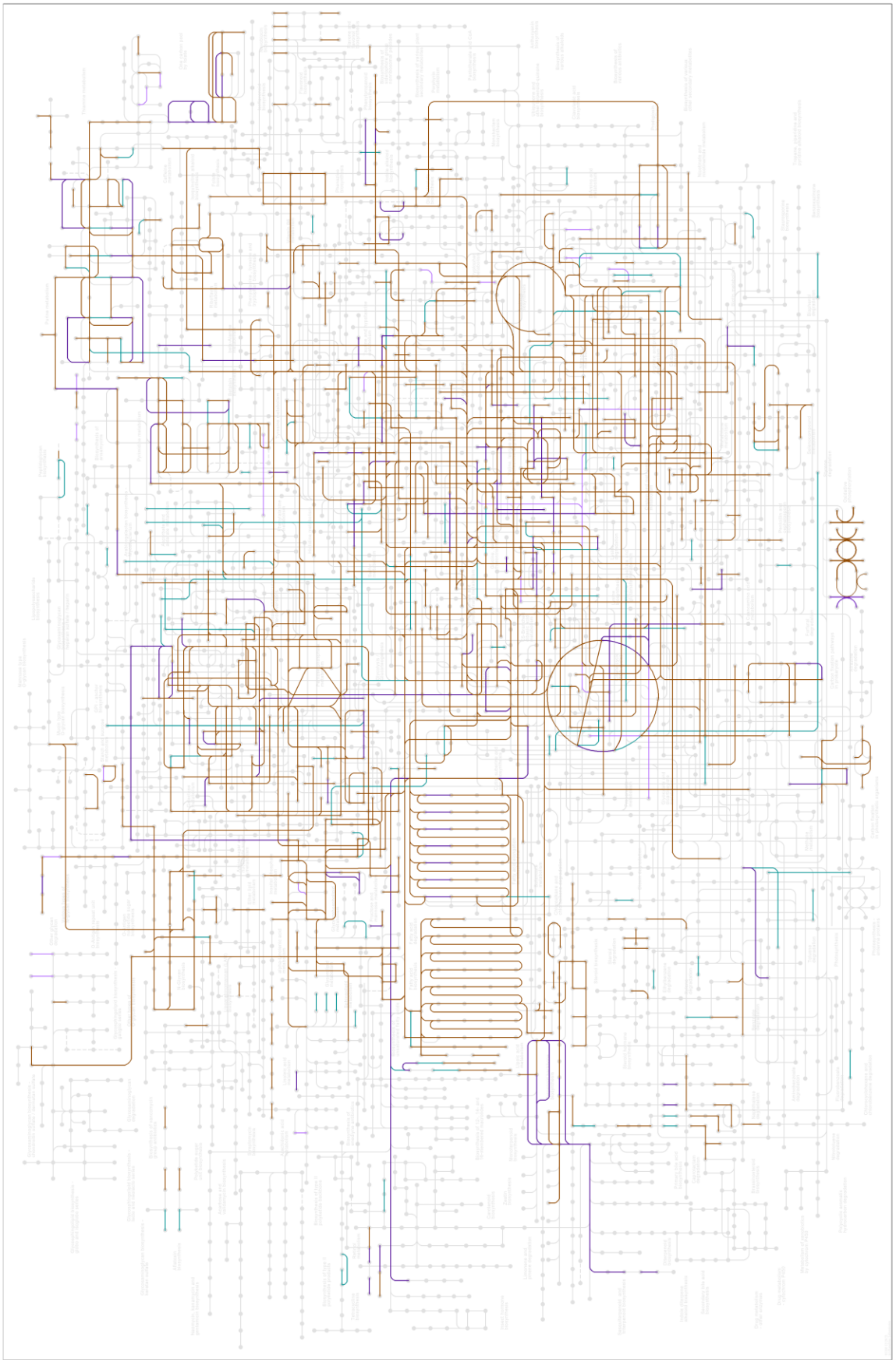
Toutefois, il est à noter que ces méthodes reposent toutes sur l'interrogation de la base de données KEGG. Ainsi, explorer des bases de données alternatives aurait potentiellement permis d'étendre la couverture des annotations. Néanmoins, en choisissant de combiner plusieurs approches, nous améliorons la fiabilité et la précision des annotations, tout en compensant les éventuelles lacunes spécifiques à chaque méthode. Cette stratégie permet de réduire les risques d'erreurs, en particulier les faux négatifs (*i.e.* annotations manquantes), et confère un poids accru à des annotations corroborées par plusieurs approches, renforçant ainsi leur taux de confiance.

L'ensemble des informations collectées a été agrégé dans des fichiers d'annotations (*i.e.* de type `*.gbk` ou `*.gff`) afin d'être interprété par Pathways Tools. Nous exploitons finalement la sortie de cet outil sous AuReMe pour produire le sous-réseau d'annotation fonctionnelle de *P. rubens*. Ce dernier est composé de 3 205 réactions soutenues par 3 390 gènes.

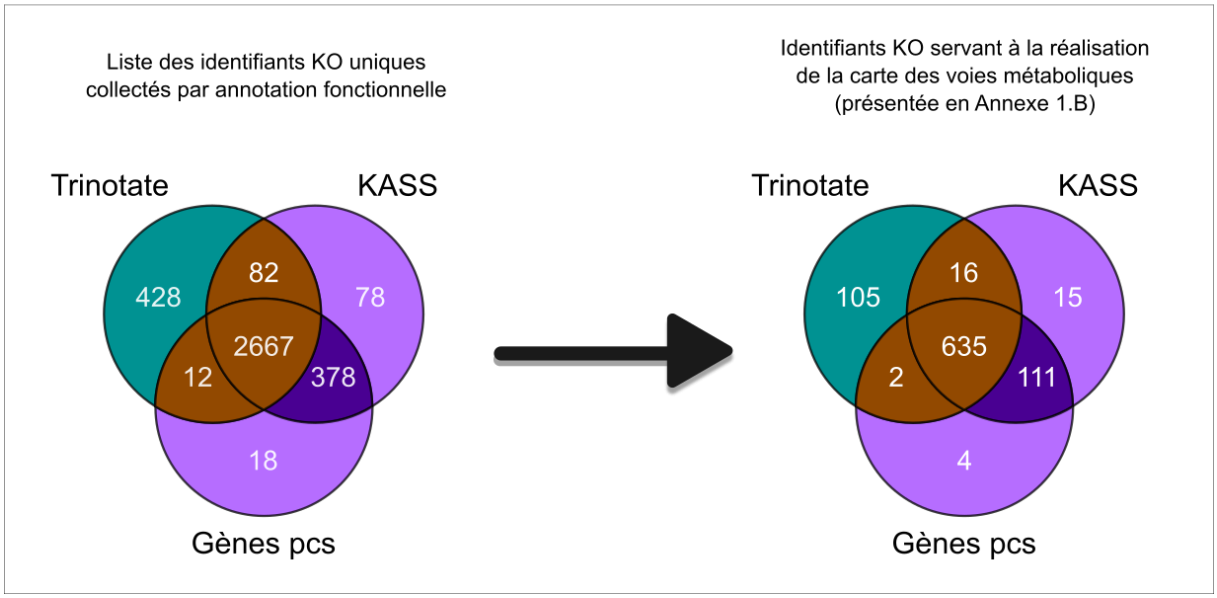
Pour conclure, le diagramme UpSet présenté en fin de cette annexe, illustre la répartition des trois catégories d'annotations, à savoir les numéros EC, les GOT, et les domaines protéiques (PFAM), pour les 8 948 séquences annotées de *P. rubens*. Parmi ces gènes, 3 390 contribuent à la présence d'au moins une réaction dans le sous-réseau d'annotation fonctionnelle. Les annotations correspondantes sont mises en évidence en bleu et les gènes concernés correspondent à la seconde ligne, matérialisée par une flèche, du diagramme UpSet.

Plus précisément, le bleu foncé (■) représente les 2 259 séquences incluses dans le réseau et annotées avec au moins un numéro EC, au moins un terme issu de la Gene Ontology, et au moins un domaine protéique PFAM. En contraste, le bleu clair (■) correspond à l'ensemble des gènes présents également dans ce sous-réseau, mais annotés seulement avec un ou deux types d'annotations. Il est à souligner que parmi l'ensemble des gènes annotés avec un numéro EC, seulement 14 séquences sont absentes du sous-réseau d'annotation fonctionnelle.

B.1 Sous-réseau d'annotation fonctionnelle (carte des voies métaboliques 01100)



B.2



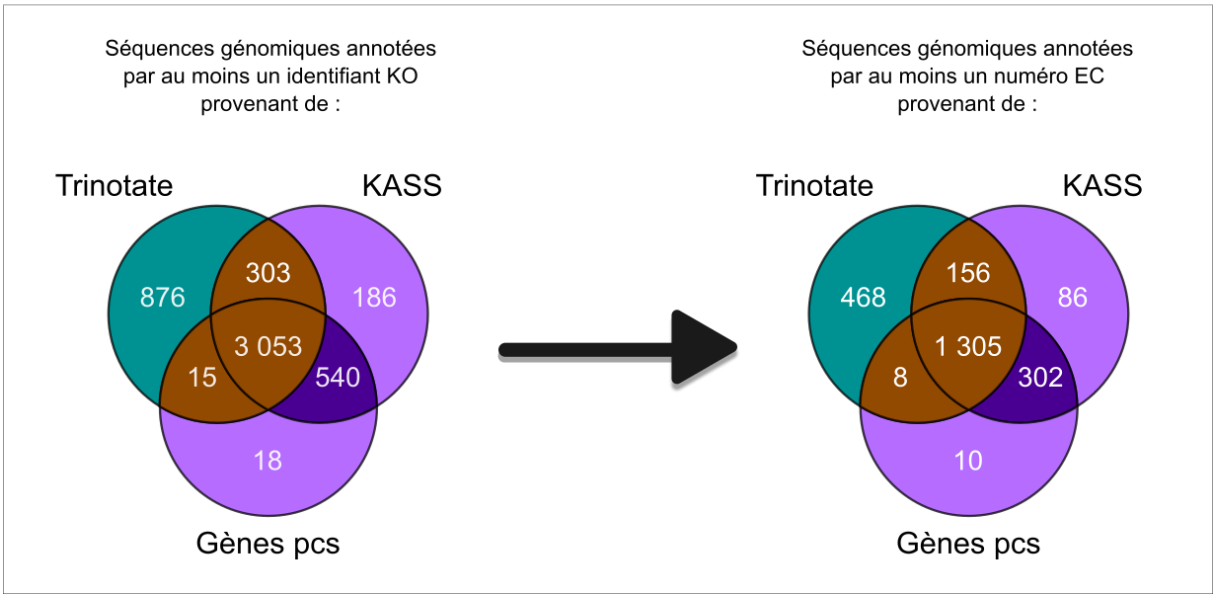
Annexe 1B : *Visualisation des réactions KEGG liées aux identifiants KO obtenus lors de l'annotation fonctionnelle du génome. B.1 Carte de référence des principales voies métaboliques obtenues avec KEGG Mapper. La couleur des réactions correspond à celle des intersections des diagrammes de Venn présentés ci-après B.2 Origine des identifiants KO détectés par annotation fonctionnelle. Le diagramme de Venn de gauche représente les 3 663 identifiants KO déterminés par annotation fonctionnelle, tandis que celui de droite se focalise sur les identifiants KO ayant servi à la génération de la carte présentée en B1.*

Avant de convertir nos identifiants KO en numéros EC en vue d'utiliser Pathways Tools pour la génération du sous-réseau d'annotation sous MetaCyc, nous avons exploité les ressources de la base de données KEGG. Les Annexes B1 et B2 ont pour objectif d'illustrer la complémentarité des approches liées à la génération du sous-réseau d'annotation fonctionnelle. En effet, selon l'approche choisie (*i.e.* pipeline Trinotate, serveur KAAS, ou exploitation des annotations préexistantes sur KEGG GENOME gènes pcs), de légères différences dans l'attribution des identifiants KO sont observées. À titre illustratif, la mise à jour de l'annotation fonctionnelle conduit à l'identification de 16 % d'identifiants KO supplémentaires.

L'annotation fonctionnelle du génome de *P. rubens* permet d'établir une liste de 3 663 identifiants KO distincts, dont 888 sont impliqués dans la création de la carte présentée en B.1. Toutefois, il convient de souligner que la présence d'un identifiant KO ne garantit pas nécessairement une correspondance avec une réaction KEGG, et encore moins avec une réaction MetaCyc. En outre, un identifiant KO, bien qu'associé à une fonction biologique spécifique, peut être lié à plusieurs réactions.

Ainsi, pour illustrer la complémentarité des approches susmentionnée, nous avons utilisé la carte de voies métaboliques la plus générale disponible avec KEGG Mapper (*i.e.* map 01100 composée de 4 587 réactions et 3 297 substances chimiques). Les réactions présentées sur cette carte résultent de la détection des identifiants KO obtenus exclusivement par Trinotate (■ - 105 KO), par KAAS ou par l'annotation des gènes pcs issus de KEGG (■ - 19 KO), par KAAS et par les gènes pcs (■ - 111 KO), ainsi que par Trinotate et par KAAS ou/et par les gènes pcs (■ - 653 KO).

C



Annexe 1C : *Diagrammes de Venn illustrant la répartition des séquences génomiques annotées par des identifiants KO et des numéros EC en fonction des sources d'annotations sélectionnées.*

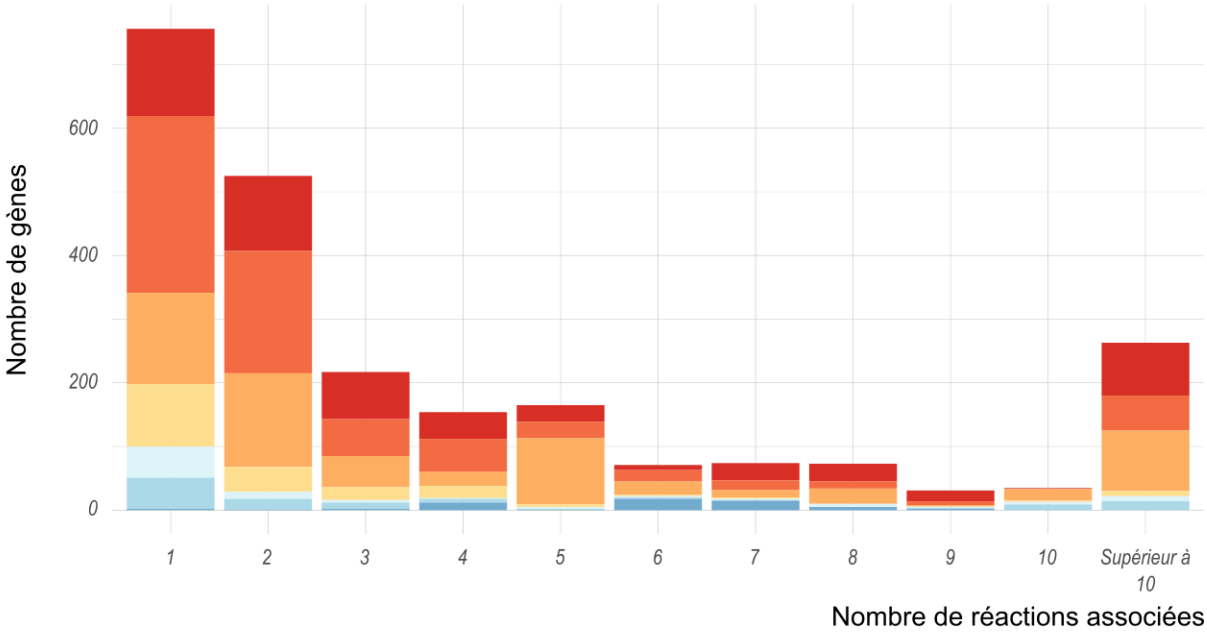
Comme mentionné précédemment, la reconstruction du sous-réseau d'annotation repose sur les numéros EC attribués aux séquences génomiques par le pipeline Trinotate, la recherche de meilleurs hits (BBH) réalisée avec KAAS et sur les annotations disponibles dans la base de données KEGG GENOME. Ces deux diagrammes de Venn permettent de quantifier le nombre de séquences génomiques annotées par au moins un identifiant KO, puis par un numéro EC en fonction de chaque approche.

La fusion de ces données aboutit à une liste de 4 991 gènes annotés avec au moins un identifiant KO, parmi lesquels 2 335 seront associés à au moins un numéro EC. Ce ratio d'environ 2:1 indique qu'un gène sur deux possédant un identifiant KO sera associé à un numéro EC.

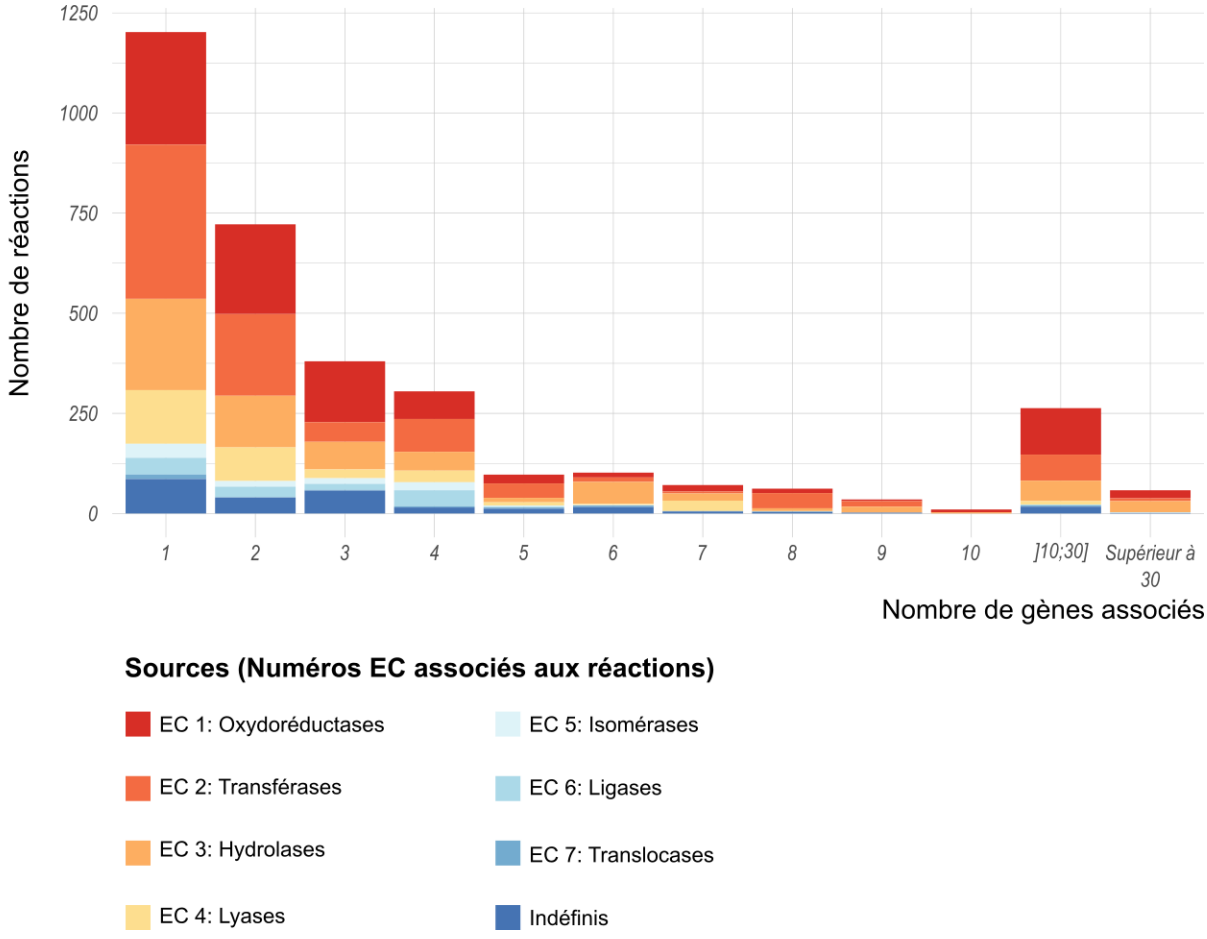
Nous constatons également que la quasi-intégralité des annotations initiales de *P. rubens* est retrouvée avec Trinotate et KAAS puisque 61 % des gènes associés à un KO et 55 % des gènes associés à un numéro EC sont communs aux trois approches. Par ailleurs, 11 % de l'ensemble des gènes avec un identifiant KO et 13 % des gènes avec un numéro EC proviennent à la fois des gènes préalablement annotés (*i.e.* gènes pcs) et des approches Trinotate ou KAAS.

