Développement de nouvelles approches de biologie systémique pour l'évaluation de la toxicité de matières premières cosmétiques

Louison Fresnais

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Infectiologie, Physipoathologie, Toxicologie, Génétique et Nutrition

---

Présentée et soutenue par :

M. LOUISON FRESNAIS

le vendredi 15 décembre 2023

Titre :

Développement de nouvelles approches de biologie systémique pour l'évaluation de la toxicité de matières premières cosmétiques

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Toxicologie Alimentaire (ToxAlim)

Directeur(s) de Thèse :

M. FABIEN JOURDAN

Rapporteurs :

MME ANNE CORLU, CNRS

MME KARINE AUDOUZE, UNIVERSITE DE PARIS

M. VINCENT LACROIX, UNIVERSITE LYON 1

Membre(s) du jury :

M. DANIEL ZALKO, INRA TOULOUSE, Président

M. ANNE RIU, L'OREAL, Invité(e)

M. FABIEN JOURDAN, INRA TOULOUSE, Membre

MME NATHALIE POUPIN, INRA TOULOUSE, Invité(e)

M. OLIVIER PERIN, L'OREAL, Membre

Résumé

L'arrêt de l'expérimentation animale pour l'évaluation de la sécurité des ingrédients cosmétiques nécessite le développement d'approches alternatives pour l'évaluation de la sécurité. Historiquement, ces approches s'appuyaient sur des données structurales pour évaluer la sécurité des matières premières cosmétiques. Cependant, les nouvelles classes de matières premières cosmétiques étant principalement des molécules naturelles ou des mélanges complexes, l'obtention de données structurales pour l'évaluation de la sécurité est devenue un facteur limitant. Dans ce contexte, la génération de données omiques sur des systèmes *in vitro* et leur analyse semblent être des approches particulièrement pertinentes pour étudier les perturbations potentielles d'un large panel de processus biologiques. Parmi ces processus, le métabolisme est particulièrement concerné par les phénomènes de toxicité suite à l'exposition à différents xénobiotiques. En effet, la perturbation de voies telles que la lipogenèse *de novo*, la β-Oxydation ainsi que plus généralement des perturbations mitochondriales sont des phénomènes métaboliques connus pour engendrer de l'hépatotoxicité. Cependant, ces impacts ne sont généralement étudiés que de manière indirecte au travers d'analyses d'enrichissement sur les données transcriptomiques uniquement. L'étude de la réponse du métabolisme cellulaire à l'exposition à ces composés peut être réalisée grâce à des méthodes de modélisation globale du réseau métabolique à l'échelle du génome. Certaines de ces méthodes basées sur des approches de modélisation sous-contraintes, permettent de reconstruire un sous-réseau métabolique représentatif d'une condition biologique en intégrant des données transcriptomiques. Cependant, la complexité du réseau métabolique global face au nombre souvent plus limité de données expérimentales conduit à l'existence d'une multitude de sous-réseaux métaboliques possibles pour représenter une condition biologique, qui peuvent être obtenus par des approches d'énumération. L'analyse de cette très grande quantité de sous-réseaux ainsi que les coûts computationnels de ces approches d'énumération représentent encore à l'heure actuelle des défis qui nécessitent le développement de nouvelles méthodes d'analyse. Ces travaux de thèse ont pour objectif d'apporter une réponse à ces limites en proposant une stratégie permettant de modéliser le métabolisme cellulaire de cellules exposées à un xénobiotique à partir de données transcriptomiques par modélisation sous-contraintes. La modélisation sous-contraintes a été couplée à une approche d'énumération de solutions suivie d'une analyse mécanistique basée sur le calcul de fréquences d'activation et d'analyse de graphes métaboliques. La première partie de cette stratégie a été appliquée à 8 molécules connues pour leurs capacités à induire des phénomènes hépatotoxiques. Cette première étape a permis d'identifier des réactions différentiellement activées (DARs) pour chacune des molécules testées que nous avons pu replacer dans le contexte des voies métaboliques. Afin de proposer une interprétation plus précise de ces réactions potentiellement modulées, nous avons développé une analyse basée sur le calcul de distances métaboliques : le principe de cette analyse est d'identifier des ensembles de DARs proches dans le réseau et d'extraire les sous-réseaux minimaux correspondant à ces ensembles. Cette deuxième partie de la stratégie a été appliquée sur 2 des 8 molécules (acide valproïque et amiodarone), ce qui nous a permis de montrer que cette stratégie était capable de proposer des mécanismes d'action métaboliques pertinents et décrits dans la littérature. La stratégie développée au cours de ces travaux constitue un premier pas vers un outil simplifié d'analyse directe de l'impact des xénobiotiques sur le métabolisme cellulaire à destination des biologistes et toxicologues. Plusieurs perspectives d'évolution pour adapter la stratégie à une utilisation dans le cadre du read-across ont été également abordées.

Abstract

The discontinuation of animal testing for the safety assessment of cosmetic ingredients calls for the development of alternative approaches for safety evaluation. For a long time, these approaches were based on structural data. However, new classes of cosmetics raw materials are mostly natural compounds or complex mixtures, therefore obtaining structural data for safety evaluation has become a limiting factor. In this context, the generation of -omics data on *in vitro* systems and their analysis are relevant approaches for studying potential disruptions to a wide range of biological processes. Among these processes, metabolism is particularly concerned with the occurrence of toxicity following exposure to xenobiotics. Indeed, disruption of metabolic pathways such as *de novo* lipogenesis, β-oxidation, and more broadly mitochondrial disruptions are metabolic phenomenon known to induce hepatotoxicity. However, the impact of these compounds on cellular metabolism is not systematically studied, or only indirectly. The study of the response of cellular metabolism to exposure to these compounds can be carried out using modelling methods on genome-scale metabolic networks. Some of these methods, based on constraint-based modelling approaches, can reconstruct the metabolic sub-network that is representative of a biological condition by integrating transcriptomic data. However, the complexity of the global metabolic network compared to the often limited amount of experimental data leads to the existence of a many possible metabolic sub-networks representing a biological condition. Interestingly, enumeration approaches are able to find a set of subnetworks, on the order of thousands subnetworks, for each of the modelled conditions. The analysis of this very large number of subnetworks, as well as the computational costs of these enumeration approaches, still represent challenges that require the development of new analysis methods. The aim of this thesis work is to provide an answer to these limitations by proposing a strategy for modelling the cellular metabolism of cells exposed to a xenobiotic from transcriptomic data using constraint-based modelling with enumeration, coupled with mechanistic analysis based on the calculation of activation frequencies and metabolic graph analysis. The first part of this strategy was applied to 8 molecules known to induce hepatotoxicity. This first step allowed us to identify differentially activated reactions (DARs) for each of the tested molecules, on which we performed over-representation analyses. In order to propose a more precise interpretation of these potentially modulated reactions, we developed an analysis based on the calculation of metabolic distances. The principle of this analysis is to identify sets of DARs close in the network and extract the minimal sub-networks corresponding to these sets. The second part of the strategy was applied to 2 of the 8 molecules (valproic acid and amiodarone), enabling us to show that this strategy was capable of proposing relevant metabolic mechanisms of action described in the literature. In conclusion, the strategy developed by combining different metabolic modelling approaches helps to analyze in greater detail the metabolic effect induced by xenobiotics, and to propose hypotheses as to possible mechanisms of action, thus providing some answers to the limitations identified in the literature. Several perspectives will also be discussed to improve this strategy and add features enabling its use in the context of read-across studies

Remerciements

Je tiens tout d'abord à remercier Vincent Lacroix, Anne Corlu, Karine Audouze et Daniel Zalko pour avoir accepté d'être membres du jury et rapporteurs de cette thèse. Je tiens également à remercier Marie-France Sagot et German Caño-Sancho pour avoir accepté de faire partie de mon comité de suivi de thèse ainsi que pour les précieux conseils qu'ils m'ont délivré à l'occasion des deux comités de suivis de thèse ayant eu lieu au cours du projet.

J'adresse également un grand merci à Fabien Jourdan, Nathalie Poupin et Olivier Perin pour leur encadrement toujours bienveillant ainsi que leurs conseils toujours pertinents. Je leur serais toujours reconnaissant pour leur disponibilité ainsi que l'écoute dont ils ont fait preuve avec toujours le bon équilibre entre liberté d'entreprendre et recentrage vers l'objectif principal en fonction des besoins du projet. Je tiens également à les remercier pour les opportunités qu'ils m'ont apporté et l'intérêt qu'ils portent à l'avenir de chacun de leurs doctorants.

J'aimerai remercier Clément Frainay, Anne Riu, Romain Grall, Bathilde Ambroise, Alban Ott et Bernard Fromenty qui ont suivi le projet du début à la fin et m'ont apporté des conseils scientifiques dans des domaines variés mais qui ont été déterminant dans la qualité des travaux réalisés mais également pour mon développement en tant que scientifique.