En conclusion, la mise à jour de l'annotation fonctionnelle, apportée exclusivement par Trinotate et/ou KAAS, correspond respectivement à un apport de 27 % et 30 % de gènes annotés avec un identifiant KO et un numéro EC, indiquant ainsi une amélioration de la couverture métabolique.

D.1 **Classification des gènes (sous-réseau annotation)**



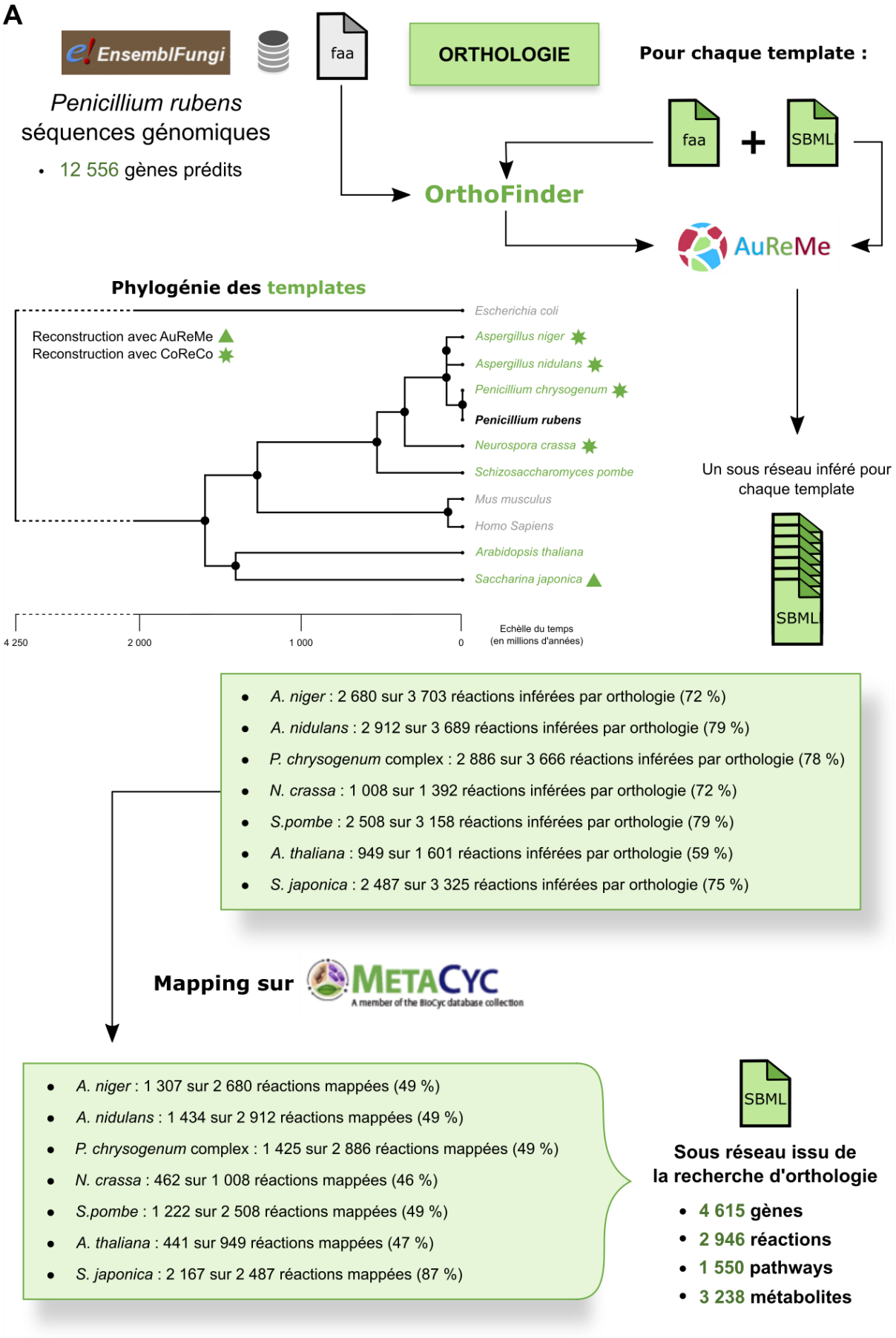
D.2 **Classification des réactions (sous-réseau annotation)**



Annexe 1D : *Répartition des classes enzymatiques dans le sous-réseau d'annotation : analyse de la distribution génique et réactionnelle. Diagrammes à barres représentant le nombre de gènes par nombre de réactions (**D.1**) et le nombre de réactions par nombre de gènes (**D.2**).*

ANNEXE 2 :

Reconstruction d'*i*Prub22 – sous-réseau issu des recherches d'orthologie



Annexe 2A : *Résumé des étapes pour la reconstruction des sous-réseaux issus de la recherche d'orthologie.*

La seconde approche abordée dans la reconstruction du **GSMN** repose sur le principe suivant : deux gènes homologues, et plus spécifiquement orthologues, partagent la même fonction. Ainsi, en effectuant une recherche d'homologie entre les gènes de *P. rubens* et ceux d'organismes pour lesquels nous disposons de **GSMNs**, il devient possible d'inférer les réactions présentes au sein de ces modèles de réseaux aux données de l'espèce étudiée.

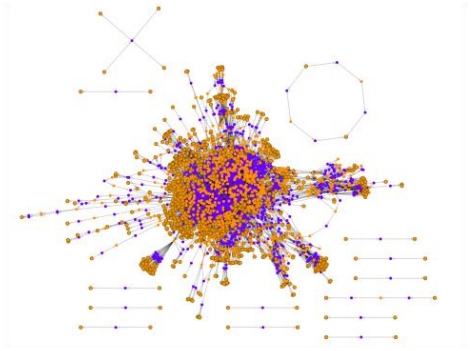
Le choix des organismes modèles, appelés *templates*, se veut le reflet d'une double diversité : biologique et computationnelle. Le premier aspect à prendre en compte est la proximité phylogénétique entre *P. rubens* et les *templates*. Opter pour des organismes phylogénétiquement éloignés des champignons, comme *A. thaliana*, facilite la mise en évidence du réactome « cœur » conservé à travers l'évolution. En opposition, nous prévoyons de faire émerger les réactions appartenant au métabolisme fongique en ciblant des modèles de champignons filamenteux, tels que les *Aspergillus*, afin de pointer les spécificités communes de ces espèces. Les sept organismes modèles sélectionnés pour cette étude sont représentés **en vert** sur l'arbre phylogénique.

Le second aspect à considérer est lié intrinsèquement à la méthode de reconstruction, aux outils utilisés et à la compatibilité et à la disponibilité des informations. Cependant, même si la diversification et la multiplication des outils augmentent le risque d'apporter du « bruit » à la topologie du réseau, la redondance étant un problème connu dans la reconstruction de **GSMNs**, ce phénomène pourra être contrebalancé, si nécessaire, lors des étapes de curation en s'appuyant sur la traçabilité des données. Cette caractéristique permettra, en outre, de détecter rapidement si les différences observées relèvent du domaine biologique ou informatique. C'est pourquoi, nous avons sélectionné le réseau de l'algue brune *S. japonica* dont le réseau a été reconstruit en suivant un protocole similaire à celui exposé ici (▲). Il est important également de noter que quatre des sept *templates* utilisés sont issus d'un processus de reconstruction automatisé à grande échelle réalisé à l'aide de CoReCo (✶).

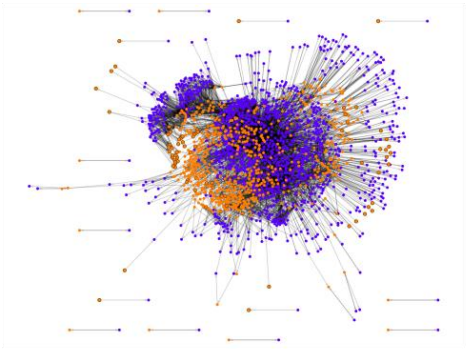
La recherche d'homologie entre les protéomes des *templates* et celui de *P. rubens* Wisconsin 54-1255 a été réalisée à l'aide d'OrthoFinder et de l'algorithme BLAST. Afin d'optimiser la recherche, nous avons extrait des protéomes des espèces d'intérêt uniquement les séquences présentes dans les **GSMNs** respectifs. Ces séquences ont été récupérées sur FungiDB et UniProt. De plus, des adaptations et un formatage dans l'écriture des fichiers **SBML** ont été effectués pour garantir l'interprétation et la correspondance des données.

La liste des protéines homologues à celles de *P. rubens* Wisconsin 54-1255 a ensuite été importée dans AuReMe afin d'inférer les réactions. Par la suite, un réseau a été généré pour chaque *template* étudié, et chacune des réactions a été mappée sur MetaCyc pour obtenir un ensemble cohérent. En conclusion, la fusion des résultats issus de la recherche d'orthologie permet d'obtenir un sous-réseau intermédiaire composé de 2 946 réactions associées à 4 615 gènes.

B

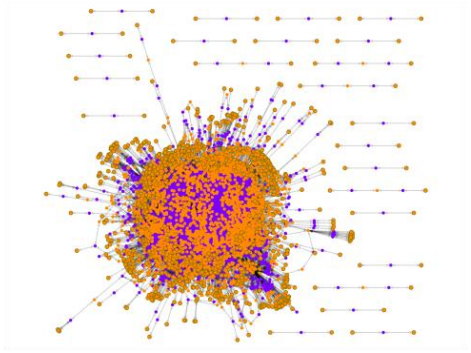


Arabidopsis thaliana
1 601 Réactions 1 737 Métabolites

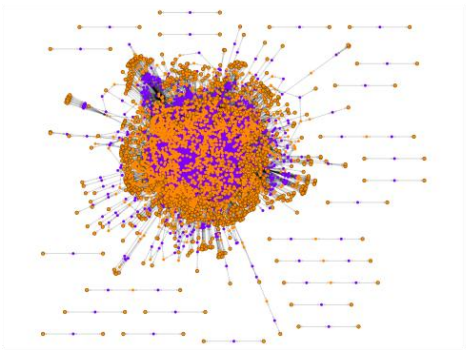


Neurospora crassa
1 392 Réactions 737 Métabolites

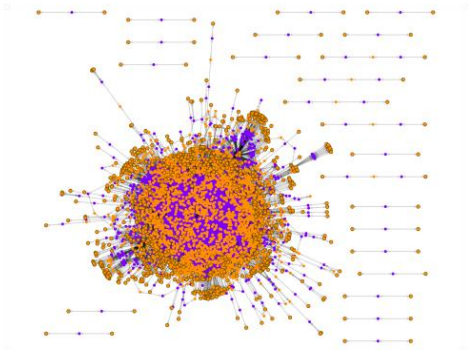
TEMPLATES



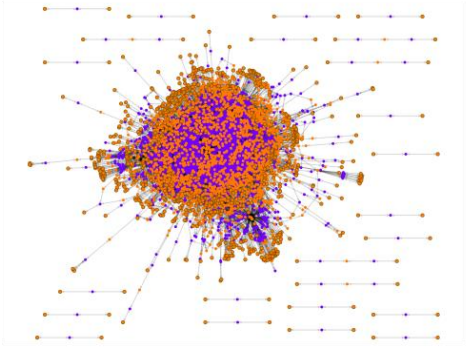
Aspergillus nidulans
3 689 Réactions 3 195 Métabolites



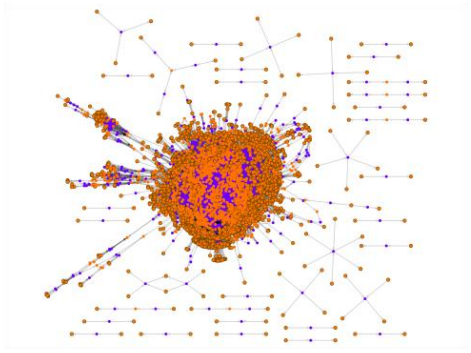
Aspergillus niger
3 703 Réactions 3 205 Métabolites



Penicillium chrysogenum
3 666 Réactions 3 175 Métabolites



Schizosaccharomyces pombe
3 158 Réactions 2 748 Métabolites



Saccharina japonica
3 325 Réactions 3 524 Métabolites

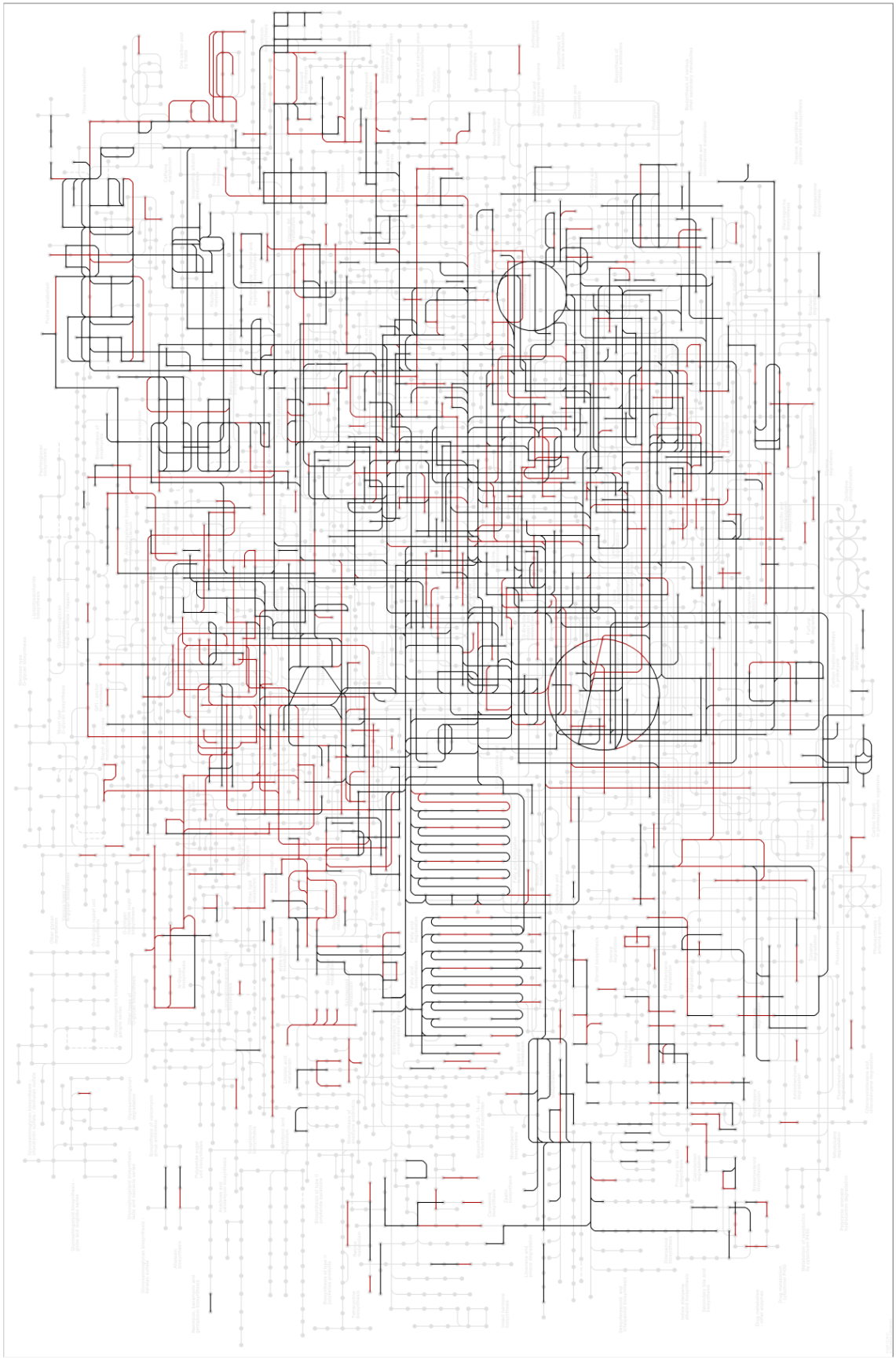
Modèle	Année de reconstruction	Base de données
Arabidopsis thaliana	2010	KEGG
Neurospora crassa	2014	Non déterminée
Aspergillus nidulans	2016	KEGG
Aspergillus niger	2016	KEGG
Penicillium chrysogenum sp.	2016	KEGG
Schizosaccharomyces pombe	2016	KEGG
Saccharina japonica	2018	MetaCyc

Annexe 2B : *Graphiques bipartites représentant la topologie des sept réseaux templates utilisés pour la génération du sous-réseau d'orthologie.*

Ces images, générées avec **ModelExplorer**, exposent la topologie des sept reconstructions employées pour la génération du sous-réseau d'orthologie. Les nœuds violets (■) symbolisent les réactions, tandis que les nœuds orange (■) représentent les métabolites. L'ensemble des **GSMNs** sélectionnés pour la recherche de liens d'orthologie ont été reconstruits entre 2010 et 2018 et proviennent, pour la majorité, de la base de données KEGG.

Les tailles approximatives en mégabases des génomes des organismes utilisés sont les suivantes : 135 Mb pour *A. thaliana*, 28 Mb pour *A. nidulans*, 34 Mb pour *A. niger*, 40 Mb pour *N. crassa*, 32 MB pour *P. chrysogenum*, 14 Mb (*i.e.* l'un des plus petits génomes eucaryotes recensés) pour *S. pombe*, et 545 Mb pour *S. japonica*. Néanmoins, bien que la taille du génome ne soit pas le seul déterminant de la complexité d'un organisme, elle peut offrir des indications sur ses capacités d'adaptation. Cependant, nous observons actuellement que la taille des réseaux semble être davantage liée à l'année de reconstruction qu'à la nature intrinsèque des organismes. Par exemple, même si le **GSMN** de *S. pombe* est légèrement plus petit que ceux des deux *Aspergillus* ou de *N. crassa*, le nombre d'entités (*i.e.* réactions et métabolites) des réseaux les plus récents est environ deux fois supérieur à ceux des réseaux d'*A. thaliana* et de *N. crassa*, reconstruits en 2010 et 2014.

C **Sous-réseau d'orthologie (carte des voies métaboliques 01100)**



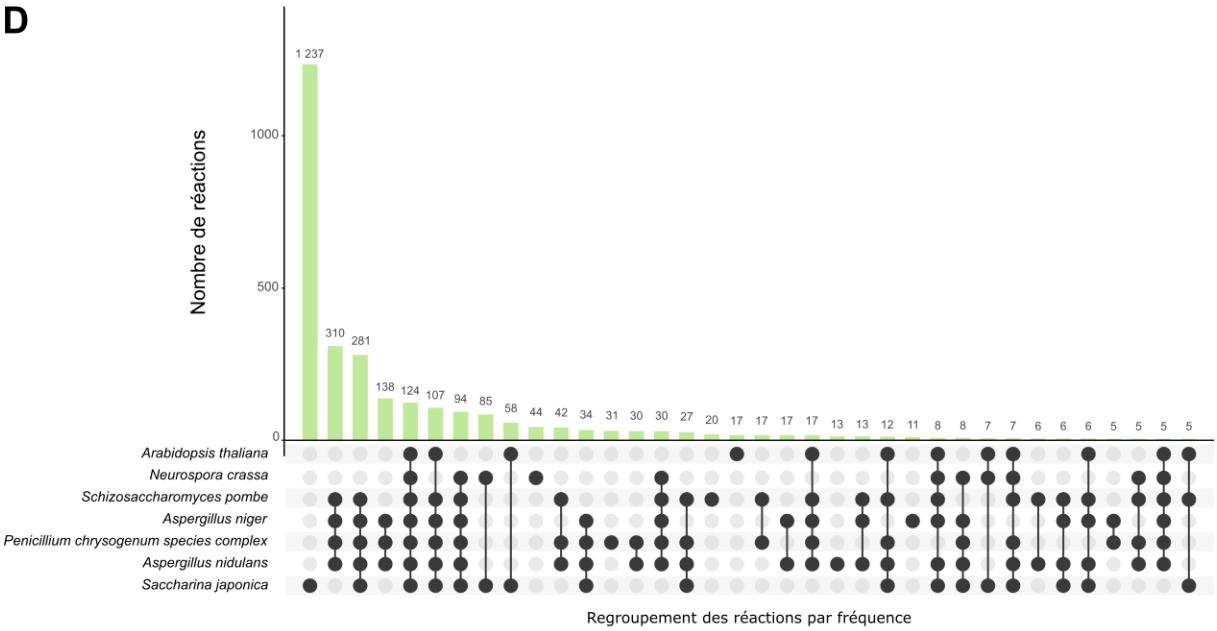
Annexe 2C : *Visualisation des réactions issues du sous-réseau d'orthologie (KEGG Mapper).*

À l'instar de la visualisation du sous-réseau d'annotation fonctionnelle, nous avons exploité la carte de voies métaboliques la plus générale disponible sur KEGG Mapper (*i.e.* map 01100), comprenant 4 587 réactions et 3 297 substances chimiques, afin de visualiser les données issues de la recherche d'orthologie sous KEGG. En revanche, la construction de la carte ne se base plus sur les identifiants KO mais sur les identifiants des réactions KEGG.

La recherche d'homologie de séquences, associée aux réactions correspondantes, permet, sur l'ensemble des sept *templates*, d'obtenir 6 401 identifiants non uniformisés de réactions. Ce mélange d'identifiants provient de diverses bases de données telles que KEGG, MetaCyc, BiGG, ainsi que d'identifiants dits « maison ».

Parmi cette diversité d'identifiants, nous recensons 2 687 identifiants appartenant à la base de données KEGG (*i.e.* identifiants de la forme `R\d{5}`). Il est ensuite essentiel d'assurer la cohérence de l'ensemble des identifiants pour obtenir un sous-réseau exploitable. Ainsi, étant donné que nous avons orienté notre reconstruction vers la base de données MetaCyc, nous procédons au mappage des identifiants vers cette dernière. Sur les 2 687 réactions KEGG, seules 1 323 d'entre elles trouvent une correspondance sur MetaCyc. Nous constatons alors que la mise en correspondance des identifiants entre les bases de données entraîne des pertes non négligeables d'informations, soulignant la nécessité de croiser les approches et les sources dans la reconstruction d'un réseau métabolique.

Enfin, la carte présentée sur cette figure est issue d'un sous-ensemble de 1 592 identifiants provenant des 2 687 identifiants de réactions KEGG obtenus par la recherche d'orthologie. Les identifiants sélectionnés correspondent à 35 % de l'ensemble des réactions présentes sur la carte 01100. Toutefois, ce pourcentage est à nuancer fortement, puisque 965 réactions seulement ont une correspondance sur MetaCyc. Ces 965 réactions sont indiquées en noir (■) sur cette figure, tandis que les 627 réactions « perdues » sont colorées en rouge (■) Ces chiffres, qui illustrent les potentialités et la complexité d'un réseau reconstruit à partir de la recherche d'homologie, soulignent également la perte d'informations inévitable liée au processus de mapping, une étape néanmoins indispensable.

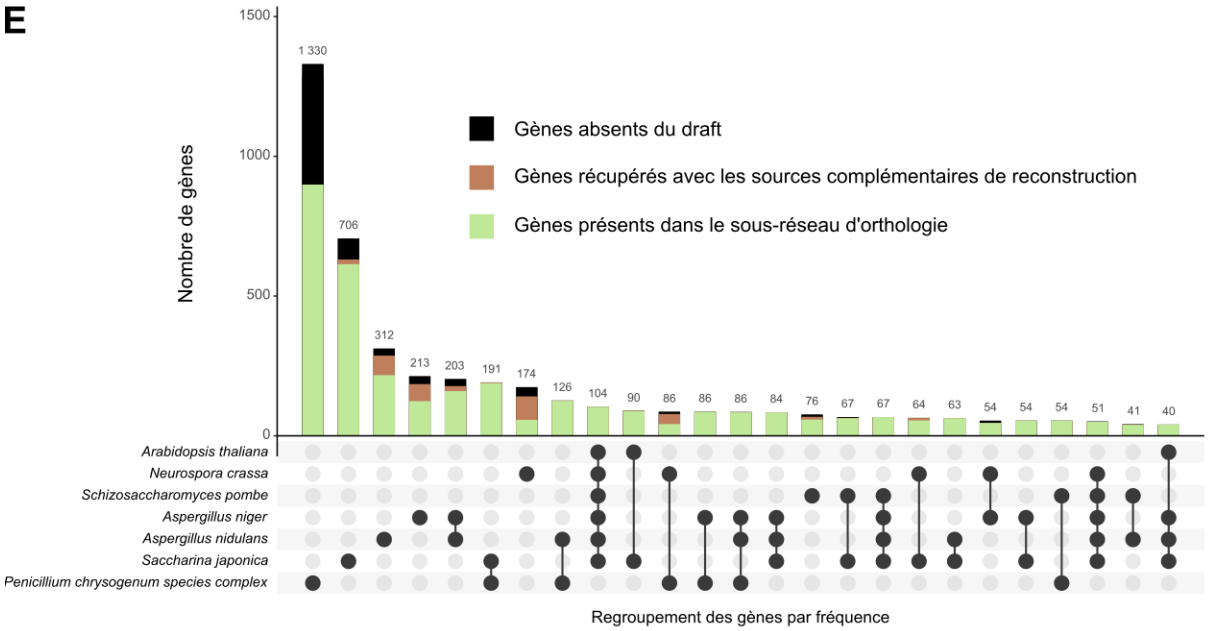


Annexe 2D : *Diagramme Upset représentant l'origine des réactions contenues dans le sous-réseau d'orthologie. Sur ce graphique, 2 877 réactions sur 2 946 sont affichées, les intersections avec moins de cinq réactions ne sont pas représentées.*

À la suite de la visualisation du sous-réseau d'orthologie, notre attention se focalise désormais sur l'origine de ces réactions. Seules celles pour lesquelles nous avons une correspondance sur MetaCyc ont été retenues. À noter également que la reconstruction axée sur la recherche d'orthologie offre l'opportunité de tirer parti d'informations vérifiées par des pairs, renforçant ainsi la pertinence de l'inclusion de certaines réactions dans le réseau.

En utilisant des organismes *templates* phylogénétiquement éloignés, notre objectif est de cibler le réactome basal, représentant le métabolisme fondamental conservé à travers l'évolution. En revanche, en optant pour des organismes plus proches sur le plan phylogénétique et en combinant les informations, nous espérons identifier, par différence, un ensemble de réactions spécifiques à un groupe d'organismes. En résumé, nous anticipons la capacité de distinguer des groupes de réactions visant, *in fine*, le métabolisme de base et le métabolisme spécialisé. De plus, en croisant les résultats provenant de différents *templates*, nous envisageons d'obtenir un soutien pour la curation manuelle. À titre illustratif, il serait cohérent d'identifier des groupes de réactions partagées exclusivement entre les *templates* de champignons (*e.g.* potentiellement l'ensemble de réactions constitué par les deuxième et quatrième intersections). En revanche, des réactions appuyées uniquement par la source *A. thaliana* devraient être interprétées avec prudence.

Cependant, l'analyse du diagramme Upset révèle que la majorité des réactions constituant le sous-réseau d'orthologie provient exclusivement de la source *S. japonica* (42 % des réactions), une algue brune bien éloignée des champignons filamenteux. Cette prédominance s'explique par l'utilisation d'un protocole de reconstruction équivalent entre le **GSMN** de *S. japonica* et le futur ***i*Prub22** (*i.e.* outils, bases de données et versions soit identiques, soit fortement similaires). En outre, comme la plupart des réactions de *S. japonica* sont affiliées à MetaCyc, la perte de réactions observée lors des phases de mappage est faible et l'utilisation de ce seul *template* aurait permis de générer un réseau composé de 2 167 réactions (*i.e.* 75 % de l'ensemble des réactions du sous-réseau d'orthologie). Ce point corrobore donc l'observation présentée en Annexe 2B, indiquant que les reconstructions réalisées dans les années 2010 semblent effectivement refléter, en premier lieu, les techniques de reconstruction, plutôt que les capacités métaboliques spécifiques d'un organisme.



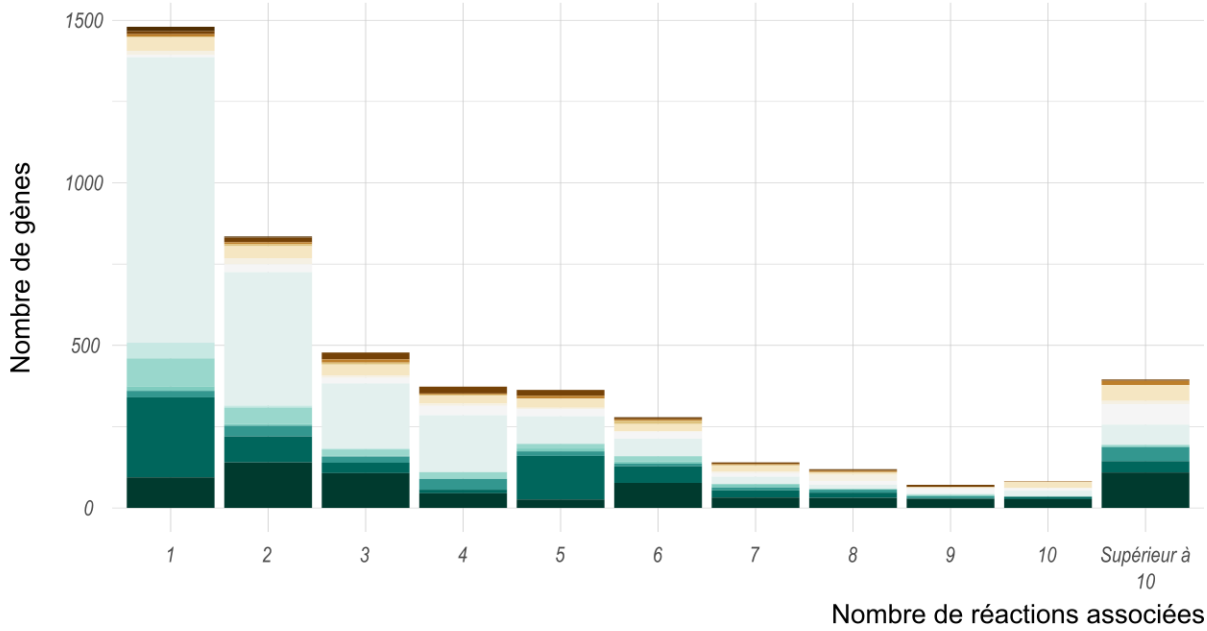
Annexe 2E : *Diagramme UpSet illustrant la distribution des séquences génomiques orthologues de P. rubens Wisconsin 54-1255 (i.e. orthogroupes). Sur ce graphique, 4 422 gènes sur 5 616 sont affichés, les intersections comportant moins de 40 gènes ne sont pas représentées.*

Le dernier aspect à aborder concerne la quantification de la couverture métabolique obtenue par la recherche d'homologie de séquences. Sur l'ensemble du protéome de *P. rubens* Wisconsin 54-1255, 5 616 protéines possèdent au moins une séquence orthologue dans l'un des sous-protéomes des sept modèles sélectionnés. Nous rappelons à cet effet que seules les séquences présentes dans les fichiers **SBML** des *templates* ont été interrogées (*i.e.* séquences génomiques qui, *de facto*, sont associées à des réactions).

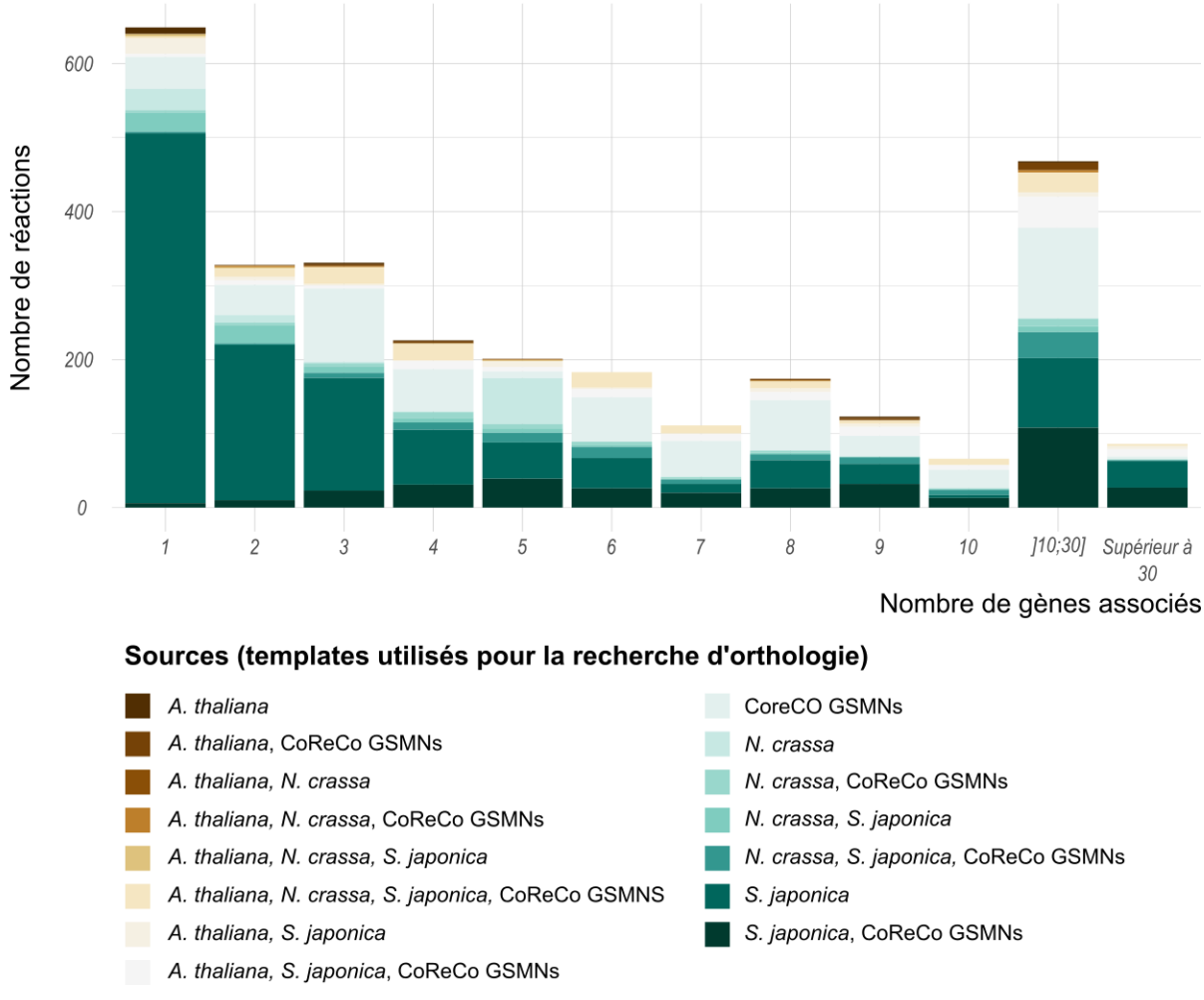
Parmi les séquences codantes associées, 1 023 ne sont pas incluses dans le sous-réseau d'orthologie en raison d'un manque d'interopérabilité entre les identifiants des réactions (■). Néanmoins, 716 d'entre elles seront ensuite récupérées et ajoutées via des sources externes ou le sous-réseau d'annotation fonctionnelle (■). Le sous-réseau d'orthologie est ainsi composé de 4 615 gènes détectés par recherche d'homologie (■).

Il est toutefois surprenant et regrettable de constater que de nombreux gènes associés à des réactions dans le *template* de *P. chrysogenum* sont perdus dans la version fusionnée du réseau (*i.e.* génération du draft suite à la fusion des données externes et des sous-réseaux d'annotation et de recherche d'orthologie).

F.1 **Classification des gènes (sous-réseau orthologie)**

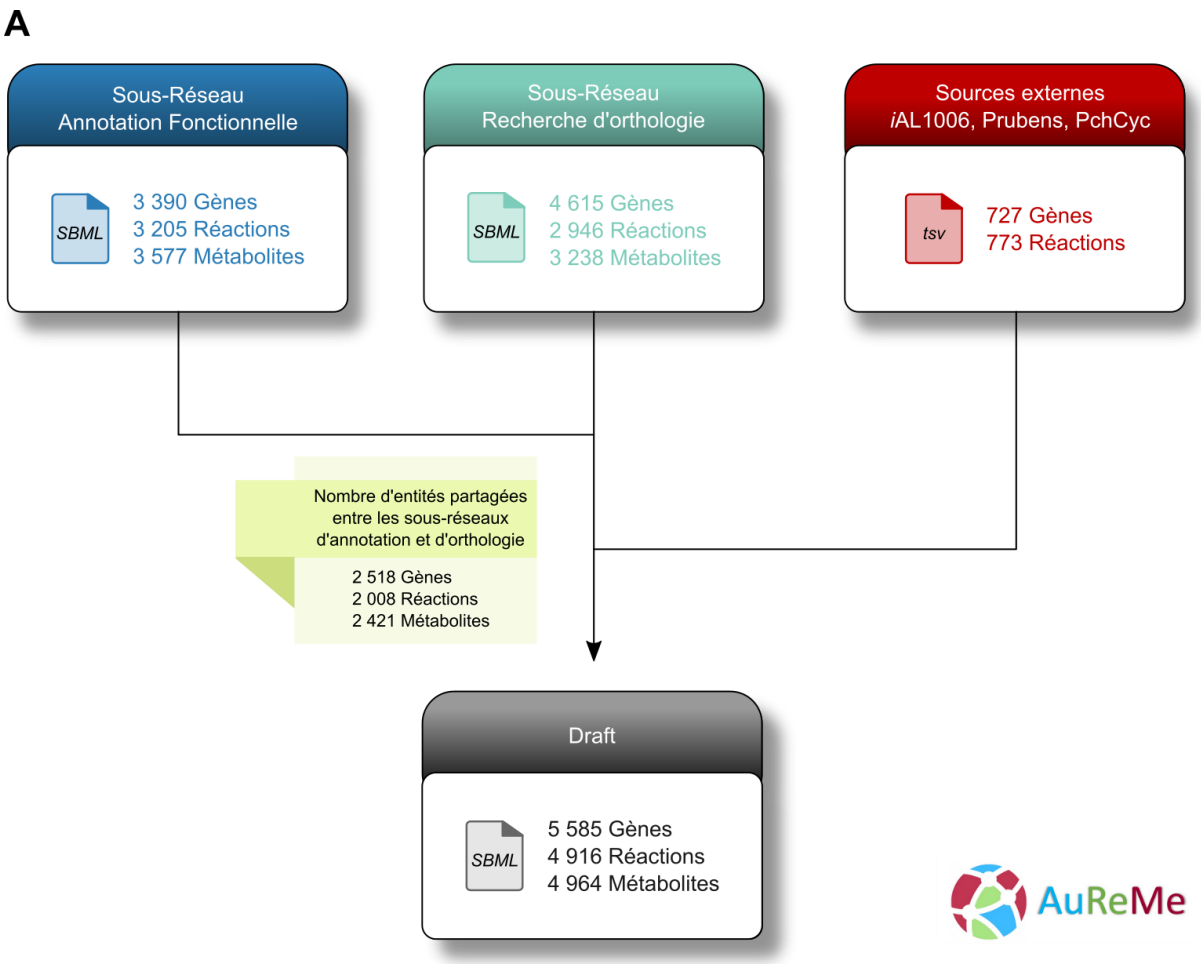


F.2 **Classification des réactions (sous-réseau orthologie)**



Annexe 2F : *Traçabilité des données et influence des templates sur le sous-réseau d'orthologie. Diagrammes à barres illustrant la distribution du nombre de gènes en fonction du nombre de réactions associées (F.1) et la distribution du nombre de réactions en fonction du nombre de gènes associés (F.2).*

ANNEXE 3 : Reconstruction d'*i*Prub22 – fusion des données et première ébauche de réseau (draft)