Je tiens à remercier l'ensemble des membres du département Digital Innovation & Transformation pour leur accueil chaleureux et avec qui j'ai apprécié discuter et passer d'excellentes pauses dej' ! Je tiens tout particulièrement à remercier Dina pour son aide précieuse pour naviguer dans l'organisation administrative et technique L'Oréal, Diane, Ismaël, Ségolène et Rémy pour leur soutient et leurs encouragements tout au long de la thèse et particulièrement ces dernières semaines. J'aimerai également remercier l'équipe de la Chair de recherche de Québec et notamment Arnaud, Mickael, Marie-Pier, Charles, Alban et Antoine avec qui j'ai eu énormément de plaisir à discuter des projets de chacun tous les vendredi soir.

Je tiens également à remercier les membres du groupe MetExplore que je n'ai pas encore eu le plaisir de remercier : Jean-Clément (le JC), Marion, Ludovic, Florence, Juliette, Maximilian, Bénédict, Maxime (le D) et Pablo. Vous avez tous contribué à rendre ces trois années de thèse riches en bons moments et en discussions scientifiques intéressantes. Je tiens également à remercier plus globalement l'ensemble de l'équipe MeX : Daniel (pour m'avoir accueilli dans son équipe), Marc, Sandrine, Nicolas, Anne, François, Elodie, Marine (le V), Olha, Cynthia, Catherine, Florence et Laurie. Merci beaucoup pour les nombreuses soirées au caporal qui étaient toujours un moment de cohésion ayant contribué à rendre certaines fins de journées fort agréables. Un immense merci à Marine pour avoir été une voisine de bureau hors-pair, tes conseils concernant les choix de couleurs de certaines figures ont certainement sauvé la vue de nombreuses personnes !

Il me tient à cœur de remercier toutes les personnes du laboratoire ToxAlim que j'ai pu croiser au cours de ces trois dernières années et avec qui j'ai toujours eu plaisir à discuter et passer de bons moments. Mention spéciale pour le groupe des graines qui participe grandement à la cohésion des post-doctorants, doctorants, CDDs et stagiaires au sein du laboratoire.
Je remercie également l'Association Nationale de la Recherche et de la Technologie ainsi que L'Oréal sans qui ce projet de thèse n'aurait pu voir le jour. Je tiens également à remercier l'ensemble des personnes qui œuvrent dans l'ombre pour la recherche. Que ce soit Dominique Pantalacci de l'école SEVAB qui m'a sauvé de quelques naufrages administratifs ou encore Gaetan qui s'est assuré de me fournir un apport calorique conséquent et n'a jamais refusé l'une de mes nombreuses demandes de rab' à la cantine !
Enfin je tiens à remercier très chaleureusement mes amis et ma famille pour leur soutien sans faille. Supporter l'absence sociale d'un doctorant en bout de course pendant ces derniers mois n'a pas du être aisé mais vous l'avez fait avec brio ! Je tiens tout particulièrement à remercier mes parents pour leur soutien d'une importance incommensurable, pour avoir toujours cru en moi, m'avoir permis de m'épanouir dans des études et un domaine professionnel qui me tient à cœur et surtout pour m'avoir donné les moyens d'arriver jusqu'ici. A tous, merci infiniment.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACP : Analyse en **C**omposante **P**rincipales

ADN : Acide **D**ésoxyribo**N**ucléique

ARN : Acide **R**ibo**N**ucléique

CIFRE : Convention **I**ndustrielle de **F**ormation par la **Re**cherche

COBRA : COnstraint-**B**ased **R**econstruction and **A**nalysis Toolbox

CoC : Center **o**f **C**ircle

CSN : Carbon **S**keleton **N**etwork

DAR : Differentially **A**ctivated **R**eaction

DEG : Differentially **E**xpressed **G**enes

DEXOM : Diversity-based **E**numeration **O**f context-specific **M**etabolic networks

DILI : Drug-**I**nduced **L**iver **I**njury

DMSO : Di**M**ethyl **S**ulf**O**xide

EDO : Equation **D**ifférentielle **O**rdinaire

FBA : Flux **B**alance **A**nalysis

FDA : Food and **D**rug **A**dministration

FDR : False **D**iscovery **R**atio

FVA : Flux **V**ariability **A**nalysis

GCN : Graph **C**onvulational **N**etworks

GNN : Graph **N**eural **N**etworks

GPR : Gène-**P**rotéine-**R**éaction

GRN : Graph **R**ecurrent **N**etworks

GSEA : Gene **S**et **E**nrichment **A**nalysis

GSMN : Genome **S**cale **M**etabolic **N**etwork

HPH : Hépatocyte **P**rimaire **H**umain

iMAT : integrative **M**etabolic **A**nalysis **T**ool

INRAE : Institut **N**ational de **R**echerche pour l'**A**griculture, l'alimentation et l'**E**nvironnement

LP : Linear **P**rogramming

MILP : Mixed-**I**nteger **L**inear **P**rogramming

mMoA : metabolic **M**echanism **o**f **A**ction

NIBIO : National **I**nstitute of **B**iomedical **I**nnovation

NIHS : National **I**nstitute of **H**ealth **S**ciences

OCDE : Organisation de **C**oopération et de **D**éveloppement **É**conomiques

PAM : Partition **A**round **M**edoids

PM : Partial **M**atch

QSAR : **Q**uantitative **S**tructure **A**ctivity **R**elationship

RMA : Robust **M**ulti-Array **A**verage

SBML : Systems **B**iology **M**arkup **L**anguage

Valorisation scientifique

Les travaux réalisés au cours de cette thèse ont pu être présentés et discutés avec la communauté scientifique à plusieurs occasions. Ces travaux, dans leur phase préliminaire, ont notamment été présentés au travers de posters dans une conférence nationale (JOBIM2022) et deux conférences internationales (COBRA2022 et SOT2023). Le poster présenté à la conférence JOBIM (Journées ouvertes en biologie, informatique et mathématiques) s'étant déroulée à Rennes est intitulé : « Improving the analysis of toxicants Mechanisms of Action with condition-specific models and network analysis. ». Pour ce poster ainsi que le poster suivant, les co-auteurs sont : « Louison Fresnais, Olivier Perin, Anne Riu, Clément Frainay, Fabien Jourdan et Nathalie Poupin ». Avec ces mêmes auteurs, j'ai également communiqué autour des travaux réalisés en me rendant à Galway pour participer à la conférence COBRA (Conference on Constraint-Based Reconstruction and Analysis) avec un poster intitulé « Combining condition-specific modelling and network topological analysis to improve the analysis and visualization of toxicants Mechanisms of Action ». Enfin, nous avons pu présenter un troisième poster la conférence annuelle de la SOT (Society of Toxicology) s'étant déroulée à Nashville et dont le titre est « Combining condition-specific modelling and network topological analysis to improve the analysis and visualization of chemicals' ». Les auteurs de ce poster sont : « Louison Fresnais, Olivier Perin, Anne Riu, Romain Grall, Alban Ott, Bernard Fromenty, Clément Frainay, Fabien Jourdan et Nathalie Poupin ».

J'ai également eu l'opportunité de présenter ces travaux de thèse et notamment la stratégie qui est décrite dans ce manuscrit via une communication orale de 20 minutes à l'ISMB/ECCB (The 31st Annual Intelligent Systems For Molecular Biology and the 22nd Annual European Conference on Computational Biology) qui s'est déroulée à Lyon. Les auteurs du résumé ayant permis d'être invité à cette communication orale sont : « Louison Fresnais, Olivier Perin, Anne Riu, Romain Grall, Alban Ott, Bernard Fromenty, Clément Frainay, Fabien Jourdan et Nathalie Poupin ».

Enfin, nous avons rédigé un article dont une première version est publiée dans BioRxiv (https://doi.org/10.1101/2023.06.30.547200). Les auteurs de cet article sont « Louison Fresnais, Olivier Perin, Anne Riu, Romain Grall, Alban Ott, Bernard Fromenty, Jean-Clément Gallardo, Maximilian Stingl, Clément Frainay, Fabien Jourdan et Nathalie Poupin ». Cet article est actuellement dans la phase d'évaluation par les relecteurs du processus de publication du journal BMC Bioinformatics.

Afin de rendre nos travaux facilement reproductibles et adaptables par d'autres scientifiques, l'ensemble du code est disponible sur GitLab (https://forgemia.inra.fr/metexplore/MANA).

L'ensemble de la stratégie qui sera présentée dans les prochains chapitre est divisée en trois « jupyters notebook » dont le lancement est géré par un « jupyter notebook » principal. Les « jupyter notebook » sont des scripts python interactifs permettant de coder, exécuter du code et visualiser les résultats de manière ergonomique. Afin de faciliter l'installation ainsi que l'utilisation de notre stratégie, un jeu de données « test » est également disponible dans le GitLab.

Les données transcriptomiques utilisées pour ces travaux sont des données publiques disponibles à cette adresse : https://dbarchive.biosciencedbc.jp/en/open-tggates/download.html. Le réseau métabolique Recon2.2 est également disponible à l'adresse suivante https://www.ebi.ac.uk/biomodels/MODEL1603150001. Enfin, les modèles calculés par notre stratégie n'ont pas été déposés en ligne mais peuvent être fournis sur demande.

Introduction générale

Le recours à l'expérimentation animale est un enjeu éthique majeur depuis de nombreuses années. En effet, dès 1959 William Russell et Rex Burch ont défini le principe des 3Rs : « Réduction, Raffinement, Remplacement » qui consiste à réduire le recours à l'expérimentation animale voir la remplacer lorsque cela est possible. Prenant en considération ces principes et fort de ses engagement éthiques, L'Oréal a donc reconstruit dès 1979 de la peau humaine en laboratoire et intensifié sa recherche de méthodes alternatives afin de pouvoir arrêter de tester ses ingrédients sur les animaux dès 1989. De son côté, l'Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAE) a pour ambition de devenir l'un des leaders mondiaux de la recherche pour répondre à des enjeux sociétaux variés (de la sécurité alimentaire à l'étude des risques naturels). Parmi ces enjeux sociétaux, la toxicologie alimentaire et environnementale sont des enjeux d'intérêt étudiés au sein de l'unité de recherche ToxAlim (UMR1331) et dont les parallèles possibles avec le développement de nouvelles approches d'évaluations de sécurité des matières premières cosmétiques sont nombreux. Fort de sa position d'acteur majeur dans la recherche et le développement de nouvelles approches alternatives à l'expérimentation animale, L'Oréal a donc trouvé en l'INRAE et plus particulièrement l'unité de recherche ToxAlim un partenaire idéal pour développer, au travers de cette thèse CIFRE, de nouvelles approches d'évaluation de la sécurité des matières premières cosmétiques.

L'évolution progressive de la régulation européenne au sujet de l'expérimentation animale dans l'industrie cosmétique est un marqueur intéressant des défis qu'a pu représenter et que représente toujours aujourd'hui l'arrêt de l'expérimentation animale. La directive de 2003 entrée en application le 11 septembre 2004 a dans un premier temps mis un terme à l'expérimentation animale pour les produits finis avant d'être complétée par le règlement CE n° 1223/2009 du Parlement européen et du Conseil du 30 novembre 2009 relatif aux produits cosmétiques(https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=celex%3A32009R1223) et notamment les articles 3 et 18 intégrant l'interdiction de l'expérimentation animale pour l'ensemble du parcours d'un produit cosmétique (des recherches initiales au produit fini).

L'interdiction de l'expérimentation animale pour l'industrie cosmétique est entrée en vigueur graduellement avec la validation et l'adoption de méthodes alternatives validées par l'OCDE (Organisation de coopération et de développement économique). Cette interdiction a été retardée au 11 mars 2013 pour les tests de toxicité par doses répétées, la toxicologie reproductive et la toxicocinétique. En effet l'étude des phénomènes toxiques à l'échelle de l'organisme représente encore aujourd'hui un défi de taille pour les nouvelles méthodes alternatives à l'expérimentation animale. Ces trois cas peuvent être regroupés sous l'égide de la toxicité systémique. La toxicité systémique fait référence à des effets toxiques dus à

l'absorption et à la distribution d'une substance dans l'organisme, l'affectant dans son ensemble [2]. Pour étudier la toxicité systémique, nous allons nous intéresser à cinq types de toxicité :

- La toxicité aigüe : toxique après une exposition ponctuelle
- La toxicité chronique : toxique après une exposition prolongée, parfois des années après l'exposition
- La carcinogénicité : favorise l'apparition de cancers
- La reprotoxicité : engendre des perturbations des fonctions de reproductions
- La génotoxicité : engendre des perturbations du fonctionnement du génome ou une altération de l'intégrité de l'ADN

Lorsqu'une molécule présente des effets toxiques à l'échelle du système, elle peut donc engendrer des effets répartis sur plusieurs organes mais on observe généralement un degré maximal de toxicité pour un ou deux organes, que l'on appelle organes cibles. Le foie étant l'un des organes cibles les plus courant, il s'agit généralement de l'organe modèle sur lequel la toxicité systémique est évaluée. L'étude de l'hépatotoxicité est donc en général utilisée comme proxy pour l'étude de la toxicité systémique.

Les approches alternatives à l'expérimentation animale peuvent être séparées en deux groupes. Le premier groupe faisant référence aux approches *in silico* et notamment les approches basées sur la comparaison structurale telles que les modèles QSAR (Quantitative Structure Activity Relationship) qui ont pour objectif de prédire l'impact d'une variation structurale sur l'activité biologique d'une molécule. L'utilisation d'empreintes moléculaires est également une approche *in silico* couramment utilisé afin de rechercher des analogues structuraux à une molécule d'intérêt [3]. Cependant, les approches alternatives basées sur la structure moléculaire sont de moins en moins adaptées aux nouvelles classes de matières premières cosmétiques qui sont de plus en plus des composés naturels ou des mélanges complexes (*i.e.* un mélange de plusieurs molécules) pour lesquels les structures ne sont pas connues. Le second groupe fait référence aux approches *in vitro*. Il existe un grand nombre d'approche *in vitro* allant de la génération de données omiques (génomique, transcriptomique, protéomique, métabolomique, …) à des tests de génotoxicité (*i.e.* des tests permettant de mesurer les dommages à l'ADN) et cytotoxicité (*i.e.* des tests dont l'objectif est de mesurer la toxicité cellulaire).

Notons que les approches *in silico* et *in vitro* sont mutuellement dépendantes les unes des autres. En effet, les approches *in silico* ne pourraient exister sans les connaissances et données générées par les approches in *vitro* et les approches *in vitro,* qui sont maintenant essentiellement à haut débit, nécessitent des outils computationnels toujours plus performants

afin d'être analysées et de décoder les informations biologiques qu'elles renferment. L'un des meilleurs exemples de ce lien *in vitro/in silico* étant le développement rapide de nombreuses approches visant à rechercher des analogues biologiques par l'analyse de données de transcriptomiques ou encore utilisant des données générées *in vitro* pour l'entrainement de modèles prédictifs de l'hépatotoxicité [4–6].

Cependant, ces approches souffrent pour le moment de quelques limites telles que l'absence de prise en compte des effets post-traductionnels, l'absence de prise en compte directe des effets métaboliques et pour la majorité ces approches, l'interprétation des gènes indépendamment les uns des autres. Il est également intéressant de noter la nature statique des données transcriptomiques qui ne permet donc pas une compréhension dynamique du phénomène de toxicité. Cette limite ne sera pas adressée au cours de nos travaux mais mérite à nos yeux d'être mentionnée car source de perspectives intéressantes pour le développement de nouvelles méthodes alternatives.

Pour tenter de répondre à ces défis, il serait intéressant de mieux comprendre l'effet des composés testés sur le métabolisme cellulaire dans l'étude de la sécurité des matières premières cosmétiques. En se plaçant au niveau du métabolisme cellulaire, nous serions au plus près du phénotype, ce qui permettrait de prendre en compte les potentiels effets post-traductionnels ainsi que d'étudier directement le métabolisme cellulaire et ses implications dans les phénomènes hépatotoxiques. Les approches de modélisation sous-contraintes à partir des réseaux métaboliques pourraient en prime permettre de ne plus considérer les gènes indépendamment les uns des autres mais plutôt de les considérer au travers de la topologie du réseau et des enzymes pour lesquels ils codent.

Après cette courte introduction permettant d'introduire le contexte des travaux réalisés au cours de ces trois dernières années, nous allons maintenant explorer les concepts majeurs nécessaires à la modélisation du métabolisme cellulaire de cellules humaines.

Chapitre 1 : Modélisation de la réponse du métabolisme cellulaire après exposition à un xénobiotique

1. Le métabolisme cellulaire

Le métabolisme cellulaire peut-être défini comme un ensemble de réactions biochimiques interconnectées et finement régulées permettant de fournir de l'énergie, maintenir l'homéostasie cellulaire et produire des métabolites nécessaires à la réalisation des fonctions cellulaires. Le métabolisme peut être représenté sous forme de réseaux afin de prendre en compte l'interconnexion existante entre les réactions et les métabolites. Les réactions biochimiques peuvent être catalysées par des enzymes ou être spontanées. Elles ont pour rôle de cataboliser (dégrader), anaboliser (produire) ou transporter des métabolites. On appelle métabolite toutes les molécules de faible poids moléculaire impliquées dans ces processus métaboliques. Le métabolisme peut être étudié par des approches omiques allant de la transcriptomique (proche du génome) jusqu'à la métabolomique, considérée comme étant l'approche omique la plus proche du phénotype [7].