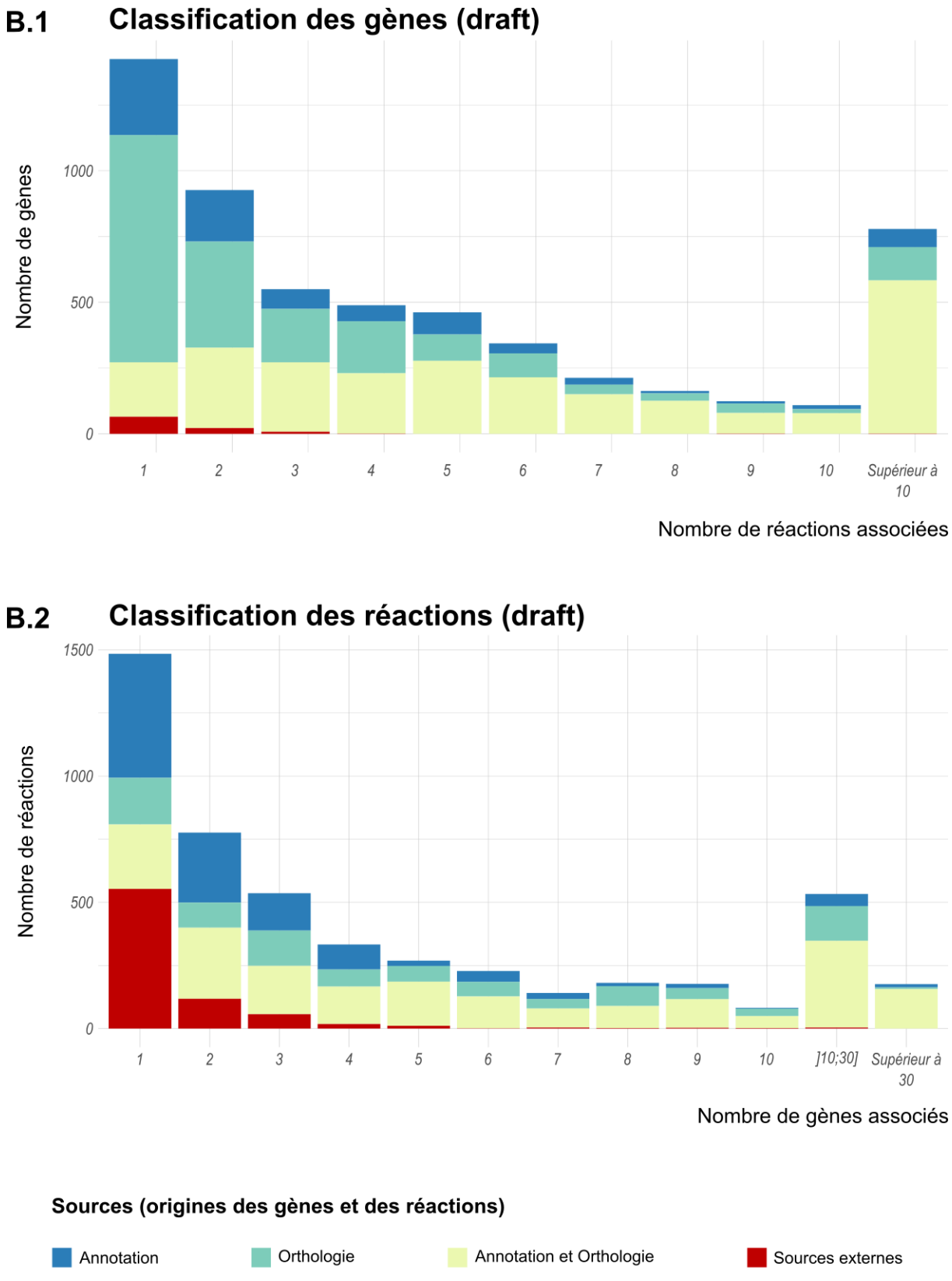


Annexe 3A : *Intégration des données et génération du draft.*

Le réseau résultant de la fusion des différentes sources est désigné sous le nom de draft. L'intégration des données s'effectue sous AuReMe, espace de travail dont l'un des attraits majeurs réside dans la possibilité de déterminer précisément l'origine des données incorporées au sein du réseau, augmentant ainsi notre niveau de confiance en celui-ci. Les sous-réseaux d'annotation et d'orthologie sont fusionnés, puis complétés par les données extraites des sources externes. Ainsi, 773 réactions, appuyées par 727 gènes - dont 98 n'avaient pas été détectés par ces deux approches - sont ajoutées au draft. À ce stade de la reconstruction, il est essentiel de souligner que toutes les réactions présentes dans le draft sont étayées par au moins une séquence génomique.

La connaissance de l'origine de chaque réaction ajoutée sert un double objectif. D'une part, elle fournit des informations sur les aspects évolutifs et phylogénétiques, mettant en évidence le génome cœur caractérisant principalement les gènes impliqués dans le métabolisme intermédiaire, la réplication et la réparation de l'ADN, ainsi que la transcription et la synthèse des protéines. D'autre part, du point de vue méthodologique et informatique, cette traçabilité permet d'appréhender la complémentarité des approches utilisées et leur pertinence.

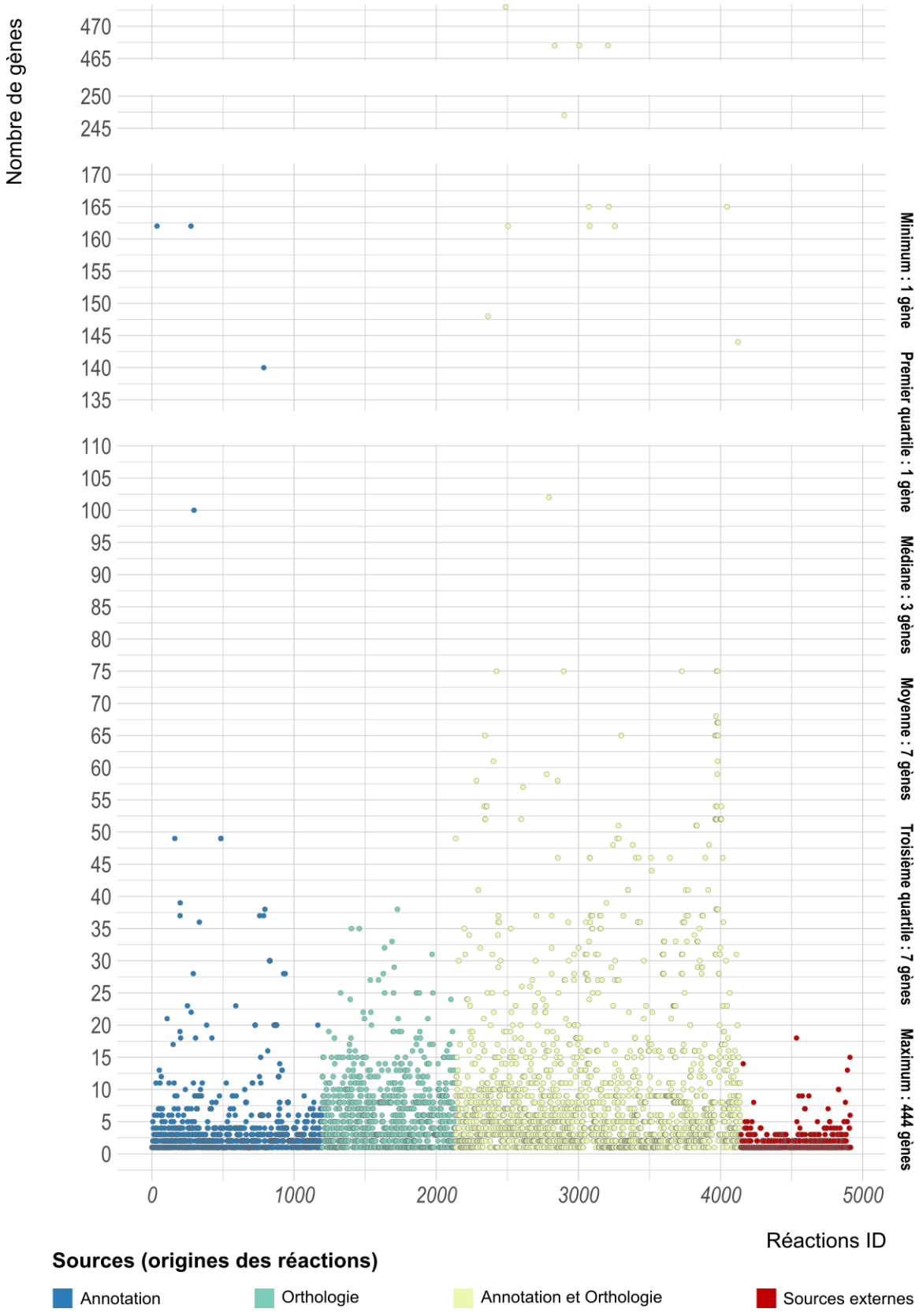
Les défis sont multiples pour refléter avec précision la complexité de la réalité biologique, particulièrement accentuée chez les organismes eucaryotes. Ainsi, diverses étapes de curation sont nécessaires afin de faire évoluer ce draft en un **GSMN** de haute qualité. Notons que le terme « curation », associé à l'amélioration et à l'optimisation d'un réseau, est générique et revêt de nombreux aspects. En effet, qu'elle soit manuelle ou partiellement automatique, la curation d'un réseau constitue un processus laborieux, chronophage et itératif.



Annexe 3B : *Diagrammes à barres représentant la classification des gènes et des réactions en fonction de leur source d'intégration dans le draft. Ces représentations détaillent les contributions de chaque sous-réseau en visualisant leur impact sur la présence des gènes et des réactions au sein du draft (i.e. 4 916 réactions et 5 585 gènes).*

C.1 Nombre de réactions associées par gène (draft)

C.2 Nombre de gènes associés par réaction (draft)



Annexe 3C : *Nuages de points illustrant la complémentarité des sources de reconstruction. Étant donné que les associations GPR ne sont pas des ensembles bijectifs, le graphique C.1 représente le nombre de réactions associées par gène, tandis que le graphique C.2 représente le nombre de gènes associés par réaction.*

ANNEXE 4 :

LiveScript MATLAB détaillant les caractéristiques d'*i*Prub22

Features of the *i*Prub22 reconstruction and model

Table of Contents

Introduction

In the following document, we present the final features of the genome-scale metabolic network ***reconstruction*** of *Penicillium rubens* Wisconsin 54-1255. We also propose a ***model*** (*i.e.* one parameterisation proposition of the reconstruction) based on the similarity of a minimum culture medium to mimic the organism growth. The model is encoded in SBML format, and its detailed characteristics are outlined in the subsequent sections.

The design and implementation of these structures follow the community standards and recommendations set out in the "**Community standards to facilitate development and address challenges in metabolic modeling**" (https://doi.org/10.15252/msb.20199235). The purpose of this article is to standardise reconstruction practice by giving a "*list as a guide to help standardize accessibility, content, and quality; however, more comprehensive documentation and more interpretable and accessible information can only improve the usability and biological relevance of the shared reconstruction.*" The recommendations suggested in this paper are followed by (▲).

0. Environment versions

To ensure the reproducibility of the data presented here, the versions of the tools used are:

- MATLAB version:

```
version()
```

```
ans = '9.5.0.1298439 (R2018b) Update 7'
```

- COBRA Toolbox version:

```
InitCobraToolbox;
```

- Gurobi version:

```
getCobraSolverVersion('gurobi') ;
```

```
> The version of GUROBI is 903.
```

1. Reconstruction

1.1 Distribution format

*i*Prub22.SBML (available here) is produced by the reconstruction process described in the Supplementary Material. It is available *via* BioModelsunder the identifier MODEL2306150001.

1.1.1 Format used consistency

SBML (Systems Biology Markup Language) is the standard file format chosen for storing and sharing our reconstruction. First of all, the format used for the model distribution is checked for consistency with **validateSBML.py** (one of the example programs demonstrating the use of different libSBML API calls available on https://synonym.caltech.edu/software/libsbml/libsbml-docs/examples/). More information is contained in **LibSBML: an API Library for SBML** (https://doi.org/10.1093/bioinformatics/btn051)

The output below shows a single warning occurring 86 times:

- **(99701 [Warning])** The SBOTerm used is not recognised by libSBML. Therefore, the appropriate parentage can not be checked. However, since libSBML is referring to a snapshot of the SBO tree, the term may now exist. Unknown SBO term 'SBO:0000672'. (detailed explanation here)

```
--------------------------------------------------------------------------------
                 filename : iPrub22.sbml
        file size (byte) : 35661786
          read time (ms) : 2819.700956
       c-check time (ms) : 4012.276888
     validation error(s) : 0
   (consistency error(s)): 0
   validation warning(s) : 86
 (consistency warning(s)): 86
```

Once the metabolic model's internal structure is correct, it is loaded into MATLAB:

```
tic
iPrub22_reconstruction = readCbModel('./Network/iPrub22.sbml');
%sparse format accelerates computations with large networks
iPrub22_reconstruction.S = sparse(iPrub22_reconstruction.S);
toc
```

```
Elapsed time is 172.196199 seconds.
```

1.1.2 Recognised naming convention (▲)

The existing model *i*Prub22 follows the recommended practice for model identifiers described in "**Community standards to facilitate development and address challenges in metabolic modeling**" (https://doi.org/10.15252/msb.20199235)

```
% ModelID is available here:
disp(iPrub22_reconstruction.modelID)
```

```
iPrub22_v1
```

NB: *i*Prub22v1 = ***in silico Penicillium rubens*** reconstruction published in 20**22**, version 1

```
% and software versioning is documented here:
disp(iPrub22_reconstruction.modelVersion)
```

```
SBML_level: 3
SBML_version: 1
fbc_version: 2
```

~~~~~~ [...] ~~~~~~

1.2 Metabolites

1.2.1 Model compartmentation

In the biological sense, the model is not compartmentalised (*i.e.* organelle modelling). Nevertheless, by convention, a GSMN is composed of an intracellular (cytosol) and extracellular compartment. This information is carried by a suffix label identifying the metabolites.

```
%Compartments characteristic
Name = iPrub22_reconstruction.compNames ;
Id = iPrub22_reconstruction.comps ;
MetaNetX = iPrub22_reconstruction.compismetanetx__46__compartmentID ;
GOT = iPrub22_reconstruction.compisgoID ;
[metId, comps] = strtok(iPrub22_reconstruction.mets,'[') ;
Number_of_metabolite = [length(findMetFromCompartment(iPrub22_reconstruction,"c"));
  length(findMetFromCompartment(iPrub22_reconstruction,"e"))];
disp(table(Name,Id,MetaNet, GOT,Number_of_metabolite))
```

Name	Id	MetaNetX	GOT	Number_of_metabolite
'cytosol'	'c'	'MNXC3'	'GO:0005829'	5213
'extracellular'	'e'	'MNXC2'	'GO:0005576'	1404
~~~~~~

```
fprintf('Total number of metabolites: <strong>%d</strong>\nNumber of unique',...
  ' metabolites: <strong>%d</strong>', length(iPrub22_reconstruction.mets),...
  length(unique(metId)))
```

```
Total number of metabolites: 5464
Number of unique metabolites: 5192
```

1.2.2 Human readable descriptive name (▲)

As the metabolite identifier is unique but not necessarily informative about the entity's nature, each metabolite is associated with a generic name to simplify the model understanding. The following example, representing the different penicillins present in the GSMN, illustrates this point. Of these four identifiers, only two are humanly understandable.

```
disp(iPrub22_reconstruction.metNames(findMetIDs(iPrub22_reconstruction,{'PENICILLIN-
G[c]','CPD-9122[c]','PENICILLIN-N[c]','CPD-9196[c]'})));
```

```
'penicillin G'
'penicillin K'
'penicillin N'
'penicillin V'
```

1.2.3 Structure identifiers (▲)

- **InChI (▲) and InChIKey strings**

InChI and InChIKey (hashed version of the full InChI) are unique descriptions and identifiers for chemical substances.

```
% Visualisation
for i = 1:2
fprintf('Metabolite name: <strong>%s</strong>\nInChI: %s\nInChIKey: %s\n\n',...
  iPrub22_reconstruction.metNames{i}, iPrub22_reconstruction.metInChIString{i},...
  iPrub22_reconstruction.metisinchikeyID{i}
end
```

```
Metabolite name: &alpha;-D-glucopyranose 1-phosphate
InChI: InChI= 1S/C6H13O9P/c7-1-2-3(8)4(9)5(10)6(14-2)15-16(11,12)13/h2-10H,1H2,(H2,11,12,13)/p-
2/t2-,3-,4+,5-,6-/m1/s1
InChIKey: HXXFSFRBOHSIMQ-VFUOTHLCSA-L

Metabolite name: &alpha;-D-glucose 6-phosphate
InChI:InChI= 1S/C6H13O9P/c7-3-2(1-14-16(11,12)13)15-6(10)5(9)4(3)8/h2-10H,1H2,(H2,11,12,13)/p-2/t2-
,3-,4+,5-,6+/m1/s1
InChIKey: NBSCHQHZLSJFNQ-DVKNGEFBSA-L
```

~~~~~~ [...] ~~~~~~

- **Charge and chemical formula (▲)**

```
% Visualisation
for i = 1:2
fprintf('Metabolite name: <strong>%s</strong>\nFormulae: %s\nCharge: %d\n\n',...
  iPrub22_reconstruction.metNames{i},iPrub22_reconstruction.metFormulas{i},...
  iPrub22_reconstruction.metCharges(i))
end
```

```
Metabolite name: &alpha;-D-glucopyranose 1-phosphate
Formulae: C6H11O9P
Charge: -2

Metabolite name: &alpha;-D-glucose 6-phosphate
Formulae: C6H11O9P
Charge: -2
```


~~~~~~

```
%Sum up
fprintf(['Of the %d metabolites present in the reconstruction:\n\n',...
  ' %d have a <strong>chemical formula</strong> (%.1f%%)\n %d have a',...
  ' <strong>charge</strong>(%.1f%%)'],...
  length(iPrub22_reconstruction.mets),
  length(iPrub22_reconstruction.metFormulas(~cellfun('isempty',
                                                iPrub22_reconstruction.metFormulas))),...
  length(iPrub22_reconstruction.metFormulas(~cellfun('isempty',
           iPrub22_reconstruction.metFormulas)))*100/ length(iPrub22_reconstruction.mets),...
  sum(~isnan(iPrub22_reconstruction.metCharges)),...
  sum(~isnan(iPrub22_reconstruction.metCharges))*100/length(iPrub22_reconstruction.mets));
```

```
Of the 5464 metabolites present in the reconstruction:

   5272 have a chemical formula (96.5%)
   5436 have a charge (99.5%)

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ [...] ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

- **SMILES (▲) and molecular mass**

SMILES and molecular weight are encoded in the notes tag.

```
% Visualisation
for i = 1:2
fprintf('Metabolite name: <strong>%s</strong>\n%s\n\n',...
  iPrub22_reconstruction.metNames{i},iPrub22_reconstruction.metNotes{i})
end
```

```
Metabolite name: &alpha;-D-glucopyranose 1-phosphate
mass: 258.119901
smiles: OC[C@H]1O[C@H](OP([O-])([O-])=O)[C@H](O)[C@@H](O)[C@@H]1O

Metabolite name: &alpha;-D-glucose 6-phosphate
mass: 258.119901
smiles: O[C@H]1O[C@H](COP([O-])([O-])=O)[C@@H](O)[C@H](O)[C@H]1O
```

```
%Sum up
fprintf(['Of the %d metabolites present in the reconstruction:\n\n' ...
  ' %d have a <strong>molecular weight</strong> (%.1f%%)\n' ...
  ' %d have a <strong>SMILES identifier</strong>(%.1f%%)'],...
  length(iPrub22_reconstruction.metNotes),...
  sum(count(iPrub22_reconstruction.metNotes,"mass")),...
  sum(count(iPrub22_reconstruction.metNotes,"mass"))*100/
                                          length(iPrub22_reconstruction.metNotes),...
  sum(count(iPrub22_reconstruction.metNotes,"smiles")),...
  sum(count(iPrub22_reconstruction.metNotes,"smiles"))*100/
                                          length(iPrub22_reconstruction.metNotes))
```

```
Of the 5464 metabolites present in the reconstruction:

5270 have a molecular weight (96.4%)
5276 have a SMILES identifier(96.6%)
```

1.2.4 At least one database identifier from a reliable resource (▲)

More cross-references are better because:

1. allows for the unambiguous identification of the entity concerned
2. allows to work more freely with different databases and facilitates exchanges
    - BioCyc
    - MetaNetX
    - (ChEBI, PubChem, ChemSpider, CAS)
    - (BiGG, KEGG, ModelSEED, SABIO-RK)
    - (LIPIDMAPS, SwissLipids, DrugBank, KNApSAcK, MetaboLights, UM-BBD, HMDB)

```
% Lines of code to generate the following table
retrieving the number of annotations per database and calculating percentages
(not displayed here, for more details please refer to the original document)
```

	Number_of_annotations		Number_of_unique_annotations	
BioCyc	5441	99.6	5171	95
ChEBI	3568	65	3311	61
Pubchem	3693	68	3484	64
Chemspider	1487	27	1331	24
MetaNetX	5436	99	5079	93
BiGG	717	13	591	11
KEGG	2371	43	2176	40
ModelSEED	4	0.073	2	0.037
LIPIDSMAPS	138	2.5	136	2.5
SABIORK	12	0.22	10	0.18
SwissLipids	26	0.48	24	0.44
CAS	959	18	816	15
DrugBank	237	4.3	202	3.7

1.2.5 SBO terms (▲)

```
fprintf('%d metabolites are annotated with <strong>%s</strong>',...
  length(iPrub22_reconstruction.metSBOTerms(~cellfun('isempty',
                                                iPrub22_reconstruction.metSBOTerms))),...
  char(unique(iPrub22_reconstruction.metSBOTerms)));
```

5464 metabolites are annotated with **SBO:0000247**

- SBO:0000247 stands for 'simple chemical'

1.3 Biochemical reactions

Reactions in *i*Prub22

- Reactions with a MetaCyc-compatible identifier
- Reactions from *i*AL1006
- Modelling reactions (artificial) (**Transport_\d{3} - Uptake_\d{3} - Demand_\d{3} - Sink_\d{3} - Production_\d{3}**)
- Specific reactions (**NGAM - Biomass_rxn - Transport_Biomass - Exchange_Biomass)**

~~~~~~ [...] ~~~~~~

1.3.1 Metadata (▲)

As for the metabolites, to ensure traceability and accessibility of data, the reactions are provided with the following features:

- **ID**
- **Name**
- **Formulae**
- **EC number**
- **GPRs associations**
- **Reconstruction sources**

Each reaction is associated with an **identifier**, a **name** and a **formula**. Most of the identifiers come from the MetaCyc database, and a MetaNetX has been done to enrich these data. Each reaction has a common name describing its nature, or failing that, gene name leading to its catalyses. **Enzyme Commission number** is s a numerical classification for enzymes based on the chemical reactions they compartmentation. Therefore, not all reactions of a model (*i.e.* exchange reactions, artificial transport) are bound to be annotated by this type of object. Similarly, not all reactions are linked to **GPR associations**. Reaction justification in the network (sources) is encoded in the notes tag. **ANNOTATION** and **ORTHOLOGY** mean that this reaction is supported by their respective subnetwork. On the other hand, the **MANUAL** tag may represent data from external sources, gap-filling or manual curations performed.

```
% Visualisation
formulas = printRxnFormula(iPrub22_reconstruction,'printFlag',false, 'gprFlag',true) ;
for i = 1:5
  % GPRs
  genes = regexp(iPrub22_reconstruction.rules{i},'x\(([0-9]+)\)', 'tokens');
  rule = iPrub22_reconstruction.rules{i} ;
  for j = 1:length(genes)
    rule = regexprep(rule,'x\([0-9]+\)',
                '${iPrub22_reconstruction.genes{str2num(char(genes{j}))}}', 'once') ;
  end
  rule = regexprep(rule,'\|','or'); rule = regexprep(rule,'&','and');
  rule = regexprep(rule,'( ','('); rule = regexprep(rule,' )',')') ;
  % Others data
  fprintf(['Reaction <strong>ID</strong>: %s\n',...
    'Reaction <strong>Name</strong>: %s\n',...
    'Reaction <strong>Formula</strong>: %s\n',...
    '<strong>Lower bound</strong>: %d\n',...
    '<strong>Upper bound</strong>: %d\n',...
    '<strong>EC number(s)</strong>: %s\n'...
    '<strong>GPR associations</strong>: %s\n',...
    '<strong>Source(s)</strong>: %s\n\n'],...
```
~~~~~~

```
    iPrub22_reconstruction.rxns{i},...
    iPrub22_reconstruction.rxnNames{i},...
    formulas{i},...
    iPrub22_reconstruction.lb(i),...
    iPrub22_reconstruction.ub(i),...
    iPrub22_reconstruction.rxnECNumbers{i},...
    rule,...
    iPrub22_reconstruction.rxnNotes{i})
end
```

```
Reaction ID: 1-ACYLGLYCEROL-3-P-ACYLTRANSFER-RXN
Reaction Name: Pc12g13010_product
Reaction Formula: ACYL-ACP[c] + ACYL-SN-GLYCEROL-3P[c]  -> ACP[c] + L-PHOSPHATIDATE[c]
Lower bound: 0
Upper bound: 1000
EC number(s): 2.3.1.51
GPR associations: (gp_Pc13g04040 or (gp_Pc16g02170 and gp_Pc16g09860 and gp_Pc16g07520) or
gp_Pc20g00970 or gp_Pc12g04190 or gp_Pc16g08280 or gp_Pc18g02500 or gp_Pc12g13010)
Source(s): CATEGORIES: ANNOTATION and ORTHOLOGY
```

~~~~~~ [...] ~~~~~~

1.3.4 SBO terms (▲)

SBO terms for reactions are extremely useful to clearly distinguish a few categories of reactions without having to rely on naming conventions.

- SBO:0000167 stands for **'biochemical or transport reaction'**
- SBO:0000176 stands for **'biochemical reaction'**
- SBO:0000185 stands for **'translocation reaction'**
- SBO:0000627 stands for **'exchange reaction'**
- SBO:0000628 stands for **'demand reaction'**
- SBO:0000629 stands for **'biomass production'**
- SBO:0000630 stands for **'ATP maintenance'**
- SBO:0000632 stands for **'sink reaction'**
- SBO:0000655 stands for **'transport reaction'**
- SBO:0000657 stands for **'active transport'** (child term of SBO:0000655)
- SBO:0000658 stands for **'passive transport'** (child term of SBO:0000655)
- SBO:0000659 stands for **'symporter-mediated transport'** (grandchild term of SBO:0000655)
- SBO:0000660 stands for **'antiporter-mediated transport'** (grandchild term of SBO:0000655)
- SBO:0000672 stands for **'spontaneous reaction'**

```
sbo = unique(iPrub22_reconstruction.rxnSBOTerms);
for i = 1:length(sbo)
  fprintf('%d reactions are annotaded with <strong>%s</strong>\n', sum(cellfun(@(x),...
    isequal(x,sbo{i}(:,:)), iPrub22_reconstruction.rxnSBOTerms)), sbo{i}(:,:));
end
```

```
     3   reactions are annotaded with SBO:0000167
  5280   reactions are annotaded with SBO:0000176
    60   reactions are annotaded with SBO:0000185
   228   reactions are annotaded with SBO:0000627
    37   reactions are annotaded with SBO:0000628
     1   reactions are annotaded with SBO:0000629
     1   reactions are annotaded with SBO:0000630
     1   reactions are annotaded with SBO:0000632
    53   reactions are annotaded with SBO:0000655
    38   reactions are annotaded with SBO:0000657
   108   reactions are annotaded with SBO:0000658
    21   reactions are annotaded with SBO:0000659
     2   reactions are annotaded with SBO:0000660
    86   reactions are annotaded with SBO:0000672
```

~~~~~~ [...] ~~~~~~

1.4 Genes

The gene identifiers for *P. chrysogenum* Wisconsin 54-1255 are in the form of **Pc\d{2}g\d{5}** where the first number is the contig number and the second is the gene number. Of the 6,171 genes referenced in the reconstruction, 468 are artificial. They have been added to the reconstruction to differentiate them from gap-filling reactions and to target specific reaction types :

- spontaneous (ID format: **s\d{3}**)
- transport (ID format: **t\d{3}**)
- demand (ID format: **d\d{3}**)
- sink (ID format: **sk\d{3}**)
- uptake (ID format: **u\d{3}**)
- production (ID format: **p\d{3}**)

```
fprintf(['<strong>Total "Genes" number in the reconstruction:</strong> %d\n\n',...
  '  <strong>Actual genes number:</strong> %d\n',...
  '  <strong>Artificial genes number:</strong> %d\n\n',...
  '    Spontaneous reactions: %d\n    Transport reactions: %d\n',...
  '    Demand reactions: %d\n    Sink reactions: %d\n    Uptake reactions: %d\n',...
  '    Production reactions: %d\n'],...
  length(iPrub22_reconstruction.genes) ,...
  sum(count(iPrub22_reconstruction.geneNames,"Pc")),...
  length(regexp(strcat(iPrub22_reconstruction.geneNames{:}),'s|t|d|u|p')),...
  sum(count(iPrub22_reconstruction.geneNames,"s"))-
                                    sum(count(iPrub22_reconstruction.geneNames,"sk")),...
  sum(count(iPrub22_reconstruction.geneNames,"t")),...
  sum(count(iPrub22_reconstruction.geneNames,"d")),...
  sum(count(iPrub22_reconstruction.geneNames,"sk")),...
  sum(count(iPrub22_reconstruction.geneNames,"u")),...
  sum(count(iPrub22_reconstruction.geneNames,"p")))
```

Total "Genes" number in the reconstruction: 6171

Actual genes number: 5703
Artificial genes number: 468

Spontaneous reactions: 86
Transport reactions: 117
Demand reactions: 37
Sink reactions: 1
Uptake reactions: 185
Production reactions: 42

1.4.1 Name(▲)

```
%Genes names
Genes_Types = ["Penicillium_chrysogenum_genes", "Spontaneous_genes", "Transport_genes",
  "Demand_genes", "Sink_genes", "Uptake_genes", "Production_genes"] ;
Pc=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                        strfind(iPrub22_reconstruction.geneNames,'Pc'))));
S=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                         strfind(iPrub22_reconstruction.geneNames,'s'))));
T=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                        (strfind(iPrub22_reconstruction.geneNames,'t'))));
D=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                        (strfind(iPrub22_reconstruction.geneNames,'d'))));
SK=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                       (strfind(iPrub22_reconstruction.geneNames,'sk'))));
SK = [SK(1);{'      '};{'      '}] ;% Only one sink recation in iPrub22
```

```
U=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                    (strfind(iPrub22_reconstruction.geneNames,'u'))));
P=iPrub22_reconstruction.geneNames(~cellfun('isempty',
                                    (strfind(iPrub22_reconstruction.geneNames,'p'))));
```

Penicillium_chrysogenum_genes	Spontaneous_genes	Transport_genes
'Pc13g04040'	's007'	't001'
'Pc16g02170'	's001'	't002'
'Pc16g09860'	's008'	't003'

Demand_genes	Sink_genes	Uptake_genes	Production_genes
'd001'	'sk001'	'u001'	'p001'
'd002'	' '	'u002'	'p002'
'd003'	' '	'u003'	'p003'

~~~~~~ [...] ~~~~~~

1.4.4 Gene products compartmentation

The proposed reconstruction is not compartmentalised at the intracellular level. Nevertheless, an annotation related to the detection of signal peptides (**SignalP - v4.1g**), transmembrane domains (**TMHMM - v2.0c**), and the prediction of the subcellular localisation of proteins (**DeepLoc - v1.0**) has been made. As access to this information is not supported by CobraToolbox, it is included in the notes tag.

```
<fbc:geneProduct sboTerm="SBO:0000243" metaid="gp_Pc21g12170" fbc:id="gp_Pc21g12170" fbc:label="Pc21g12170">
  <notes>
    <body xmlns="http://www.w3.org/1999/xhtml">
      <p>DeepLoc: Mitochondrion</p>
      <p>TMHMM: Topology=o</p>
      <p>SignalP: 0,497</p>
    </body>
  </notes>
  <annotation>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
      <rdf:Description rdf:about="#gp_Pc21g12170">
        <bqbiol:is>
          <rdf:Bag>
            <rdf:li rdf:resource="https://identifiers.org/ncbigene/8307889"/>
            <rdf:li rdf:resource="https://identifiers.org/ncbiprotein/XP_002568248"/>
            <rdf:li rdf:resource="https://identifiers.org/kegg.genes/pcs:Pc21g12170"/>
            <rdf:li rdf:resource="https://identifiers.org/uniprot/B6HLA3"/>
            <rdf:li rdf:resource="https://identifiers.org/refseq/XP_002568248"/>
          </rdf:Bag>
        </bqbiol:is>
      </rdf:Description>
    </rdf:RDF>
  </annotation>
</fbc:geneProduct>
```
~~~~~~

2. Model

From a reconstruction viewpoint, obtaining a consistent flux model requires some adjustments (*e.g.* reversibility, redundancy and non-reducibility being common issues). Therefore, the model presented below represents **a specification of the reconstruction**. We have chosen **not to suppress any reactions**, so to remedy the problems commonly encountered, we have **set up our model to block them**. Section 2.1 of this document summarises the actions taken (*i.e.* modifications have been directly incorporated into *i*Prub22; this document serves as a comprehensive reference for tracking and documenting these changes), and section 2.2 outlines some characteristics of the resulting model.

```
% Environnement definition
solverName = 'gurobi' ;
solverType = 'LP' ;
changeCobraSolver(solverName,solverType) ;
```

```
> changeCobraSolver: Gurobi interface added to MATLAB path.
```

```
iPrub22_model = iPrub22_reconstruction ;
%sparse format accelerates computations with large networks
iPrub22_model.S = sparse(iPrub22_model.S) ;
```

2.1 Reconstruction modifications: from reconstruction to model

SBOTerm - model

- SBO:0000624 stands for '**flux balance framework**': "Modelling approach, typically used for metabolic models, where the flow of metabolites (flux) through a network can be calculated. This approach will generally produce a set of solutions (solution space), which may be reduced using objective functions and constraints on individual fluxes".

```
<sbml xmlns="http://www.sbml.org/sbml/level3/version1/core"
 xmlns:fbc="http://www.sbml.org/sbml/level3/version1/fbc/version2"
level="3" version="1" fbc:required="false">
```

2.1.1 List of parameters

For improved usability, we have added human-readable labels to identify the boundaries of the relevant reactions. We hope these labels enhance the user experience by providing easily understandable information about the scope and context of the reactions of interest.

SBOTerm - listOfParameters

- SBO:0000626 stands for '**default flux bound**'
- SBO:0000625 stands for '**flux bound**'

```
<listOfParameters>
  <parameter sboTerm="SBO:0000626" id="default_lb" name="arbitrary lower bound - reversible reaction" units="mmol_per_gDW_per_hr" value="-1000" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="default_ub" name="arbitrary upper bound" units="mmol_per_gDW_per_hr" value="1000" constant="true"/>
  <parameter sboTerm="SBO:0000625" id="irrLow" name="arbitrary lower bound - irreversible reaction" units="mmol_per_gDW_per_hr" value="0" constant="true"/>
  <parameter sboTerm="SBO:0000625" id="NGAM_BOUND" name="lower and upper bounds for NGAM reaction" units="mmol_per_gDW_per_hr" value="1" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="Exchange_Biomass_lb" name="lower bounds for exchange biomass reaction" units="mmol_per_gDW_per_hr" value="0" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="Exchange_Biomass_ub" name="upper bounds for exchange biomass reaction" units="mmol_per_gDW_per_hr" value="1000" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="correction_reversibility_BOUND" name="Reversibility update from MetaCyc data" units="mmol_per_gDW_per_hr" value="0" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="blocked_BOUND" name="Reaction boundaries to be blocked (e.g. inconsistency, redudancy)" units="mmol_per_gDW_per_hr" value="0" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="unbalanced_reaction_BOUND" name="Unbalanced reaction - results of the checkMassChargeBalance function" units="mmol_per_gDW_per_hr" value="0" constant="true"/>
  <parameter sboTerm="SBO:0000626" id="duplicate_reaction_BOUND" name="Reaction closed due to redundancy " units="mmol_per_gDW_per_hr" value="0" constant="true"/>
  <parameter sboTerm="SBO:0000625" id="ub_uptake_001" name="Uptake of (S)-lactate" units="mmol_per_gDW_per_hr" value="0" constant="true"/>

  ...

  <parameter sboTerm="SBO:0000626" id="ub_production_032" name="Production of Meleagrin" units="mmol_per_gDW_per_hr" value="1000" constant="true"/>
</listOfParameters>
```

To facilitate the tracking of changes and their justifications, we have introduced four distinct labels within the list of parameters. In cases where a reaction had multiple labels, we have selected the most informative label to ensure clarity and transparency in documenting the modifications.

- `blocked_BOUND` - see **2.1.2 Errors during model loading & 2.1.4 Reconcilation and duplication**
- `correction_reversibility_BOUND` - see **2.1.3 Reversibility**
- `duplicate_reaction_BOUND` - **see 2.1.4 Reconcilation and duplication**
- `imbalanced_reaction_BOUND` - **see 2.1.5 Unbalanced reactions**

```
Parameter_Id = ['default_ub';'default_lb';"irrLow";'Exchange_Biomass_lb';...
  'Exchange_Biomass_ub';'NGAM_BOUND';'correction_reversibility_BOUND';'blocked_BOUND';...
  'imbalanced_reaction_BOUND';'duplicate_reaction_BOUND';'ub_uptake_...'; ...
  'ub_production_...'];
Value = [1000;-1000;0;0;1000;1;0;0;0;0;"depends on media modelling conditions";...
  "depends on media modelling conditions"];
Number_of_ub = [4358;0;0;0;1;1;180;111;961;80;185;42];
Number_of_lb = [0;618;3914;1;0;1;233;111;961;80;0;0];
disp(table(Parameter_Id,Value,Number_of_ub,Number_of_lb))
```

Parameter_Id	Value	Number_of_ub	Number_of_lb
"default_ub"	"1000"	4358	0
"default_lb"	"-1000"	0	618
"irrLow"	"0"	0	3914
"Exchange_Biomass_lb"	"0"	0	1
"Exchange_Biomass_ub"	"1000"	1	0
"NGAM_BOUND"	"1"	1	1
"correction_reversibility_BOUND"	"0"	180	233
"blocked_BOUND"	"0"	111	111
"imbalanced_reaction_BOUND"	"0"	961	961
"duplicate_reaction_BOUND"	"0"	80	80
"ub_uptake_..."	"depends on media modelling conditions"	185	0
"ub_production_..."	"depends on media modelling conditions"	42	0

~~~~~~ [...] ~~~~~~

2.1.5 Unbalanced reactions

```
[massImbalance, imBalancedMass, imBalancedCharge, imBalancedRxnBool, Elements,
  missingFormulaeBool, balancedMetBool] = checkMassChargeBalance(iPrub22_model);
fprintf(['Number of reactions: <strong>%d\n</strong>',...
  '    Reactions with mass-imbalance: <strong>%d (%.2f%%)</strong>\n',...
  '    Reactions with charge imbalance: <strong>%d (%.2f%%)</strong>\n',...
  '    Imbalance reactions (exchange reactions are included): <strong>%d
                                                    (%.2f%%)</strong>\n\n',...
  'Number of metabolites: <strong>%d</strong>\n',...
  '    Metabolites without formulae: <strong>%d (%.2f%%)</strong>\n',...
  '    Metabolites exclusively involved in balanced reactions: <strong>%d
                                                     (%.2f%%)</strong>\n'],...
  length(iPrub22_model.rxns),...
  length(imBalancedMass(~cellfun('isempty',imBalancedMass))),...
  length(imBalancedMass(~cellfun('isempty',imBalancedMass)))*100/
                                                  length(iPrub22_model.rxns),...
```
~~~~~~

```
  sum(imBalancedCharge ~= 0),sum(imBalancedCharge ~= 0)*100/length(iPrub22_model.rxns),...
  sum(imBalancedRxnBool),sum(imBalancedRxnBool)*100/length(iPrub22_model.rxns),...
  length(iPrub22_model.mets),...
  sum(missingFormulaeBool), ...
  sum(missingFormulaeBool)*100/length(iPrub22_model.mets), ...
  sum(balancedMetBool),...
  sum(balancedMetBool)*100/length(iPrub22_model.mets)) ;
```

Number of reactions: **5919**
Reactions with mass-imbalance: **1398 (23.62%)**
Reactions with charge imbalance: **969 (16.37%)**
Imbalance reactions (exchange reactions are included): **1402 (23.69%)**

Number of metabolites: **5464**
Metabolites without formulae: **192 (3.51%)**
Metabolites exclusively involved in balanced reactions: **3517 (64.37%)**

```
% Remove the exchange reactions from the set of unbalanced reactions
Uptake = strmatch('Uptake',iPrub22_model.rxns) ;
Production = strmatch('Production',iPrub22_model.rxns) ;
Sink = strmatch('Sink',iPrub22_model.rxns) ;
Exchange = [Uptake ; Production ; Sink ] ;
Imbalanced_reactions = setdiff(iPrub22_model.rxns(find(imBalancedRxnBool==1)),
                                                        iPrub22_model.rxns(Exchange));