1.1. Régulation du métabolisme

Le métabolisme cellulaire varie de manière importante en fonction du type cellulaire considéré, des contraintes environnementales telles que l'exposition à un composé chimique, mais également en fonction de l'absence ou au contraire de l'abondance d'un ou plusieurs nutriments. Par exemple, les cellules immunitaires peuvent passer d'un stade dormant à un stade d'importante prolifération suite à la réponse à un stimuli extérieur [8]. De tels cas de figures existent pour de nombreux autres types cellulaires ou conditions environnementales tels que l'activation des métabolismes aérobie ou anaérobie dans les cellules musculaires [9] ou la réponse de cellules hépatiques à la présence de substances exogènes dans la circulation sanguine. Plusieurs grands types de mécanismes de régulation interviennent dans la régulation du métabolisme est identifié. En premier lieu, des modifications de l'ADN telles que des mutations ou des modifications de la structure de la chromatine peuvent impacter la transcription des gènes codant pour des enzymes du métabolisme. L'expression des gènes du métabolisme peut également être régulée par des facteurs de transcription qui peuvent être des récepteurs nucléaires[10] sensibles à des molécules exogènes (*e.g*. des contaminants alimentaires ou chimiques, des molécules endogènes, …). Par exemple, le PFOA (une molécule que l'on retrouve notamment dans les revêtements anti-adhésif) serait un potentiel facteur d'activation de PPARalpha, un récepteur nucléaire impliqué dans la régulation du métabolisme des lipides au niveau hépatique [11] Enfin la régulation du métabolisme peut également avoir lieu à l'échelle enzymatique au travers de boucles de rétrocontrôles qui vont engendrer une inactivation enzymatique en présence d'un excès de substrat [12] ou un excès de produit [13]. La modulation des activités enzymatiques ainsi que la disponibilité des nutriments provenant du milieu extérieur impactent également les flux de métabolites exerçant ainsi cette action de rétrocontrôle sur le métabolisme cellulaire.

1.2. Métabolisme cellulaire des hépatocytes

Au cours de nos travaux, nous avons focalisé notre attention sur la modélisation du métabolisme hépatique. En effet, le foie est un organe majeur pour les mécanismes de détoxification des molécules endogènes mais également dans la régulation du métabolisme endogène. C'est pourquoi nous allons aborder plus précisément le métabolisme cellulaire des hépatocytes au cours de cette section. Les hépatocytes représentent 65% de la totalité des cellules du foie et occupent plus de 80% du volume de l'organe. Les hépatocytes ont deux rôles principaux : d'une part ils jouent un rôle prépondérant dans la régulation et le stockage énergétique et d'autre part jouent les premiers rôles en ce qui concerne la détoxification de composés exogènes présents dans la circulation sanguine. Les hépatocytes sont en effet capables de stocker le glucose sous forme de glycogène lorsqu'il est présent en excès et de le libérer dans la circulation sanguine via l'activation de la glycolyse lorsque la concentration sanguine de glucose est trop basse [14]. Un mécanisme similaire peut avoir lieu avec le stockage de lipides (sous forme de triglycérides) dans les hépatocytes et leur déstockage via la β-oxydation des acides gras en cas de besoin en apport énergétique [14].

Les hépatocytes sont également responsables de la majorité des capacités de biotransformation des molécules exogènes (médicaments, alcool, …) grâce à l'activité d'enzymes de la famille des Cytochromes P450 [15].

L'ensemble de ces fonctions est essentiel au maintien de l'homéostasie métabolique et au bon fonctionnement de l'organisme, ce qui fait des hépatocytes (et du foie dans son ensemble) des cellules clés du métabolisme.

1.3. Modèles cellulaires pour l'étude du métabolisme cellulaire hépatique

1.3.1. Hépatocytes Primaires Humains

Les hépatocytes sont des cellules modèles pour étudier les maladies métaboliques ainsi que les phénomènes de toxicité systémique induits par des médicaments [16]. L'étude *in vitro* du métabolisme des hépatocytes et de l'impact des xénobiotiques, des molécules pharmaceutiques, industrielles ou cosmétiques, sur ces derniers peut être réalisée en isolant des hépatocytes à partir d'échantillons de foies (*i.e.* provenant de chirurgies ou d'expérimentations *in vivo*). Les hépatocytes ainsi obtenus sont appelés des hépatocytes primaires. Lorsque ces hépatocytes sont obtenus à partir de foies humains, on parle d'hépatocytes primaires humains (HPH). Les HPHs conservent la structure ainsi que la plupart des fonctions de leurs équivalents *in vivo* mais perdent cependant certaines fonctions membranaires [17]. Leur principale limite est que leur durée de vie n'est que de quelques heures à quelques jours en culture cellulaire. De plus, en conservant la majorité des fonctions

des hépatocytes *in vivo*, les HPH conservent également les spécificités propres au donneur et donc il existe une variabilité forte liée au « donneur » lors de l'analyse de données obtenues sur culture d'HPH. Dans une étude de Lee *et al* [18], il a été démontré que la viabilité des HPH prélevés dépendait majoritairement des caractéristiques du donneur : âge, indice de masse corporelle, contenu lipidique du foie, dommages hépatiques, données biologiques sanguines. Des facteurs techniques de variabilité liés aux différents protocoles d'isolation et de filtration cellulaires utilisés lors de l'obtention des HPH sont également à prendre en compte. Cette variabilité inter-individuelle peut être un atout lorsque l'on conduit une étude toxicologique pour un composé en particulier car il est connu que la variabilité hépatique inter-individuelle peut-être à l'origine de différences importantes en termes d'effets secondaires et donc de toxicité [19–21]. Dans ce cas précis, l'utilisation de différents lots d'HPH, provenant de différents donneurs, permettra d'étudier un large panel de réponses métaboliques possibles suite à l'exposition à un composé chimique, donc une caractérisation et une compréhension plus exhaustive du mécanisme d'action et du potentiel hépatotoxique du composé en population générale. Il est également intéressant de noter que les hépatocytes primaires humains et les hépatocytes primaires de rongeurs présentent des différences importantes sur des fonctions impliquées dans la détoxification telles que les cytochromes P450[17]. Ce type de différence entre ces deux modèles cellulaires n'est pas négligeable pour l'interprétation des études de toxicité réalisées sur ces modèles.

1.3.2. Lignées hépatocytaires « immortelles »

En raison des limites évoquées précédemment (effet donneur et durée de survie limitée) mais également à cause de la difficulté d'obtention des HPH, des lignées cellulaires « immortelles » d'hépatocytes sont couramment utilisées, notamment les cellules HepG2 [22] et HepaRG [23] , qui sont issues de lignées cellulaires tumorales. Ces lignées ont l'avantage de pouvoir se répliquer à l'infini et donc d'être théoriquement immortelles. Cependant, il a été observé que ces lignées perdent graduellement, au fur et à mesure des repiquages cellulaires, les fonctions métaboliques spécifiques au foie que la lignée initiale était capable d'exprimer. Il est donc conseillé de caractériser le potentiel métabolique de la lignée cellulaire en culture avant de pouvoir interpréter les résultats issus de ces cellules [17]. Les cellules HepaRG sont des cellules possédant des fonctionnalités proches de celles des HPH tout en évitant les problèmes de variabilités liées à ce modèle [24] De plus, les cellules HepaRG conservent ces fonctionnalités de manière stable pendant plusieurs semaines, ce qui en fait un modèle adapté pour l'étude de phénomènes d'exposition prolongée à de faibles doses de xénobiotiques [25]. Ce modèle cellulaire est par ailleurs reconnu et recommandé par l'OCDE ainsi que Tox21 (un programme américain d'évaluation de la toxicité). La lignée HepaRG est issue d'une lignée tumorale et est capable de proliférer jusqu'à atteindre une certaine densité de cellules en culture (*i.e.* la

confluence) puis de se différencier en deux types cellulaires : des cellules proches des hépatocytes ainsi que des cellules proches des cellules biliaires. Une fois activées par du DMSO, ces cellules expriment de nombreuses fonctions cellulaires et une bonne similarité avec les hépatocytes primaires humains [17] mais en ayant une durée de vie en culture supérieure.

2. Etude de la toxicité systémique *in vitro* par les approches omiques

Comme nous avons pu l'aborder en introduction générale, la toxicité systémique est l'ensemble des effets toxiques dus à l'absorption et à la distribution d'une substance et ses métabolites dans l'organisme et se manifestant dans l'ensemble de l'organisme. L'évaluation de la toxicité systémique est un élément essentiel avant toute mise sur le marché de produit pharmaceutique ou cosmétique. Elle peut également avoir lieu dans un contexte de recherche fondamentale ou *a posteriori* pour des substances dont des effets toxiques sont suspectés et nécessitant donc une étude approfondie comme c'est le cas pour de nombreux contaminants alimentaires, qu'ils soient d'origine naturelle comme certaines mycotoxines ou d'origine industrielle comme les Bisphénols ou les composés perfluorés. Historiquement l'étude de la toxicité systémique est réalisée *in vivo* car il s'agit du modèle le plus adéquat pour étudier un phénomène aussi multi-factoriel et généralisé [26], permettant d'intégrer les processus de régulation ainsi que les effets sur les différents tissus. Cependant, l'extrapolation des résultats d'évaluation de la toxicité obtenus sur un organisme modèle (*e.g.* le rat) vers l'organisme cible (*e.g.* l'humain) peut donner lieu à des différences importantes en termes de toxicité [27,28] et pour des raisons éthiques, la réglementation 3R nécessite maintenant de réduire au maximum voire l'arrêt complet de l'expérimentation *in vivo*. C'est pourquoi de nombreuses approches omiques et des tests cytotoxiques *in vitro* sont développées afin d'estimer la toxicité systémique en se focalisant sur l'étude de la toxicité d'organes cibles tels que le foie ou les reins. Ces approches peuvent être utilisées en combinaison avec des test *in vivo* ou comme méthode d'évaluation principale de la toxicité comme c'est le cas pour l'industrie cosmétique (se référer à l'introduction générale). Le terme « omique » est dérivé du suffixe des différents niveaux d'exploration d'un système biologique des sciences de la vie tels que la génomique, la transcriptomique, la protéomique, la métabolomique, la lipidomique, etc [29]. La génération de données omiques fait appel à des méthodes de génération de données à haut débit et implique donc des grandes quantités de données à stocker et analyser. Il existe un grand nombre de base de données publiques permettant de stocker et d'analyser des données omiques. Parmi les bases de données omiques les plus connues, on peut citer Ensembl [30] pour les données issues d'analyse génomiques, GEO [31]pour la transcriptomique, HPA [32] pour la protéomique et MetaboLights [33] pour la métabolomique.