% Remove from the set of unbalanced reactions the biomass and its assimilated reactions
Biomass_rxns_and_assimilate = {'Biomass_rxn';'Exchange_Biomass';'Transport_Biomass';
  'r1455';'r1456';'r1457';'r1458';'r1459';'r1460';'r1465'};
Imbalanced_reactions = setdiff(Imbalanced_reactions,Biomass_rxns_and_assimilate);

% Remove from the set of unbalanced reactions the set of reactions that have already been
% closed
Already_blocked = unique([Concerned_rxns ; internal_transport ; block_ub ; block_lb ;
                                          block_both ; duplicate_rxns ; redundant_rxns]) ;
Blocked_in_previous_curation = intersect(Imbalanced_reactions,Already_blocked ) ;
Imbalanced_reactions = setdiff(Imbalanced_reactions,Blocked_in_previous_curation);

% Remove from the set of unbalanced reactions the reactions involved in the biosynthesis of
% specialised metabolites (penicillin, roquefortin, isoepoxydon, toluquinol, ferrichrome,
% etc.)
Biosynthesis_specialised_metabolites = {'METHYLGLUTACONYL-COA-HYDRATASE-RXN';...
  'RXN-15470';'RXN-15472';'RXN-15473';'RXN-15480';'RXN-15950';'RXN-16125';'RXN-16128';...
  'RXN-16140';'RXN-16141';'RXN-16142';'RXN-16143';'RXN-16144';'RXN-16147'; ...
  'RXN-8809';'RXN-9479';'RXN-9481';'RXN-9484';'RXN-9486';'RXN-9487';'RXN-9489';...
  'RXN-9494';'RXN-9497'};
Imbalanced_reactions = setdiff(Imbalanced_reactions,Biosynthesis_specialised_metabolites);

% Remove from the set of unbalanced reactions those that prevent biomass production when
% all the uptakes are open
Retained_reaction = {'ADENYLOSUCCINATE-SYNTHASE-RXN';'AIRCARBOXY-RXN'; ...
  'AIRS-RXN';'ATPPHOSPHORIBOSYLTRANS-RXN';'DETHIOBIOTIN-SYN-RXN';'FGAMSYN-RXN'; ...
  'GART-RXN';'HISTPRATPHYD-RXN';'PGPPHOSPHA-RXN';'RXN-19329';'RXN-20084';'RXN-6081'; ...
  'RXN-8141';'RXN-9384';'RXN3O-130';'RXN3O-75';'SAICARSYN-RXN'; ...
  'THIOSULFATE--THIOL-SULFURTRANSFERASE-RXN'; 'r0479';'r0480';'r0491';'r0729';'r0995';...
  'r1125';'NGAM'} ;
Imbalanced_reactions = setdiff(Imbalanced_reactions,Retained_reaction);
```

```
% Sum up
fprintf(['Of the <strong>%d</strong> unbalanced reactions detected by the
                                                checkMassChargeBalance function:\n\n',...
  '  <strong>%d</strong> are exchange reactions\n',...
  '  <strong>%d</strong>  are the biomass and its assimilated reactions\n',...
  '  <strong>%d</strong> have been closed during previous curations\n',...
  '  <strong>%d</strong>  are required to produce specialised metabolites\n',..
  '  <strong>%d</strong>  are required to maintain biomass production\n',...
  '  <strong>%d</strong> are blocked and annotated with
                                    <strong>imbalanced_reaction_bound</strong> tag\n'],...
    length(iPrub22_model.rxns(find(imBalancedRxnBool==1))),...
    length(Exchange),length(Biomass_rxns_and_assimilate),...
    length(Blocked_in_previous_curation),length(Biosynthesis_specialised_metabolites),...
    length(Retained_reaction),length(Imbalanced_reactions)) ;
```

```
Of the 1402 unbalanced reactions detected by the checkMassChargeBalance function:

   228  are exchange reactions
    10  are the biomass and its assimilated reactions
   170  have been closed during previous curations
    23  are required to produce specialised metabolite
    25  are required to maintain biomass production
   946  are blocked and annotated with imbalanced_reaction_bound tag
```

NB. To improve traceability, we preferred to keep mainly the most informative tags for previously closed reactions: `correction_reversibility_BOUND`, `blocked_BOUND`, or `duplicate_reaction_BOUND`. It ensures a more comprehensive and informative representation of the modifications made to these reactions. Thus, of the 170 reactions blocked at previous stages, only 15 had their label changed to `imbalanced_reaction_bound`.

~~~~~~ [...] ~~~~~~

→ Overall, **a total of 1,607 reactions**, which accounts for approximately **27% of the network**, have undergone **modifications to at least one of their bounds**.

2.2 Model Characteristics

~~~~~~ [...] ~~~~~~

2.2.2 Identify metabolic dead-ends

The `detectDeadEnds` function identifies dead-end metabolites that cannot be produced or consumed by any other reaction in the metabolic network. It checks the stoichiometric coefficients to determine if a metabolite is solely produced or solely consumed in a reaction. The function also detects metabolites involved in only one reaction, indicating a lack of support from other reactions in the network.

```
% Detection of dead-end metabolites:
DeadEnds_close = closed_model.mets(detectDeadEnds(closed_model));
DeadEnds_default = default_model.mets(detectDeadEnds(default_model));
DeadEnds_open = open_model.mets(detectDeadEnds(open_model));

% Identification of associated reactions
[rxnList_close, ~] = findRxnsFromMets(iPrub22_model, DeadEnds_close);
[rxnList_default, ~] = findRxnsFromMets(iPrub22_model, DeadEnds_default);
[rxnList_open, ~] = findRxnsFromMets(iPrub22_model, DeadEnds_open);
```

```
fprintf(['<strong>Number of dead-ends detected:</strong>\n   Closed model: %d (%.1f%%)\n
                            Default model: %d (%.1f%%)\n   Open model: %d (%.1f%%)\n\n',...
  '<strong>Number of associated reactions:</strong>\n   Closed model: %d (%.1f%%)\n
                           Default model: %d (%.1f%%)\n   Open model: %d (%.1f%%)\n\n'],...
  length(DeadEnds_close), length(DeadEnds_close)*100/length(iPrub22_model.mets),...
  length(DeadEnds_default), length(DeadEnds_default)*100/length(iPrub22_model.mets),...
  length(DeadEnds_open), length(DeadEnds_open)*100/length(iPrub22_model.mets),...
  length(rxnList_close), length(rxnList_close)*100/length(iPrub22_model.rxns),...
  length(rxnList_default), length(rxnList_default)*100/length(iPrub22_model.rxns),...
  length(rxnList_open), length(rxnList_open)*100/length(iPrub22_model.rxns))
```

Number of dead-ends detected:
Closed model: 3411 (62.4%)
Default model: 3379 (61.8%)
Open model: 3336 (61.1%)

Number of associated reactions:
Closed model: 3173 (53.6%)
Default model: 3109 (52.5%)
Open model: 3023 (51.1%)

2.2.3 Identify blocked reactions

The `identifyBlockedRxns` function detects reactions that are unable to generate flux within the constraints of the model. Blocked reactions provide information about the feasibility of the model. They highlight potential errors, connectivity issues, or gaps in our understanding of metabolism, as they cannot contribute to metabolic fluxes.

```
tic
BlockedReactions_default = findBlockedReaction(default_model);
BlockedReactions_open = findBlockedReaction(open_model);
toc
```

```
Elapsed time is 152.888550 seconds.
```

```
fprintf(['<strong>Number of blocked reactions:</strong>\n',...
  '   Default model: %d (percentage of active reactions: %.1f%%)\n',...
  '   Open model: %d (percentage of active reactions: %.1f%%)\n'],...
  length(BlockedReactions_default),...
  100-length(BlockedReactions_default)*100/length(iPrub22_model.rxns),...
  length(BlockedReactions_open),...
  100-length(BlockedReactions_open)*100/length(iPrub22_model.rxns))
```

Number of blocked reactions:

```
Default model: 4483 (percentage of active reactions: 24.3%)
Open model: 4252 (percentage of active reactions: 28.2%)
```

~~~~~~ [...] ~~~~~~
~~~~~~

2.3 Flux simulation

Minimum constraints for biomass production

1. Uptake_015: Alpha-glucose (Carbon source)
2. Uptake_062: FE2+
3. Uptake_130: NH3 (Nitrogen source)
4. Uptake_136: Oxygen-molecule or Uptake_077: Hydrogen-peroxide
5. Uptake_146: Phosphate
6. Uptake_157: Riboflavin or Uptake_065: FMN
7. Uptake_169: Elemental-Sulfur
8. Uptake_171: Thiamine

Constraints for specialised metabolites production

- Production of Isoepoxydon, Roquefortins and assimilates, Averufin, Versicolorin B and Versiconol

  Uptake_155: Red-NADPH-Hemoprotein-Reductases

- Production of Penicillin K

  Uptake_044: Decanoate or Uptake_137: Octanoate

- Production of Peniciilin V

  Uptake_144: Phenoxyacetate

- Production of Sterigmatocystin and Aflatoxin B1

  Uptake_008: 6-demethylsterigmatocystin

  Uptake_155: Red-NADPH-Hemoprotein-Reductases

2.3.1 Default model

- **Check the objective function and identify the constraints**

```
checkObjective(iPrub22_model) ;
```

summaryT = 12×5 table

	Coefficient	Metabolite	metID	Reaction	RxnID
1	-104.0000	WATER[c]	6	Biomass_rxn	810
2	-104.0000	ATP[c]	43	Biomass_rxn	810
3	104.0000	ADP[c]	50	Biomass_rxn	810
4	104.0000	Pi[c]	167	Biomass_rxn	810
5	-0.0001	COF[c]	2325	Biomass_rxn	810
6	1.0000	Biomass[c]	3182	Biomass_rxn	810
7	-0.4500	PROTEIN[c]	3572	Biomass_rxn	810
8	-0.0800	RNA[c]	3576	Biomass_rxn	810
9	-0.0100	DNA[c]	3601	Biomass_rxn	810
10	-0.0400	AAPOOL[c]	3645	Biomass_rxn	810

```
printConstraints(iPrub22_model,-1000,1000)
```

```
MinConstraints:
NGAM          1

maxConstraints:
Uptake_008   10
Uptake_015   15
Uptake_044   10
Uptake_062   10
Uptake_130    5
Uptake_144   10
Uptake_146   10
Uptake_155   10
Uptake_157   10
Uptake_169   10
Uptake_171   10
NGAM          1
```

- **Flux Balance Analysis result**

```
FBAsolution_iPrub22 = optimizeCbModel(iPrub22_model,'max') ;
fprintf('Solution: %d\nStatut: %d', FBAsolution_iPrub22.f, FBAsolution_iPrub22.stat)
```

```
Solution: 2.944472e-01
Statut: 1
```

- **List of monitored metabolites**

Penicillins Biosynthesis

~~~~~~ [...] ~~~~~~

```
Isopenicillin_N = findMetIDs(iPrub22_model,'ISOPENICILLIN-N[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
            <strong>%s</strong>\n  Exchange reaction: Production_010\n  Equation: %s\n',...
  iPrub22_reconstruction.metNames{Isopenicillin_N},...
  iPrub22_reconstruction.metFormulas{Isopenicillin_N},...
  iPrub22_reconstruction.metCharges(Isopenicillin_N),...
  iPrub22_reconstruction.mets{Isopenicillin_N},...
  formulas{findRxnIDs(iPrub22_model,'Production_010')})
```

```
Metabolite name: isopenicillin N
  Formulae: C14H20N3O6S
  Charge: -1
Metabolite ID: ISOPENICILLIN-N[c]
  Exchange reaction: Production_010
  Equation: ISOPENICILLIN-N[e] ->
```

```
Penicillin_K = findMetIDs(iPrub22_model,'CPD-9122[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
            <strong>%s</strong>\n  Exchange reaction: Production_012\n  Equation: %s\n',...
  iPrub22_reconstruction.metNames{Penicillin_K},...
  iPrub22_reconstruction.metFormulas{Penicillin_K},...
  iPrub22_reconstruction.metCharges(Penicillin_K),...
  iPrub22_reconstruction.mets{Penicillin_K},...
  formulas{findRxnIDs(iPrub22_model,'Production_012')})
```

```
Metabolite name: penicillin K
  Formulae: C16H26N2O4S
  Charge: 0
Metabolite  ID:  CPD-9122[c]
  Exchange reaction: Production_012
  Equation: CPD-9122[e] ->
```


~~~~~~

```
Penicillin_V = findMetIDs(iPrub22_model,'CPD-9196[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
            <strong>%s</strong>\n  Exchange reaction: Production_013\n  Equation: %s\n',...
  iPrub22_reconstruction.metNames{Penicillin_V},...
  iPrub22_reconstruction.metFormulas{Penicillin_V},...
  iPrub22_reconstruction.metCharges(Penicillin_V),...
  iPrub22_reconstruction.mets{Penicillin_V},...
  formulas{findRxnIDs(iPrub22_model,'Production_013')})
```

```
Metabolite name: penicillin V
  Formulae: C16H18N2O5S
  Charge: 0
Metabolite  ID: CPD-9196[c]
  Exchange reaction: Production_013
  Equation: CPD-9196[e] ->
```

~~~~~~ [...] ~~~~~~

Roquefortines Biosynthesis

~~~~~~ [...] ~~~~~~

```
Roquefortine_C = findMetIDs(iPrub22_model,'CPD-17381[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
            <strong>%s</strong>\n  Exchange reaction: Production_025\n  Equation: %s\n',...
  iPrub22_reconstruction.metNames{Roquefortine_C},...
  iPrub22_reconstruction.metFormulas{Roquefortine_C},...
  iPrub22_reconstruction.metCharges(Roquefortine_C),...
  iPrub22_reconstruction.mets{Roquefortine_C},...
  formulas{findRxnIDs(iPrub22_model,'Production_025')})
```

```
Metabolite name: roquefortine C
Formulae: C22H23N5O2
  Charge: 0
Metabolite ID: CPD-17381[c]
  Exchange reaction: Production_025
  Equation: CPD-17381[e] ->
```

~~~~~~ [...] ~~~~~~

```
Meleagrin = findMetIDs(iPrub22_model,'CPD-17391[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
            <strong>%s</strong>\n  Exchange reaction: Production_032\n  Equation: %s\n',...
  iPrub22_reconstruction.metNames{Meleagrin},...
  iPrub22_reconstruction.metFormulas{Meleagrin},...
  iPrub22_reconstruction.metCharges(Meleagrin),...
  iPrub22_reconstruction.mets{Meleagrin},...
  formulas{findRxnIDs(iPrub22_model,'Production_032')})
```

```
Metabolite name: meleagrin
  Formulae: C23H23N5O4
  Charge: 0
Metabolite ID: CPD-17391[c]
  Exchange reaction: Production_032
  Equation: CPD-17391[e] ->
```

~~~~~~ [...] ~~~~~~

Isoepoxydon biosynthesis and toluquinol precursor

~~~~~~ [...] ~~~~~~

```
Isoepoxydon = findMetIDs(iPrub22_model,'CPD-16723[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
            <strong>%s</strong>\n  Exchange reaction: Production_020\n  Equation: %s\n',...
  iPrub22_reconstruction.metNames{Isoepoxydon},...
  iPrub22_reconstruction.metFormulas{Isoepoxydon},...
  iPrub22_reconstruction.metCharges(Isoepoxydon),...
  iPrub22_reconstruction.mets{Isoepoxydon},...
  formulas{findRxnIDs(iPrub22_model,'Production_020')})
```

```
Metabolite name: isoepoxydon
  Formulae: C7H8O4
  Charge: 0
Metabolite ID: CPD-16723[c]
  Exchange reaction: Production_020
  Equation: CPD-16723[e] ->
```

```
Toluquinol = findMetIDs(iPrub22_model,'CPD-16720[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
           <strong>%s</strong>\n  Exchange reaction: Production_021\n  Equation: %s\n', ...
  iPrub22_reconstruction.metNames{Toluquinol},...
  iPrub22_reconstruction.metFormulas{Toluquinol},...
  iPrub22_reconstruction.metCharges(Toluquinol),... ...
  iPrub22_reconstruction.mets{Toluquinol},...
  formulas{findRxnIDs(iPrub22_model,'Production_021')})
```

```
Metabolite name: toluquinol
  Formulae: C7H8O2
  Charge: 0
Metabolite ID: CPD-16720[c]
  Exchange reaction: Production_021
  Equation: CPD-16720[e] ->
```

Precursor of DHN-melanin

```
YWA1 = findMetIDs(iPrub22_model,'CPD-18255[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
           <strong>%s</strong>\n  Exchange reaction: Production_033\n  Equation: %s\n', ...
  iPrub22_reconstruction.metNames{YWA1},...
  iPrub22_reconstruction.metFormulas{YWA1},...
  iPrub22_reconstruction.metCharges(YWA1), ...
  iPrub22_reconstruction.mets{YWA1},...
  formulas{findRxnIDs(iPrub22_model,'Production_033')})
```

```
Metabolite name: YWA1
  Formulae: C14H12O6
  Charge: 0
Metabolite ID: CPD-18255[c]
  Exchange reaction: Production_033
  Equation: CPD-18255[e] ->
```
~~~~~~

Ferrichrome biosynthesis

```
Ferrichrome = findMetIDs(iPrub22_model,'CPD0-2241[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
          <strong>%s</strong>\n  Exchange reaction: Production_034\n  Equation: %s\n', ...
  iPrub22_reconstruction.metNames{Ferrichrome},...
  iPrub22_reconstruction.metFormulas{Ferrichrome},...
  iPrub22_reconstruction.metCharges(Ferrichrome),...
  iPrub22_reconstruction.mets{Ferrichrome},...
  formulas{findRxnIDs(iPrub22_model,'Production_034')})
```

```
Metabolite name: ferrichrome
  Formulae: C27H42FeN9O12
  Charge: 0
Metabolite ID: CPD0-2241[c]
  Exchange reaction: Production_034
  Equation: CPD0-2241[e] ->
```

~~~~~~ [...] ~~~~~~

Precursor of Pr-Toxin

```
Aristolochene = findMetIDs(iPrub22_model,'ARISTOLOCHENE[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
          <strong>%s</strong>\n  Exchange reaction: Production_036\n  Equation: %s\n', ...
  iPrub22_reconstruction.metNames{Aristolochene},...
  iPrub22_reconstruction.metFormulas{Aristolochene},...
  iPrub22_reconstruction.metCharges(Aristolochene),...
  iPrub22_reconstruction.mets{Aristolochene},...
  formulas{findRxnIDs(iPrub22_model,'Production_036')})
```

```
Metabolite name: aristolochene
  Formulae: C15H24
  Charge: 0
Metabolite ID: ARISTOLOCHENE[c]
  Exchange reaction: Production_036
  Equation: ARISTOLOCHENE[e] ->
```

Precursor of Andrastin A

```
DMOA = findMetIDs(iPrub22_model,'CPD-17316[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
          <strong>%s</strong>\n  Exchange reaction: Production_037\n  Equation: %s\n', ...
  iPrub22_reconstruction.metNames{DMOA},...
  iPrub22_reconstruction.metFormulas{DMOA},...
  iPrub22_reconstruction.metCharges(DMOA), ...
  iPrub22_reconstruction.mets{DMOA},...
  formulas{findRxnIDs(iPrub22_model,'Production_037')})
```

```
Metabolite name: 3,5-dimethylorsellinate
  Formulae: C10H11O4
  Charge: -1
Metabolite ID: CPD-17316[c]
  Exchange reaction: Production_037
  Equation: CPD-17316[e] ->
```


~~~~~~

Biosynthesis of Averufin/Versiconol/Versicolorin B/Sterigmatocystin/Aflatoxin B1

```
Averufin = findMetIDs(iPrub22_model,'CPD-10167[c]');
fprintf('Metabolite name: <strong>%s</strong>\n  Formulae: %s\n  Charge: %d\nMetabolite ID:
          <strong>%s</strong>\n  Exchange reaction: Production_038\n  Equation: %s\n', ...
  iPrub22_reconstruction.metNames{Averufin},...
  iPrub22_reconstruction.metFormulas{Averufin},
  iPrub22_reconstruction.metCharges(Averufin), ...
  iPrub22_reconstruction.mets{Averufin},...
  formulas{findRxnIDs(iPrub22_model,'Production_038')})
```