2.1. Des bases de données publiques d'exposition à des xénobiotiques

La recherche scientifique tend de plus en plus à être ouverte et accessible par le plus grand nombre. Cette ouverture de la recherche se fait au travers d'un accès ouvert aux publications mais également aux codes et données [34]. Des besoins de stockage permettant le libre accès à des données de qualité ont donc émergé sous la forme de bases de données publiques. Etant donné qu'il existe des bases de données pour une large variété de données, nous allons focaliser notre attention sur les bases de données contenant des données adaptées à notre sujet d'étude, à savoir des bases de données en toxicologie.

L'utilisation de bases de données publiques permet de réutiliser des données générées pour d'autres expériences et ainsi de maximiser la valeur scientifique de ces données générées. La majorité de ces bases contient des données d'expression génique obtenues avec différentes technologies. On peut par exemple citer Open TG-Gates [35] ainsi que la DrugMatrix [36] qui recensent des données transcriptomiques générées sur puces à ADN après exposition à des xénobiotiques. Open TG-Gates a l'avantage de combiner des données *in vitro* générées sur hépatocytes primaires humain et hépatocytes primaires de rat avec des données générées *in vivo* chez le rat alors que la base de données DrugMatrix ne contient que des données générées *in vivo* et *in vitro* chez le rat. La base de données CMap [37] et son extension avec la base de données LINCS [38] constitue l'une des plus grandes bases de données transcriptomique d'exposition à des xénobiotiques. La base de données LINCS a été générée grâce au protocole L1000 [38] qui consiste à mesurer l'intensité d'expression d'environ 1000 gènes représentatif du niveau d'expression de tous les gènes connus du génome humain. Ce protocole a notamment permis une réduction des coûts et ainsi de tester plus de 29 668 composés et modifications génétiques sur 98 lignées cellulaires. La création d'une telle base de données représente une avancée considérable pour la disponibilité de données omiques permettant l'évaluation de la toxicité sur un grand nombre de composés mais présente des limites de reproductibilité [39]. Par exemple, les données transcriptomiques générées dans la première version de la base de données CMap sont faiblement corrélées avec les données générées dans la deuxième version de la CMap (LINCS), à condition identique [39]. Ce qui indique que les données contenues dans chacune de ces bases sont difficilement comparables entres elles mais également avec des données générées dans le cadre d'autres études. Cependant, le grand nombre de bases de données omiques implique des chevauchements importants en termes de contenu entre ces différentes bases et également des données de qualité variable qu'il convient de prendre en compte lors de l'utilisation de ces bases de données. Enfin, il est important de disposer de bases de données permettant de recenser les effets secondaires connus, observés et publiés suite à l'exposition à des xénobiotiques. En effet, il s'agit en général du type d'effet que l'on cherchera à prédire et donc ce type d'information est utile pour l'entrainement de

modèle prédictif. On peut notamment obtenir ce type d'informations dans la base de données SIDER [40] ainsi que la base de données CTD [41] qui tirent toutes deux parties de la grande quantité d'information disponible dans la littérature.

2.2. Faciliter l'interprétation des données omiques par l'enrichissement fonctionnel

La complexité ainsi que la grande dimension des données omiques et particulièrement des données transcriptomiques implique d'utiliser des méthodes bio-informatiques afin de pouvoir faciliter l'interprétation de ces données et donc leur utilisation pour améliorer l'évaluation de la toxicité systémique. L'une des approches les plus couramment utilisée est l'enrichissement fonctionnel : cette approche consiste à identifier des voies biologiques ou des groupes de gènes impliqués dans une fonction commune (*e.g.* les termes de la Gene Ontology [42] par exemple) significativement surreprésentés dans les données. Pour réaliser un enrichissement fonctionnel à partir de données transcriptomiques (Fig 1), il faut dans un premier temps obtenir une liste de gènes d'intérêt. Afin de déterminer cette liste de gènes, une étude d'expression différentielle est généralement réalisée afin d'identifier les gènes étant significativement dérégulés entre deux conditions. Après avoir obtenu cette liste de gènes d'intérêt, il conviendra de choisir la base de voies biologiques ou de d'ensembles de gènes à laquelle comparer la liste de gènes d'intérêt puis de réaliser un test statistique afin de déterminer si certaines voies biologiques, représentées par des listes de gènes, sont significativement associées à la liste de gènes d'intérêt. Le test statistique généralement utilisé pour les analyses de sur-représentation est le test exact de Fisher mais d'autres alternatives telles que la GSEA (Gene Set Enrichment Analysis) [43] sont également classiquement utilisées.

De nombreux packages R et serveurs web tels que EnrichR [44] ou g:Profiler [45] permettent de réaliser ce type d'analyse rapidement et plutôt facilement.

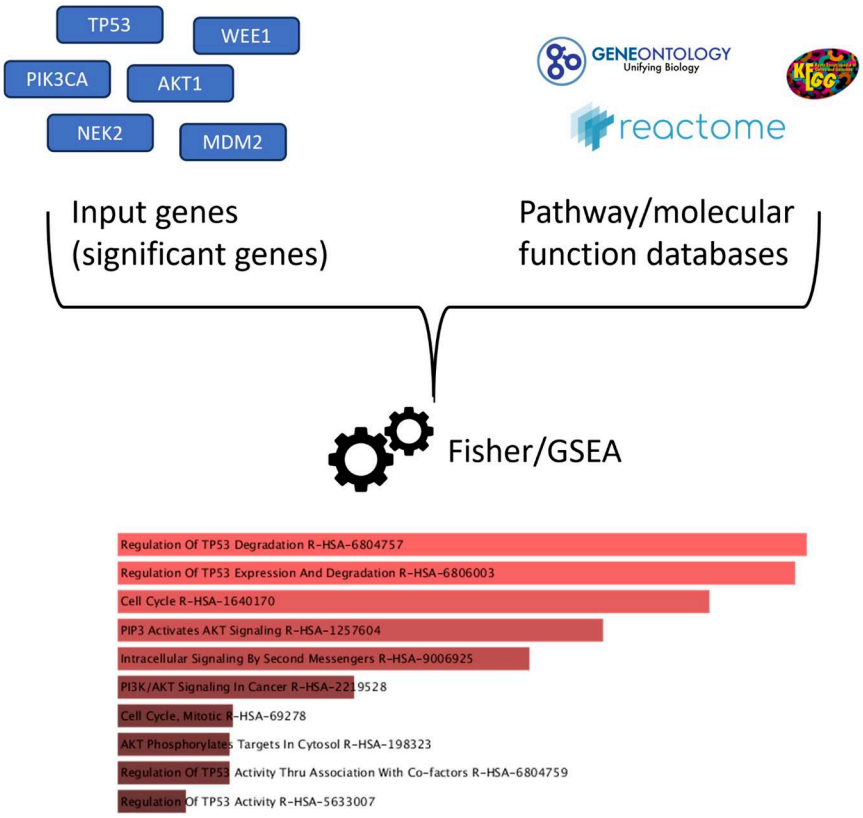

Figure 1 : Protocole d'enrichissement fonctionnel d'une signature transcriptomique.

L'enrichissement fonctionnel est donc une étape importante dans l'interprétation des données transcriptomiques en permettant de passer de milliers de gènes différentiellement exprimés difficilement interprétables à des fonctions perturbées et donc une meilleure compréhension de la réponse cellulaire d'une condition à une autre. Cependant, le choix de la base de données de référence peut influencer de manière importante les résultats d'enrichissement fonctionnel car les approches de sur-représentation sont sensibles à la taille des groupes fonctionnels définies dans ces bases. Un groupe fonctionnel de petite taille sera plus susceptible d'être significative alors qu'un groupe fonctionnel de grande taille le sera moins [46,47]. En utilisant des tests statistiques tels que le test exact de Fisher qui fait l'hypothèse d'indépendance entre les variables (les gènes), ces approches ne prennent pas en compte l'action de facteurs de transcription et d'autres mécanismes de régulation affectant plusieurs gènes simultanément. Néanmoins les approches d'enrichissement fonctionnel constituent une première étape dans la caractérisation et la compréhension des perturbations induites par des composés chimique faisant le lien avec les approches de read-across que nous allons aborder dans la prochaine section.