```
Metabolite name: (1'S,5'S)-averufin
Formulae: C20H16O7
  Charge: 0
Metabolite ID: CPD-10167[c]
  Exchange reaction: Production_038
  Equation: CPD-10167[e] ->
```

~~~~~~ [...] ~~~~~~

- **Flux variance analysis**

```
Production_specialised_metabolites = {'Production_001';'Production_004';'Production_010';
   'Production_012';'Production_013';'Production_015';... %Penicillins
   'Production_016';'Production_020';'Production_021';... %Isoepoxydon/Toluquinol
   'Production_022';'Production_023';'Production_024';'Production_025';'Production_026';...
   'Production_027';'Production_028';'Production_029';'Production_030';...
   'Production_031';'Production_032';...%Roquefortines
   'Production_033';'Production_034';'Production_035';... %YWA1 & Ferrichrome
   'Production_036';'Production_037';... %Aristolochene & DMOA
   'Production_038';'Production_039';'Production_040';'Production_041';...
   'Production_042'} ; %Averufin/Versiconol/Versicolrin B/Sterigmatocystin/Aflatoxin B1
for i = 1 : length(Production_specialised_metabolites)
  [min, max, ~, ~] = fluxVariability(iPrub22_model, 80, 'max',
                                                  Production_specialised_metabolites(i));
  fprintf('Production of <strong>%s</strong>: minimum flux: %d - ',...
    'Maximum flux <strong>%d</strong>\n',...
    strjoin(iPrub22_model.metNames(findMetIDs(iPrub22_model,
            (findMetsFromRxns(iPrub22_model, Production_specialised_metabolites{i})))),...
    min, max);
end
```

```
Production of 6-aminopenicillanate: minimum flux: 0 - Maximum flux 1.675307e+00
Production of penicillin G: minimum flux: 0 - Maximum flux 1.167432e+00
Production of isopenicillin N: minimum flux: 0 - Maximum flux 1.116872e+00
Production of penicillin K: minimum flux: 0 - Maximum flux 1.675307e+00
Production of penicillin V: minimum flux: 0 - Maximum flux 1.675307e+00
Production of penicillin N: minimum flux: 0 - Maximum flux 1.116872e+00
Production of 3-methylphenol: minimum flux: 0 - Maximum flux 2.334864e+00
Production of isoepoxydon: minimum flux: 0 - Maximum flux 2.334864e+00
Production of toluquinol: minimum flux: 0 - Maximum flux 2.334864e+00
Production of histidyltryptophyldiketopiperazine: minimum flux: 0 - Maximum flux 6.701229e-01
Production of dehydrohistidyltryptophyldiketopiperazine: minimum flux: 0 - Maximum flux 6.701229e-01
Production of roquefortine D: minimum flux: 0 - Maximum flux 6.701229e-01
Production of roquefortine C: minimum flux: 0 - Maximum flux 6.701229e-01
Production of glandicoline A: minimum flux: 0 - Maximum flux 6.701229e-01
Production of N1-hydroxy-roquefortine C: minimum flux: 0 - Maximum flux 6.701229e-01
Production of roquefortine F: minimum flux: 0 - Maximum flux 6.701229e-01
Production of neoxaline: minimum flux: 0 - Maximum flux 6.701229e-01
Production of roquefortine L: minimum flux: 0 - Maximum flux 6.701229e-01
Production of glandicoline B: minimum flux: 0 - Maximum flux 6.701229e-01
Production of meleagrin: minimum flux: 0 - Maximum flux 6.701229e-01
Production of YWA1: minimum flux: 0 - Maximum flux 1.167432e+00
Production of ferrichrome: minimum flux: 0 - Maximum flux 3.722905e-01
Production of ferrichrome A: minimum flux: 0 - Maximum flux 3.722905e-01
Production of aristolochene: minimum flux: 0 - Maximum flux 1.089603e+00
Production of 3,5-dimethylorsellinate: minimum flux: 0 - Maximum flux 2.043006e+00
```
~~~~~~

2.3.2 Growth on different media

- **Unchanged constraints**

```
%% Unsubstitutable
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_062', 10 , 'u'); % FE2+
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_171', 10 , 'u'); % THIAMINE
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_146', 10 , 'u'); % Phosphate
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_169', 10 , 'u'); % Elemental-Sulfur
%% Sustitutable
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_157', 10 , 'u'); % RIBOFLAVIN
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_065', 0 , 'u'); % FMN
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_136', 1000 , 'u'); % OXYGEN-MOLECULE
% iPrub22_model=changeRxnBounds(iPrub22_model, 'Uptake_077', 0 , 'u') ; % Hydrogen-peroxide
```

- **Closure of uptakes not belonging to the minimum required for growth (linked to specialised metabolism)**

```
iPrub22_model = changeRxnBounds(iPrub22_model, 'Uptake_155', 0 , 'u');
iPrub22_model = changeRxnBounds(iPrub22_model, 'Uptake_044', 0 , 'u');
iPrub22_model = changeRxnBounds(iPrub22_model, 'Uptake_144', 0 , 'u');
iPrub22_model = changeRxnBounds(iPrub22_model, 'Uptake_008', 0 , 'u');
```

- **Modification of the carbon source**

```
% nitrogen source = NH3[0;5]
iProb22_model = changeRxnBounds(iProb22_model, 'Uptake_130', 5 , 'u') ;
% closing the alpha-glucose boundary
iProb22_model = changeRxnBounds(iProb22_model, 'Uptake_015', 0 , 'u') ;
% Definition of potential sources of carbon
Carbon_source = {'Uptake_001';'Uptake_005';'Uptake_014';'Uptake_015';'Uptake_016';...
  'Uptake_026';'Uptake_028';'Uptake_033';'Uptake_043';'Uptake_045';'Uptake_046';...
  'Uptake_048';'Uptake_049';'Uptake_054';'Uptake_060';'Uptake_072';'Uptake_085';...
  'Uptake_086';'Uptake_088';'Uptake_098';'Uptake_104';'Uptake_108';'Uptake_118';...
  'Uptake_119';'Uptake_122';'Uptake_123';'Uptake_149';'Uptake_151';'Uptake_152';...
  'Uptake_154'; 'Uptake_165';'Uptake_182'} ;
for i = 1 : length(Carbon_source)
  iProb22_model = changeRxnBounds(iProb22_model, Carbon_source(i), 15 , 'u');
  FBAsolution_iProb22 = optimizeCbModel(iProb22_model,'max');
  if FBAsolution_iProb22.f < 0.3 && FBAsolution_iProb22.f > 0.1
    fprintf('<strong>%s</strong> as Carbon source: Objective value: <strong>%d</strong> -
                                                                Solver statut: %d\n',...
      strjoin(iProb22_model.metNames(findMetIDs(iProb22_model,
                                findMetsFromRxns(iProb22_model, Carbon_source{i})))),...
      FBAsolution_iProb22.f, FBAsolution_iProb22.stat)
  else
      fprintf('<strong>%s</strong> as Carbon source: Objective value: %d - Solver statut:
                                                                        %d\n',...
      strjoin(iProb22_model.metNames(findMetIDs(iProb22_model,
                               (findMetsFromRxns(iProb22_model, Carbon_source{i})))),...
      FBAsolution_iProb22.f, FBAsolution_iProb22.stat)
  end
  iProb22_model = changeRxnBounds(iProb22_model, Carbon_source(i), 0 , 'u') ;
end
```

(S)-lactate as Carbon source: Objective value: 2.421299e-02 - Solver statut: 1
2-oxoglutarate as Carbon source: Objective value: 2.421299e-02 - Solver statut: 1
α,α-trehalose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
α-D-glucopyranose as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
aldehydo-L-arabinose as Carbon source: Objective value: **2.404004e-01** - Solver statut: 1
β-D-glucopyranose as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
butanoate as Carbon source: Objective value: 7.825983e-02 - Solver statut: 1
β-D-cellobiose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
D-arabinitol as Carbon source: Objective value: **2.404004e-01** - Solver statut: 1
β-D-fructofuranose as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
keto-D-fructose as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
α-D-galactopyranose as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
D-glucono-1,5-lactone as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
D-mannitol as Carbon source: Objective value: **2.944472e-01** - Solver statut: 1
ethanol as Carbon source: Objective value: 7.825983e-02 - Solver statut: 1
glycerol as Carbon source: Objective value: **1.323067e-01** - Solver statut: 1
α-lactose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
lactose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
L-arabinitol as Carbon source: Objective value: **2.404004e-01** - Solver statut: 1
L-homocysteine as Carbon source: Objective value: 2.428381e-02 - Solver statut: 1
L-methionine as Carbon source: Objective value: 7.848873e-02 - Solver statut: 1
L-ribulose as Carbon source: Objective value: **2.404004e-01** - Solver statut: 1
maltose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
maltotriose as Carbon source: Objective value: 7.140770e-01 - Solver statut: 1
melibiose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
methanol as Carbon source: Objective value: 2.421302e-02 - Solver statut: 1
propanoate as Carbon source: Objective value: 2.421299e-02 - Solver statut: 1
pyruvate as Carbon source: Objective value: 2.421299e-02 - Solver statut: 1
L-quinate as Carbon source: Objective value: **2.132570e-01** - Solver statut: 1
raffinose as Carbon source: Objective value: 7.140770e-01 - Solver statut: 1
sucrose as Carbon source: Objective value: 6.187283e-01 - Solver statut: 1
xylitol as Carbon source: Objective value: **2.404004e-01** - Solver statut: 1

- **Modification of the nitrogen source**

```
% carbon source = ALPHA-GLUCOSE[0;15]
iPrub22_model = changeRxnBounds(iPrub22_model, 'Uptake_015', 15 , 'u');
% closing the NH3 boundary
iPrub22_model = changeRxnBounds(iPrub22_model, 'Uptake_130', 0 , 'u');
% Definition of potential sources of nitrogen
Nitrogen_source = {'Uptake_012';'Uptake_013';'Uptake_019';'Uptake_039';'Uptake_042';...
  'Uptake_050';'Uptake_061';'Uptake_069';'Uptake_071';'Uptake_073';'Uptake_076';...
  'Uptake_080';'Uptake_084';'Uptake_087';'Uptake_089';'Uptake_090';'Uptake_091';...
  'Uptake_095';'Uptake_096';'Uptake_103';'Uptake_106';'Uptake_107';'Uptake_109';...
  'Uptake_111';'Uptake_113';'Uptake_130';'Uptake_133';'Uptake_176';'Uptake_180'};
for i = 1 : length(Nitrogen_source)
  iPrub22_model = changeRxnBounds(iPrub22_model, Nitrogen_source(i), 5 , 'u');
  FBAsolution_iPrub22 = optimizeCbModel(iPrub22_model,'max');
  if FBAsolution_iPrub22.f < 0.3 && FBAsolution_iPrub22.f > 0.1
    fprintf('<strong>%s</strong> as Nitrogen source: Objective value: <strong>%d</strong> -
                                                                Solver statut: %d\n',...
    strjoin(iPrub22_model.metNames(findMetIDs(iPrub22_model,
                                (findMetsFromRxns(iPrub22_model, Nitrogen_source{i})))),...
    FBAsolution_iPrub22.f, FBAsolution_iPrub22.stat)
  else
    fprintf('<strong>%s</strong> as Nitrogen source: Objective value: %d - Solver statut:
                                                                        %d\n',...
    strjoin(iPrub22_model.metNames(findMetIDs(iPrub22_model,
                                (findMetsFromRxns(iPrub22_model, Nitrogen_source{i})))),...
    FBAsolution_iPrub22.f, FBAsolution_iPrub22.stat)
  end
  iPrub22_model = changeRxnBounds(iPrub22_model, Nitrogen_source(i), 0 , 'u');
end
```

adenine as Nitrogen source: Objective value: 3.438235e-01 - Solver statut: 1
adenosine as Nitrogen source: Objective value: 4.350524e-01 - Solver statut: 1
ammonium as Nitrogen source: Objective value: 3.255777e-01 - Solver statut: 1
choline as Nitrogen source: Objective value: 4.168066e-01 - Solver statut: 1
cyanate as Nitrogen source: Objective value: **2.872013e-01** - Solver statut: 1
D-glucosamine as Nitrogen source: Objective value: 4.350524e-01 - Solver statut: 1
aci-nitroethane as Nitrogen source: Objective value: 3.255777e-01 - Solver statut: 1
4-aminobutanoate as Nitrogen source: Objective value: 3.620692e-01 - Solver statut: 1
glutathione as Nitrogen source: Objective value: 3.721080e-01 - Solver statut: 1
glycine as Nitrogen source: Objective value: 3.255777e-01 - Solver statut: 1
guanine as Nitrogen source: Objective value: 3.438235e-01 - Solver statut: 1
hypoxanthine as Nitrogen source: Objective value: 3.529463e-01 - Solver statut: 1
L-2-aminoadipate as Nitrogen source: Objective value: 3.620692e-01 - Solver statut: 1
L-alanine as Nitrogen source: Objective value: 3.438235e-01 - Solver statut: 1
L-arginine as Nitrogen source: Objective value: 4.076837e-01 - Solver statut: 1
L-asparagine as Nitrogen source: Objective value: 3.529463e-01 - Solver statut: 1
L-aspartate as Nitrogen source: Objective value: 3.255777e-01 - Solver statut: 1
L-glutamate as Nitrogen source: Objective value: 3.438235e-01 - Solver statut: 1
L-glutamine as Nitrogen source: Objective value: 3.711921e-01 - Solver statut: 1
L-lysine as Nitrogen source: Objective value: 4.259295e-01 - Solver statut: 1
L-phenylalanine as Nitrogen source: Objective value: 4.532981e-01 - Solver statut: 1
L-proline as Nitrogen source: Objective value: 3.803150e-01 - Solver statut: 1
L-serine as Nitrogen source: Objective value: 3.438235e-01 - Solver statut: 1
L-threonine as Nitrogen source: Objective value: 3.620692e-01 - Solver statut: 1
L-tyrosine as Nitrogen source: Objective value: 4.532981e-01 - Solver statut: 1
ammonia as Nitrogen source: Objective value: **2.944472e-01** - Solver statut: 1
nitrate as Nitrogen source: Objective value: **2.803035e-01** - Solver statut: 1
urea as Nitrogen source: Objective value: **2.982090e-01** - Solver statut: 1
xanthine as Nitrogen source: Objective value: 3.529463e-01 - Solver statut: 1

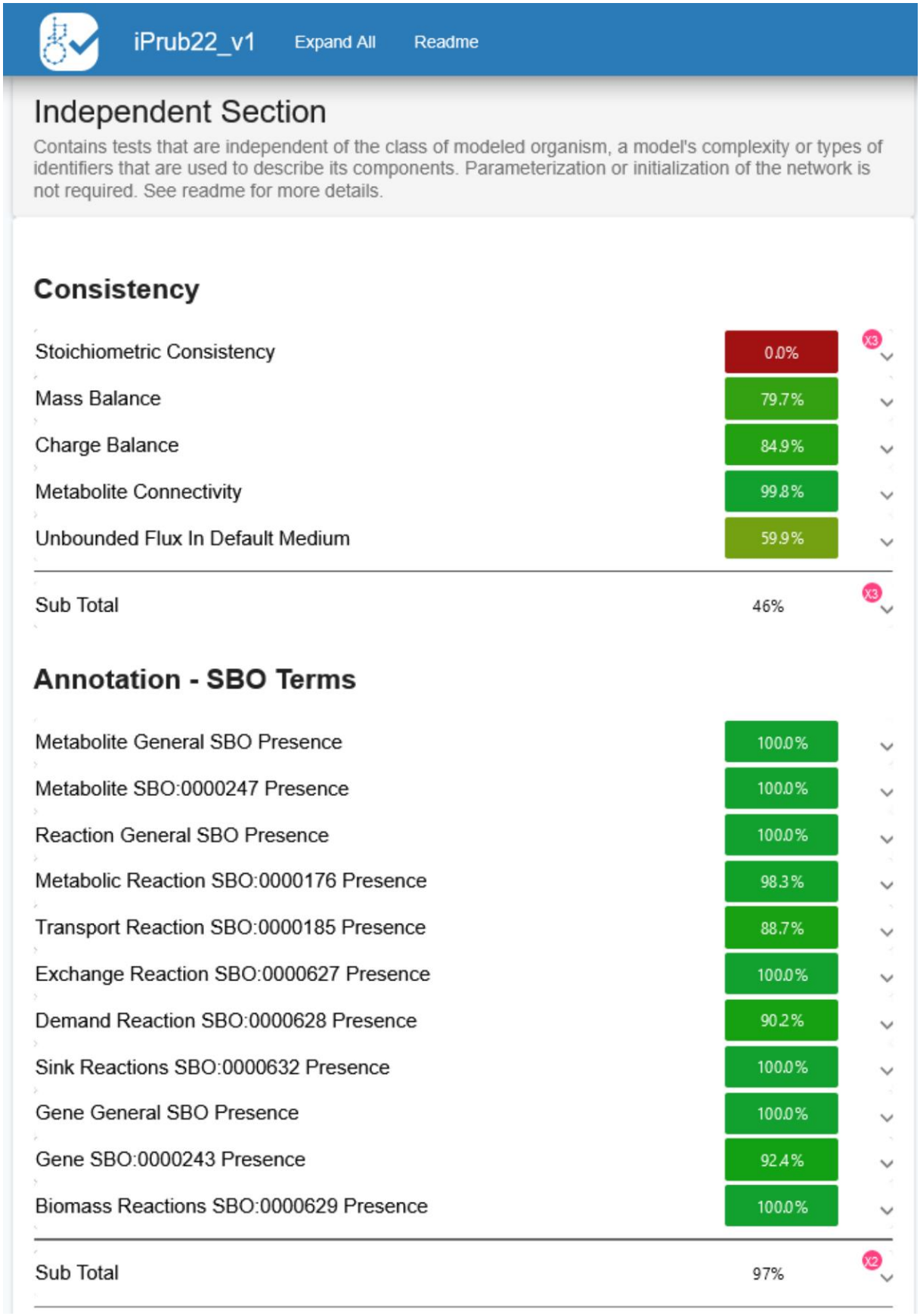
- **Amino acids as a source of carbon and nitrogen**