2.3. Approches de Read-Across pour l'évaluation de la toxicité systémique

L'objectif du « Read-Across » consiste à évaluer la toxicité d'une molécule en se basant sur les effets connus d'une molécule de structure chimique proche. Historiquement, le read-across était basé uniquement sur la recherche d'analogues structuraux en considérant que deux composés structurellement proches induisaient des effets biologiques similaires [48,49]. Cependant, il a été démontré que certains composés ayant des structures très similaires, pouvant donc être considérés comme de très bons analogues structuraux, pouvait engendrer des effets toxiques très différents [48]. Cette différence de toxicité pour des molécules structurellement proches voir identiques a été mis au jour avec le scandale du Thalidomide qui a eu lieu au début des années 60. Le Thalidomide est un médicament qui était prescrit pour soigner des troubles du sommeil, de l'anxiété et des nausées. Le thalidomide possède deux énantiomères présent en quantité équivalentes en solution, le S-Thalidomide qui engendre une inhibition de TNFα (ce qui n'est pas l'effet recherché) et le R-Thalidomide qui engendre les effets sédatifs attendus. Il s'est avéré que l'énantiomère R-Thalidomide présente des effets tératogènes importants [50] alors que les deux structures sont identiques [51] Ce type de phénomène invite donc à considérer le read-across basé sur la structure uniquement avec précaution. De plus, de nombreux principes actifs sont en réalité des mélanges complexes et/ou des composés naturels dont on ne connaît pas la structure, rendant impossible le read-across structural.

Une solution a donc été de tirer parti de l'essor des données omiques pour rechercher des analogues biologiques en plus d'analogues structuraux [52–55]. Un analogue biologique est une molécule ayant des effets biologiques mesurés par une ou plusieurs technologies omique similaires à la molécule d'intérêt dont on cherche à connaitre les potentiels effets secondaires. Le read-across biologique constitue donc un champ d'applications des données omiques prometteurs et un domaine de recherche très actif. Le read-across est d'ores et déjà une des approches les plus fréquemment utilisées pour l'évaluation du risque [56]. Cependant, de nombreux défis subsistent. En effet, étant donné que le read-across biologique a pour objectif d'identifier des molécules similaires grâce aux données omiques, il est essentiel de garantir que ces données omiques ne comportent pas de biais techniques (expérimentateur, choix méthodologiques, …) et biologiques (lot de cellule, types cellulaires, …) auquel cas on risquerait d'identifier comme analogues biologiques des données présentant seulement des biais similaires et non pas une réelle similarité biologique. L'un des grands défis à résoudre concernant le read-across biologique, et la toxicogénomique de manière générale, est donc l'harmonisation des données entre les bases de données publiques ainsi qu'entre les différentes études réalisées au cours du temps. En l'absence de méthode d'harmonisation des données éprouvée il serait risqué de combiner ces bases de données sans prendre le risque d'identifier

de « faux » analogues biologiques. C'est pourquoi avons choisi de travailler avec une seule base de données, que nous décrirons au cours du prochain chapitre. Comme nous avons pu l'aborder au cours de cette section, il existe de nombreuses approches, principalement basées sur des données transcriptomiques ou structurales, permettant d'étudier la toxicité *in vitro*. Cependant, les effets toxiques sur le métabolisme cellulaire ne sont pas directement étudiés par ces approches. Les réseaux métaboliques à l'échelle du génome dont nous allons discuter dans la prochaine section s'avèrent être de bons candidats pour étudier des effets toxiques sur le métabolisme cellulaire.

3. Modélisation du métabolisme cellulaire, de la modélisation statistique aux réseaux métaboliques

3.1. Introduction à la modélisation du métabolisme cellulaire

Modéliser le métabolisme c'est représenter le métabolisme d'une cellule, d'un tissu, d'un organisme ou même de communautés au travers d'une représentation mathématique. Le métabolisme peut être modélisé de manière simplifiée par des modèles statistiques [57]. L'objectif de ces modèles est d'identifier des métabolites « biomarqueurs » à partir de jeux de données de métabolomique [58]. Ces modèles sont des modèles de statistiques multivariées telles que la régression linéaire ou la PLS-DA qui est l'une des approches les plus utilisées pour l'analyse des données de métabolomiques [59,60].

Il est également possible de modéliser le métabolisme par un ensemble d'équations linéaires ou différentielles. La modélisation par des équations linéaires est notamment utilisée par les approches de modélisation sous-contraintes et que nous développerons en détail au cours des prochaines sections. La modélisation par des équations différentielles est utilisée pour les approches de modélisation cinétiques. Enfin, il est également possible de modéliser le métabolisme par des approches de graphes.

Au cours de ce projet, nous avons combiné une approche de modélisation sous-contraintes à une approche de modélisation par des graphes. Avant de détailler ces deux types de modélisation, nous allons rapidement aborder la modélisation cinétique qui est un type de modélisation du métabolisme largement décris dans la littérature mais que nous n'avons pas utilisé au cours de ce projet.

3.1.1. Modélisation cinétique

La modélisation cinétique permet de simuler l'évolution des concentrations de métabolites au cours du temps [61]. Pour ce faire, les modèles cinétiques du métabolisme représentent les mécanismes ayant lieu dans une cellule à l'aide d'un ensemble d'équations différentielles ordinaires (EDO). Chaque équation différentielle du modèle décrit la variation de quantité pour chaque métabolite au travers d'une équation cinétique dont au moins l'un des paramètres cinétiques est inconnu. Les modèles cinétiques sont constitués de plusieurs réactions métaboliques (de quelques dizaines à quelques centaines) décrivant l'équilibre de masses dans le modèle [62]. Selon l'objectif de l'étude, on peut avoir des modèles cinétiques ne représentant qu'une voie métabolique, comme le modèle de la voie de l'aspartate dans les chloroplastes d'*Arabidopsis Thaliana* qui contient seulement 13 enzymes [63]. Il existe également des modèles cinétiques couvrant plusieurs voies métaboliques tel que le modèle représentant le métabolisme du glucose hépatique publié dans [64] et contenant 49 métabolites, 36 réactions et 3 compartiments. Enfin, pour les organismes plus simples, il existe des modèles cinétiques dit « à l'échelle du génome » comme le modèle « k-ecoli457 » contenant 457 réactions et 337 métabolites [65].

Après avoir défini les métabolites et réactions (enzymes) représentant l'organisme, le tissu ou la voie métabolique à modéliser, il est nécessaire de définir l'équation différentielle de chaque réaction (Fig 2.1 et 2.2.1). Il existe plusieurs formalismes mathématiques pour décrire ces réactions : Canonique, Approximé et Mécanistique. Le formalisme canonique se base exclusivement sur la structure du réseau pour décrire les réactions enzymatiques, le formalisme approximé fait appel à des lois enzymatiques telles que la loi de Michaelis-Menten. Enfin le formalisme mécanistique est le formalisme le plus précis et est basé sur la description mathématique du mécanisme de chaque réaction au travers d'équations décrivant la conservation des masses ainsi que les lois de thermodynamiques [62]. La construction du modèle cinétique nécessite donc de choisir le formalisme adapté au niveau de connaissance du système [62]. Le niveau de connaissance du système et notamment les différents paramètres cinétiques peut être améliorer grâce à des mesures *in vitro,* des données provenant de la littérature ou grâce à des approches d'apprentissage machine permettant de prédire ces paramètres [66]. L'une des phases critiques de la construction d'un modèle cinétique est la détermination des paramètres (Fig 2.3), paramètres qui peuvent être inconnus car n'ayant pas encore été mesurés. Lorsque les paramètres cinétiques ont été mesurés *in vitro*, il est nécessaire d'optimiser ces paramètres afin d'obtenir un modèle cinétique le plus précis possible. Cependant l'optimisation et l'acquisition de ces paramètres cinétiques pour les modèles de grande taille (plusieurs milliers de réactions) est une des limites actuelles de la

modélisation cinétique et fait l'objet de nombreux développements algorithmiques par la communauté [67,68].

Une fois les équations, les paramètres et les valeurs initiales définies, le modèle cinétique (Fig 2) peut être utilisé pour prédire l'évolution des concentrations de métabolites et des flux propres à chaque réaction au cours du temps.

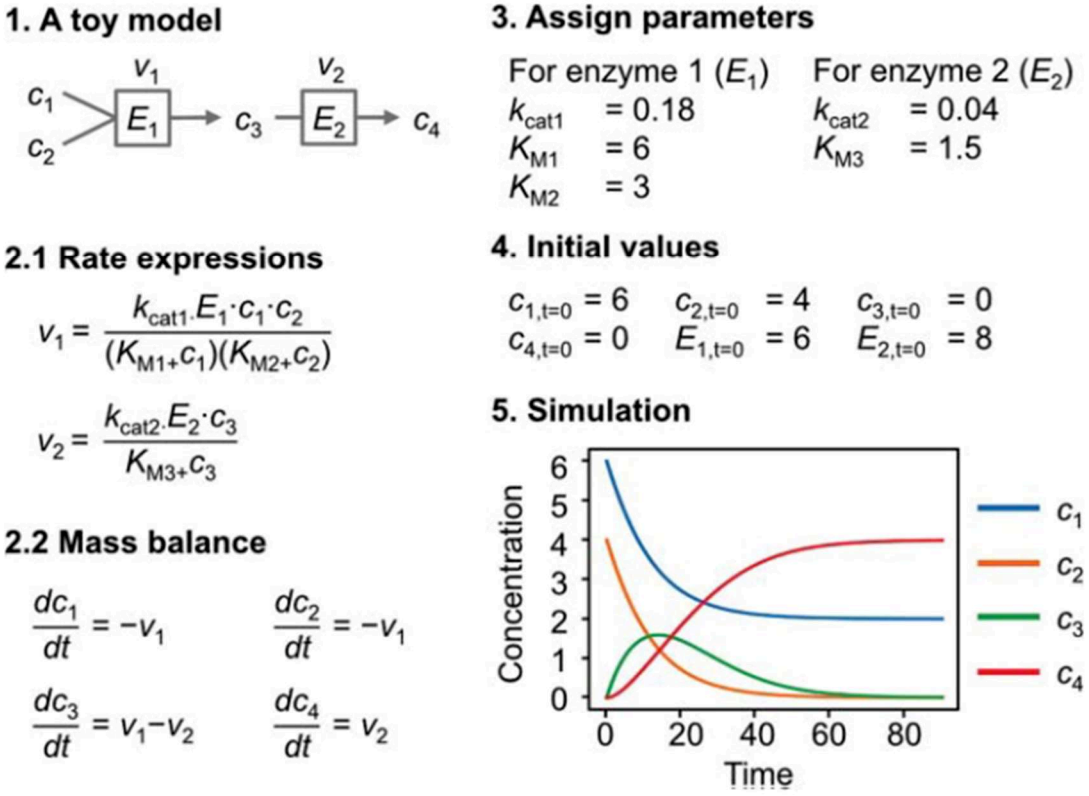

Figure 2: Exemple jouet de la construction d'un modèle cinétique. *Figure provenant de* [61]

Dans la prochaine section, nous allons détailler l'un des éléments centraux de la modélisation du métabolisme cellulaire : le réseau métabolique. Ce formalisme est particulièrement adapté à la représentation de données complexes (plusieurs milliers de réactions et métabolites en interaction) comme celles représentant le métabolisme cellulaire.

3.2. Modélisation du métabolisme cellulaire et réseaux métaboliques

Les réseaux métaboliques à l'échelle du génome (ou Genome Scale Metabolic Networks - GSMN) sont des réseaux contenant l'ensemble des réactions métaboliques connues pour un organisme donné. Comme leur nom le suggère, ces réseaux sont reconstruits à partir du génome annoté de l'organisme. Pour des organismes multicellulaires de grande taille tels que l'être humain, ces réseaux sont constitués de plusieurs milliers de réactions et de métabolites répartis entre 1 ou plusieurs compartiments cellulaires afin de représenter la localisation des

processus biochimiques de la cellule à modéliser. Initialement ces réseaux ont été créés pour représenter le métabolisme de souches bactériennes puis cette approche a été graduellement étendue à d'autres organismes tels que les archées, les levures, les plantes et aux mammifères dont l'Homme [69].

Dans ces réseaux, une réaction est définie par un ou plusieurs métabolites qu'elle consomme, également appelés « substrats », afin de produire un ou plusieurs métabolites appelés « produits ». Si la réaction est catalysée par une enzyme alors cette enzyme peut être associée à un ou plusieurs gènes codants. Par exemple, sur la Figure 3A, les métabolites (représentés par des cercles) D-Glucose et ATP sont des substrats de la réaction (représentée par un carré) catalysée par une hexokinase et qui a pour produits le D-Glucose-6-Phosphate, de l'ADP ainsi qu'un proton. Ces produits sont consommés par d'autres réactions du réseau métabolique connectant ces réactions les unes aux autres. En appliquant ce principe à l'ensemble des réactions et métabolites connus du métabolisme cellulaire d'un organisme donné, il est possible d'obtenir un réseau représentant l'ensemble des interactions (Fig 3B) entre toutes les réactions biochimiques du métabolisme de cet organisme, en l'occurrence l'humain.

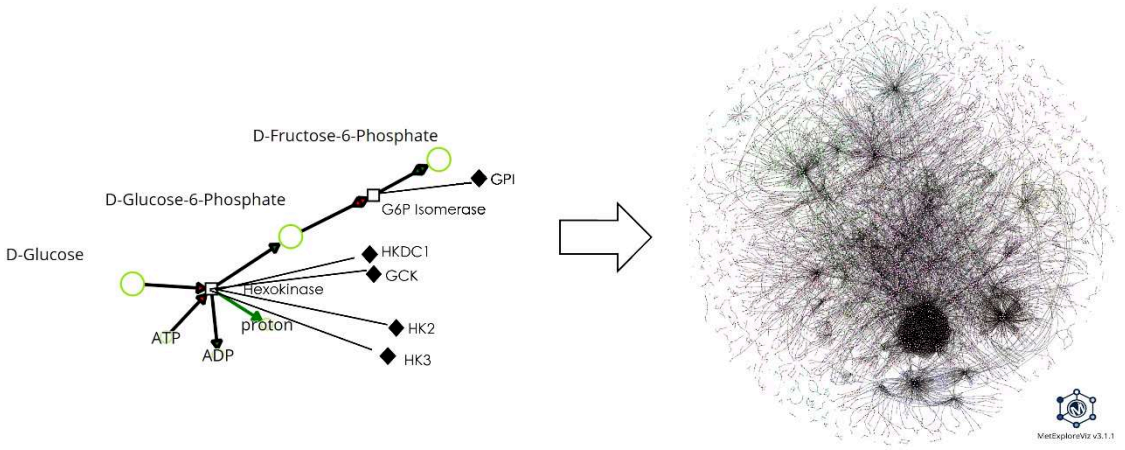

Figure 3 : Exemple de la structure d'un GSMN*. La figure 3A correspond à un zoom sur 2 réactions du réseau Recon2.2 visualisées avec MetExploreViz* [70]*. Les métabolites sont représentés par des cercles et les réactions par des carrés. Les losanges représentent les gènes codant pour les enzymes auxquelles ils sont reliés et les flèches indiquent le sens de la réaction (quel métabolite est consommé et quel métabolite est produit). La figure 3B est une visualisation de Recon2.2, un GSMN de l'humain constitué de 7785 réactions et 5323 métabolites.*

L'organisation Gène-Protéine-Réaction (ou GPR) (Fig 4) est une structure importante permettant d'interconnecter tous les éléments d'un GSMN. Elle fait le lien entre le ou les gènes codant pour une enzyme qui catalyse une ou plusieurs réactions. Ce lien entre les gènes et les protéines est un lien booléen. Par exemple, sur la Figure 4, si le gène b est actif alors les

réactions 2 et 3 catalysées par l'enzyme B seront actives. A l'inverse, si le gène b n'est pas exprimé alors l'enzyme B ne sera pas synthétisée et les réactions 2 et 3 seront inactives. Certaines enzymes sont constituées de plusieurs sous unités. Par exemple, sur la Figure 4, la réaction 4 est catalysée par une enzyme constituée de l'enzyme C et de l'enzyme D. Dans ce cas la GPR de cette réaction a une règle booléenne « AND ». C'est-à-dire que le gène c ET le gène d doivent être tous les deux actifs pour que les deux sous-unités enzymatiques soient traduites et que la réaction 4 puisse être catalysée. Il existe également des règles booléennes de type « OR », ces règles représentent les cas ou plusieurs isoenzymes peuvent catalyser la même réaction. Ce cas correspond à la réaction 5 (Fig 4) pour laquelle il suffit que le gène e OU le gène f soit actif pour que l'une des deux isoenzymes soit synthétisée et que la réaction 5 soit active. Enfin, pour les cas plus complexes ou plusieurs enzymes composées de plusieurs sous-unités enzymatiques permettent de catalyser une réaction, des règles « AND » et « OR » sont combinées. Comme nous le verrons dans les prochains chapitres, les GPRs sont essentielles pour intégrer des données omiques aux GSMNs.