```
% closing carbon source boundary
iPrub22_model = changeRxnBounds(iPrub22_model, Carbon_source, 0 , 'u');
% closing nitrogen source boundary
iPrub22_model = changeRxnBounds(iPrub22_model, Nitrogen_source, 0 , 'u');
% List of amino acids as carbon and nitrogen sources
amino_acids = {'Uptake_073';'Uptake_087';'Uptake_089';'Uptake_090';'Uptake_091';...
  'Uptake_095';'Uptake_096';'Uptake_103';'Uptake_106';'Uptake_107';'Uptake_109';...
  'Uptake_111';'Uptake_113'};
for i = 1 : length(amino_acids)
  iPrub22_model = changeRxnBounds(iPrub22_model, amino_acids(i), 15 , 'u');
  FBAsolution_iPrub22 = optimizeCbModel(iPrub22_model,'max');
  if FBAsolution_iPrub22.f < 0.3 && FBAsolution_iPrub22.f > 0.1
    fprintf('<strong>%s</strong> as Carbon and Nitrogen source: Objective value:
                                          <strong>%d</strong> - Solver statut: %d\n',...
    strjoin(iPrub22_model.metNames(findMetIDs(iPrub22_model,
                                  findMetsFromRxns(iPrub22_model, amino_acids{i})))),...
    FBAsolution_iPrub22.f, FBAsolution_iPrub22.stat)
  else
    fprintf('<strong>%s</strong> as Carbon and Nitrogen source: Objective value: %d -
                                                        Solver statut: %d\n',...
    strjoin(iPrub22_model.metNames(findMetIDs(iPrub22_model,
                                 (findMetsFromRxns(iPrub22_model, amino_acids{i})))),...
    FBAsolution_iPrub22.f, FBAsolution_iPrub22.stat)
  end
    iPrub22_model = changeRxnBounds(iPrub22_model, amino_acids(i), 0 , 'u');
end
```

glycine as Carbon and Nitrogen source: Objective value: 5.189100e-02 - Solver statut: 1
L-alanine as Carbon and Nitrogen source: Objective value: **1.066283e-01** - Solver statut: 1
L-arginine as Carbon and Nitrogen source: Objective value: **2.982090e-01** - Solver statut: 1
L-asparagine as Carbon and Nitrogen source: Objective value: **1.339970e-01** - Solver statut: 1
L-aspartate as Carbon and Nitrogen source: Objective value: 5.189100e-02 - Solver statut: 1
L-glutamate as Carbon and Nitrogen source: Objective value: **1.066283e-01** - Solver statut: 1
L-glutamine as Carbon and Nitrogen source: Objective value: **1.887343e-01** - Solver statut: 1
L-lysine as Carbon and Nitrogen source: Objective value: 3.529463e-01 - Solver statut: 1
L-phenylalanine as Carbon and Nitrogen source: Objective value: 4.350524e-01 - Solver statut: 1
L-proline as Carbon and Nitrogen source: Objective value: **2.161030e-01** - Solver statut: 1
L-serine as Carbon and Nitrogen source: Objective value: **1.066283e-01** - Solver statut: 1
L-threonine as Carbon and Nitrogen source: Objective value: **1.613657e-01** - Solver statut: 1
L-tyrosine as Carbon and Nitrogen source: Objective value: 4.350524e-01 - Solver statut: 1

ANNEXE 5 :

Rapport MEMOTE


iPrub22_v1
Expand All
Readme
Independent Section
Contains tests that are independent of the class of modeled organism, a model's complexity or types of identifiers that are used to describe its components. Parameterization or initialization of the network is not required. See readme for more details.
Consistency
Stoichiometric Consistency
0.0%
Mass Balance
79.7%
Charge Balance
84.9%
Metabolite Connectivity
99.8%
Unbounded Flux In Default Medium
59.9%
Sub Total
46%
Annotation - SBO Terms
Metabolite General SBO Presence
100.0%
Metabolite SBO:0000247 Presence
100.0%
Reaction General SBO Presence
100.0%
Metabolic Reaction SBO:0000176 Presence
98.3%
Transport Reaction SBO:0000185 Presence
88.7%
Exchange Reaction SBO:0000627 Presence
100.0%
Demand Reaction SBO:0000628 Presence
90.2%
Sink Reactions SBO:0000632 Presence
100.0%
Gene General SBO Presence
100.0%
Gene SBO:0000243 Presence
92.4%
Biomass Reactions SBO:0000629 Presence
100.0%
Sub Total
97%

Annotation - Metabolites

Presence of Metabolite Annotation	100.0%
Metabolite Annotations Per Database	Info
pubchem.compound	67.6%
kegg.compound	43.0%
seed.compound	0.1%
inchikey	71.1%
inchi	71.1%
chebi	65.3%
hmdb	28.3%
reactome	0.0%
metanetx.chemical	99.5%
bigg.metabolite	13.1%
biocyc	99.6%
Metabolite Annotation Conformity Per Database	Info
pubchem.compound	100.0%
kegg.compound	100.0%
seed.compound	100.0%
inchikey	100.0%
inchi	100.0%
chebi	100.0%
hmdb	100.0%
reactome	100.0%
metanetx.chemical	100.0%
bigg.metabolite	100.0%
biocyc	100.0%
Uniform Metabolite Identifier Namespace	84.4%
Sub Total	84%

Annotation - Reactions

Presence of Reaction Annotation	100.0%
Reaction Annotations Per Database	Info
rhea	48.5%
kegg.reaction	41.7%
seed.reaction	22.5%
metanetx.reaction	85.1%
bigg.reaction	0.1%
reactome	0.0%
ec-code	73.4%
brenda	73.4%
biocyc	87.2%
Reaction Annotation Conformity Per Database	Info
rhea	100.0%
kegg.reaction	100.0%
seed.reaction	100.0%
metanetx.reaction	100.0%
bigg.reaction	100.0%
reactome	100.0%
ec-code	100.0%
brenda	100.0%
biocyc	100.0%
Uniform Reaction Identifier Namespace	87.2%
Sub Total	84%

Annotation - Genes

Presence of Gene Annotation	100.0%
Gene Annotations Per Database	Info
refseq	92.4%
uniprot	92.4%
ecogene	0.0%
kegg.genes	92.4%
ncbigi	0.0%
ncbigene	92.4%
ncbiprotein	92.4%
ccds	0.0%
hprd	0.0%
asap	0.0%
Gene Annotation Conformity Per Database	Info
refseq	100.0%
uniprot	100.0%
ecogene	0.0%
kegg.genes	100.0%
ncbigi	0.0%
ncbigene	100.0%
ncbiprotein	100.0%
ccds	0.0%
hprd	0.0%
asap	0.0%
Sub Total	65%

Total Score

74%

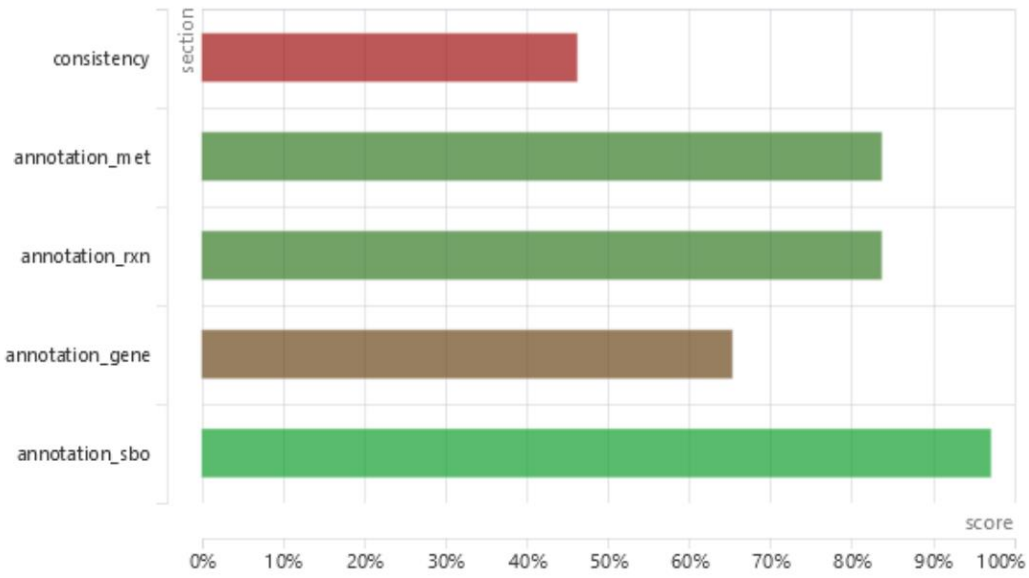
The Total Score is the result of the following calculation. For more information please click on "Readme" in the top left of the report.

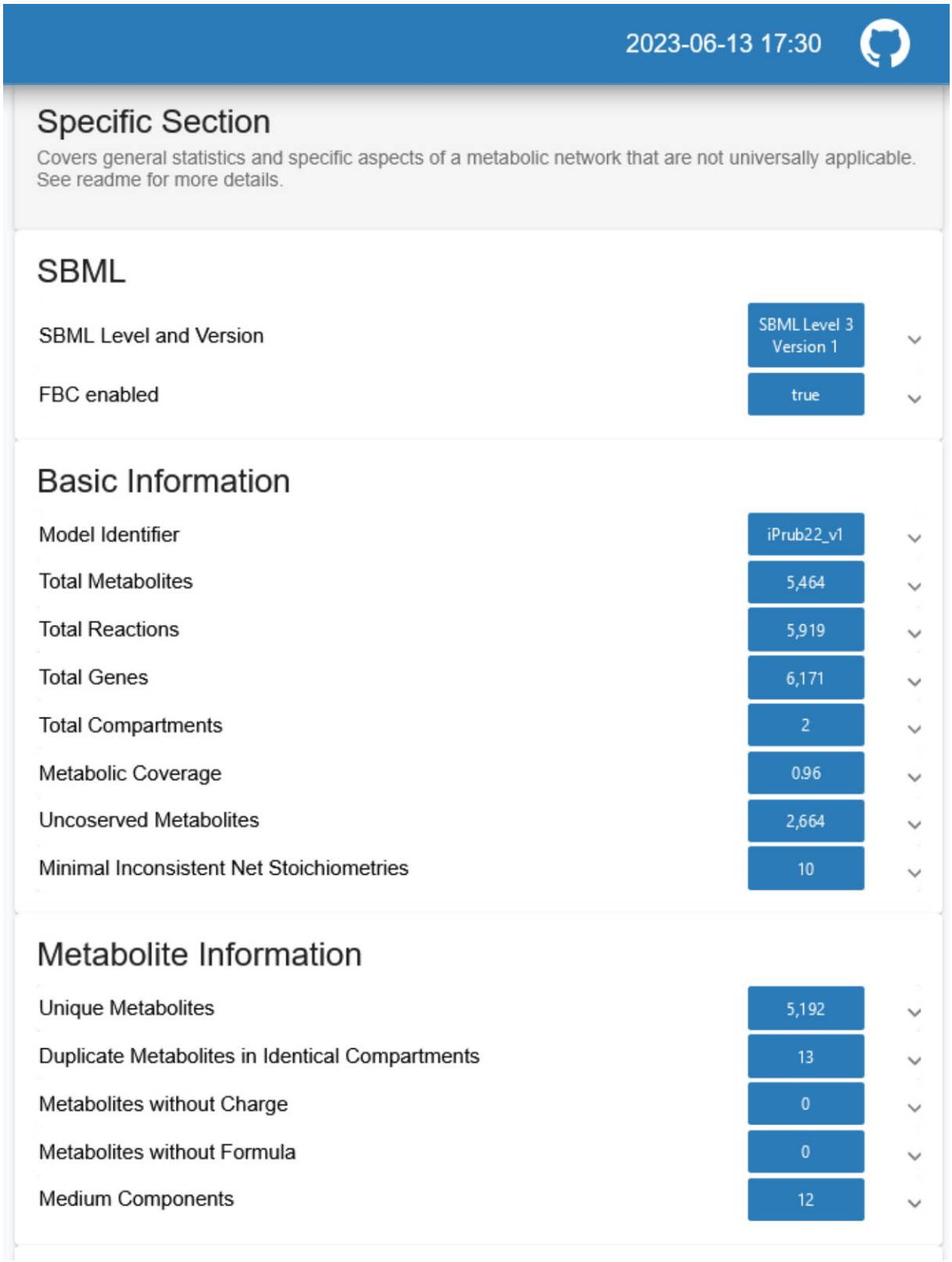
$$\frac{(3 \cdot 46.32) + (1 \cdot 83.80) + (1 \cdot 83.79) + (1 \cdot 65.40) + (2 \cdot 97.24)}{(3 \cdot 100) + (1 \cdot 100) + (1 \cdot 100) + (1 \cdot 100) + (2 \cdot 100)} = 73.68$$

Total Score

74%

Score per Category





2023-06-13 17:30

Specific Section

Covers general statistics and specific aspects of a metabolic network that are not universally applicable. See readme for more details.

SBML

SBML Level and Version	SBML Level 3 Version 1
FBC enabled	true

Basic Information

Model Identifier	iPrub22_v1
Total Metabolites	5,464
Total Reactions	5,919
Total Genes	6,171
Total Compartments	2
Metabolic Coverage	0.96
Uncoserved Metabolites	2,664
Minimal Inconsistent Net Stoichiometries	10

Metabolite Information

Unique Metabolites	5,192
Duplicate Metabolites in Identical Compartments	13
Metabolites without Charge	0
Metabolites without Formula	0
Medium Components	12

Reaction Information

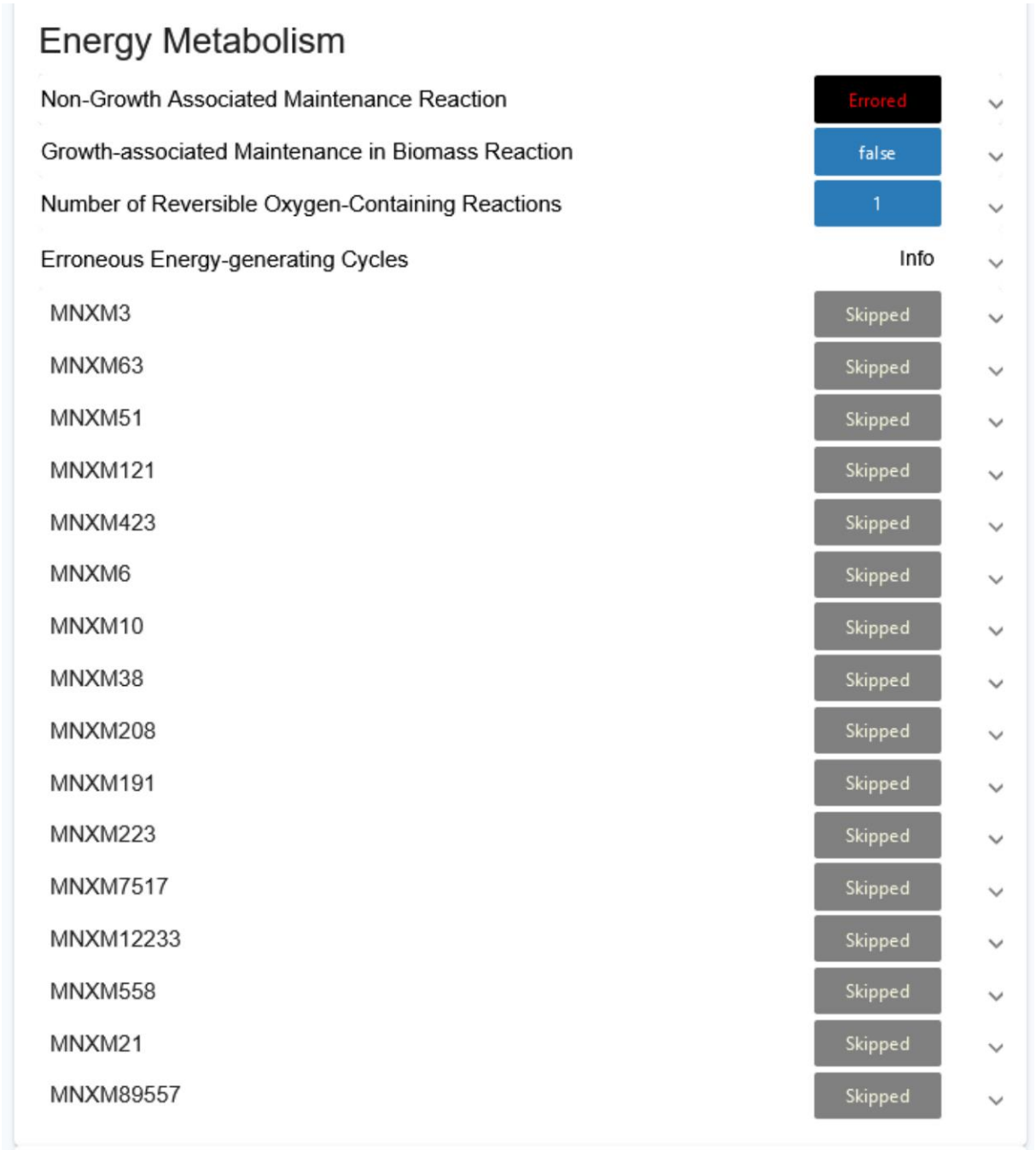
Purely Metabolic Reactions	5,367
Purely Metabolic Reactions with Constraints	1,302
Transport Reactions	318
Transport Reactions with Constraints	27
Reactions With Partially Identical Annotations	0.54
Duplicate Reactions	Errored
Reactions With Identical Genes	0.55

Gene-Protein-Reaction (GPR) Associations

Reactions without GPR	416
Fraction of Transport Reactions without GPR	0.00
Enzyme Complexes	1,549

Biomass

Biomass Reactions Identified	1
Biomass Consistency	0.10
Biomass Production In Default Medium	0.29
Unrealistic Growth Rate In Default Medium	false
Biomass Production In Complete Medium	90.71
Blocked Biomass Precursors In Default Medium	0
Blocked Biomass Precursors In Complete Medium	0
Ratio of Direct Metabolites in Biomass Reaction	0.00
Number of Missing Essential Biomass Precursors	2


Energy Metabolism
Non-Growth Associated Maintenance Reaction
Errored
Growth-associated Maintenance in Biomass Reaction
false
Number of Reversible Oxygen-Containing Reactions
1
Erroneous Energy-generating Cycles
Info
MNXM3
Skipped
MNXM63
Skipped
MNXM51
Skipped
MNXM121
Skipped
MNXM423
Skipped
MNXM6
Skipped
MNXM10
Skipped
MNXM38
Skipped
MNXM208
Skipped
MNXM191
Skipped
MNXM223
Skipped
MNXM7517
Skipped
MNXM12233
Skipped
MNXM558
Skipped
MNXM21
Skipped
MNXM89557
Skipped

Network Topology

Universally Blocked Reactions	4,226
Orphan Metabolites	855
Dead-end Metabolites	1,030
Stoichiometrically Balanced Cycles	828
Metabolite Production In Complete Medium	3,807
Metabolite Consumption In Complete Medium	4,259

Matrix Conditioning

Ratio Min/Max Non-Zero Coefficients	0.00
Independent Conservation Relations	997
Rank	4467
Degrees Of Freedom	1452

Experimental Data Comparison

Growth Prediction	Skipped
Gene Essentiality Prediction	Skipped

Misc. Tests

Environment

Python Version	3.7.10
Platform	Linux
Memote Version	0.13.0

Titre : Rationalisation de l'Accès aux Produits Naturels Fongiques par une Approche OSMAC *in silco*. Cas d'étude avec la modélisation du métabolisme de *Penicillium rubens.*

Mots clés : Réseau métabolique à l'échelle du génome (GSMN), métabolites spécialisés, traçabilité, reconstruction et simulations de flux

Résumé : Face à la résistance accrue aux antibiotiques menaçant la santé publique, la prospection de nouvelles molécules biologiquement actives est pressante. Les champignons filamenteux se distinguent par leur capacité à synthétiser une large gamme de produits naturels, sous l'influence de clusters de gènes biosynthétiques (*BGC*) qui orchestrent la production de métabolites spécialisés. Toutefois, de nombreux produits issus de ces BGCs n'ont pas encore été caractérisés et leur chimiodiversité demeure sous-explorée en raison de l'incapacité à activer l'ensemble de leur potentiel en laboratoire. L'approche OSMAC (*One Strain Many Compounds*) permet de solliciter ce potentiel en variant les conditions de culture. Cependant, cette méthode reste complexe et coûteuse en raison de son caractère aléatoire et du grand nombre d'expérimentations nécessaires. L'optimisation de ces processus nécessite l'intégration de stratégies plus rationnelles et efficaces. À l'aide d'approches systémiques liées à la biologie des systèmes, les réseaux métaboliques à l'échelle du génome (GSMN) offrent une modélisation détaillée des voies métaboliques, des enzymes impliquées et des gènes associés, fournissant un aperçu précis du métabolisme.
Dans ce cadre, nous proposons une stratégie alternative : l'OSMAC *in silico*. En reconstruisant un GSMN actualisé pour *Penicillium rubens*, nous avons pu étudier les réponses de son métabolisme sous divers scénarios nutritionnels. Cette modélisation a permis d'évaluer l'influence de différentes sources de carbone et d'azote sur sa croissance et la production de métabolites spécialisés, ouvrant ainsi de nouvelles perspectives pour optimiser la production de produits naturels.

Title: Rationalising Access to Fungal Natural Products through an *in silico* OSMAC Approach. A Case Study with the Metabolism Modelling of *Penicillium rubens.*

Keywords: Genome-Scale Metabolic Network (GSMN), specialised metabolites, traceability, reconstruction and flux simulations

Abstract: Given the pressing issue of increasing antibiotic resistance threatening public health, new biologically active molecule research is urgent. Filamentous fungi are characterised by their ability to synthesise a wide range of natural products, driven by biosynthetic gene clusters (BGCs) that orchestrate the production of specialised metabolites. However, many products derived from these BGCs remain uncharacterised, and their chemodiversity is underexplored due to the inability to activate their full potential in laboratory settings. The OSMAC (One Strain Many Compounds) approach seeks to harness this potential through culture condition variations. Nevertheless, this method remains complex and costly due to its randomness and vast number of experiments required. Therefore, optimising these processes needs the integration of more rational and efficient strategies. Using systems biology approaches, genome-scale metabolic networks (GSMNs) provide detailed modelling of metabolic pathways, involved enzymes, and associated genes, offering a precise overview of metabolism.
In this context, we propose an alternative strategy: *in silico* OSMAC. By reconstructing an updated GSMN for *Penicillium rubens*, we studied its metabolic responses under various nutritional scenarios. This modelling enabled us to assess the influence of different carbon and nitrogen sources on growth and the production of specialised metabolites, thereby opening new prospects for optimising the production of natural products.