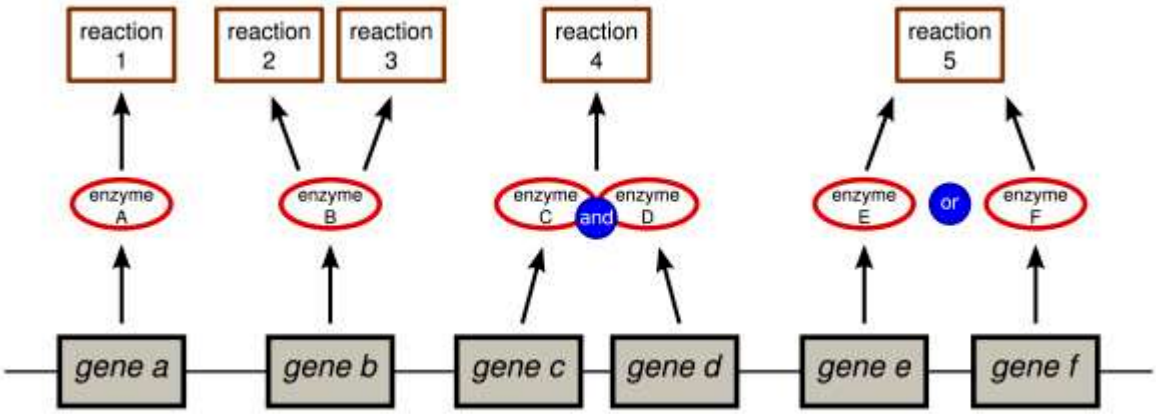

Figure 4: Schéma des différents types basiques d'association Gène-Protéine-Réaction existant dans un GSMN. *Figure provenant de* [71]

Afin d'avoir un réseau métabolique représentatif du fonctionnement du métabolisme d'un organisme, les coefficients stœchiométriques des réactions du réseau sont stockés dans une matrice de stœchiométrie (Fig 5). Chaque ligne de cette matrice représente un métabolite du réseau et chaque colonne représente une des réactions du réseau. Si le métabolite n'est pas impliqué dans une réaction alors il aura la valeur 0 dans la colonne correspondant à cette réaction. Si ce métabolite est impliqué en tant que substrat il aura une valeur négative correspondant à la quantité consommée par la réaction (*e.g. -1)* et s'il est impliqué en tant que produit de cette réaction, il aura une valeur positive correspondant à la quantité produite par la réaction (*e.g. 1)*.

Réactions

R1 :	2M1 ->M3
R2 :	M2 ->M1 + M4
R3 :	2M3 + M1->M2 + M5
R4 :	M3 + M4 ->2M2
R5 :	M1 + M4->M2
R6 :	M3+ M4 ->M5

Matrice de Stoechiométrie

	R1	R2	R3	R4	R5	R6
M1	-2	1	-1	0	-1	0
M2	0	-1	1	2	1	0
M3	1	0	-2	-1	0	-1
M4	0	1	0	-1	-1	-1
M5	0	0	1	0	0	1

Figure 5 : Exemple jouet d'un ensemble de réactions et de la matrice stœchiométrique correspondante.

3.2.1. Construction et validation des réseaux métaboliques à l'échelle du génome

La construction d'un GSMN est un processus long et chronophage impliquant généralement plusieurs personnes et groupes [72]. La construction peut être divisée en 3 étapes (Fig 6) : la reconstruction automatique, la curation manuelle et enfin la validation du modèle.

L'étape de reconstruction automatique permet de construire une première version du GSMN à partir des informations issues de l'annotation du génome (réalisée au préalable ou disponible dans des bases de données (https://www.ncbi.nlm.nih.gov/genome/)) et l'interrogation de bases de données telles que KEGG [73]. Il existe de nombreux outils permettant de réaliser cette étape de reconstruction automatique, notamment ModelSeed [74], RAVEN [75,76], KBase [77] et CarveMe [78] qui est dédié à la reconstruction de GSMN bactériens et de communautés bactériennes. A noter que certains GSMNs sont reconstruits par analogie avec le réseau d'organismes proches. C'est notamment le cas des réseaux Mouse-GEM, Zebrafish-GEM, Worm-GEM, Rat-GEM, Fruitfly-GEM [79].

Idéalement cette étape de reconstruction automatique est couplée à une étape de reconstruction manuelle visant à améliorer la qualité du modèle en vérifiant la pertinence de tout ou partie des réactions (selon la taille du réseau, le nombre de réaction à vérifier peut être important). Vérifier la pertinence d'une réaction dans le réseau revient notamment à vérifier la validité des coefficients stœchiométriques ainsi que le sens de la réaction [80]. Après chaque étape de curation manuelle, il est nécessaire de vérifier que les modifications réalisées ont permis d'améliorer les capacités du modèle. Cette étape de validation est généralement aidée par des approches *in silico* intégrant des protocoles de contrôle qualité tels qu'un contrôle des capacités fonctionnelles du réseau, l'identification et la correction d'éléments manquants dans la reconstruction (gap-filling en anglais) afin d'assister l'expert lors de la reconstruction manuelle et statuer sur la pertinence des réactions ajoutées au fil des itérations [81,82]. Il convient également de vérifier la faisabilité thermodynamique des réactions du modèle ainsi que la définition de réactions de transport entre les différents compartiments cellulaires.

Les étapes de curation manuelle et de validation sont réalisées successivement un certain nombre de fois afin de converger vers un modèle qui soit le plus qualitatif possible [71]. Par exemple, la construction de Recon2.2, un GSMN de l'humain, s'est appuyé sur Recon2.1 [72]. A partir de Recon2.1, une série de modifications manuelles a été réalisée afin de corriger des limites identifiées par de précédents travaux [73,74] et couplée à une annotation semi-automatique des identifiants de gènes et de métabolites. Certaines voies métaboliques ont été décrites de manière plus détaillée, l'équilibre des masses et des charges entre les réactions a également été corrigé ainsi que de nombreuses autres modifications avec à chaque itération, l'objectif d'améliorer la qualité du modèle.

Selon l'organisme dont on souhaite reconstruire le GSMN, différents types de données et de méthodes *in silico* seront disponibles pour curer et valider le modèle. Par exemple, lors de la curation de GSMNs bactériens, les propriétés phénotypiques de consommation et production de métabolites peuvent être prédites *in silico* via les approches de modélisation sous-contraintes (se reporter à la section suivante) et comparées aux valeurs mesurées *in vitro* à l'aide de culture cellulaires [83].

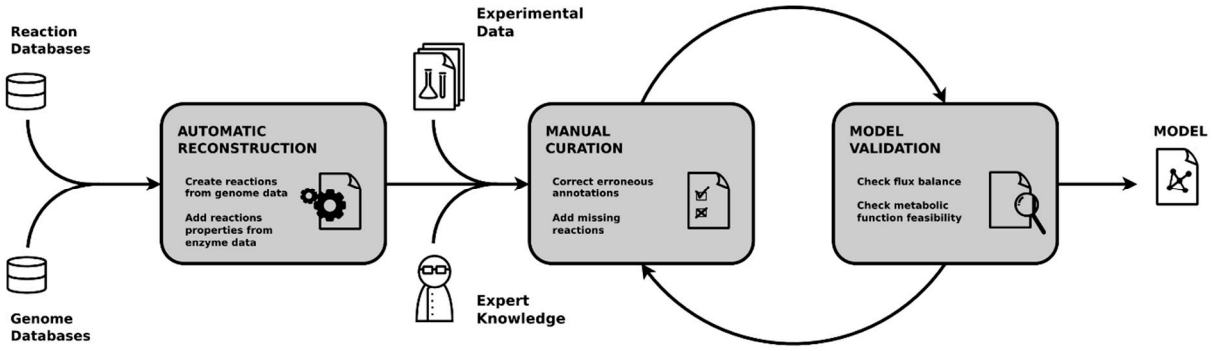

Figure 6 : Schéma du processus de construction et de validation des réseaux métaboliques à l'échelle du génome. *Figure de* [84].

Comme en atteste les statistiques d'évolution du nombre de séquences disponibles dans la base de données GenBank [85,86] au fil des années, le nombre d'organismes dont le génome a été séquencé a très fortement augmenté (3278 génomes animaux uniques dans la base de données GenBank en 2020 [87]) depuis le séquençage complet du premier génome humain. De fait, le nombre d'organismes pour lequel un GSMN a été reconstruit, manuellement ou automatiquement, a de fait lui aussi augmenté, dans une moindre mesure, au fil des années. A l'heure actuelle, on dénombre plus de 3368 modèles mathématiques dans la base de données BioModels [88], qui est l'une des plus grandes bases de données de modèles mathématiques. Bien que d'intérêt, cette base présente quelques limites dans l'annotation (*i.e.* type de modèle, informations concernant la qualité des modèles, …) rendant la recherche de modèle potentiellement fastidieuse. Il convient donc de lire les références bibliographiques des

modèles afin de s'assurer de leurs capacités et de leur qualité avant de pouvoir les utiliser en toute confiance. On peut également citer MetExplore [89] ainsi que la base de données BiGG [90] comme autres sources de GSMNs. La base de données BiGG contient plus de 70 GSMNs publiés avec une attention particulière portée à la standardisation des identifiants via l'introduction d'identifiants BiGG et de liens avec des bases telles que KEGG [73], PubChem [91], etc. MetExplore (285 disponibles publiquement [1]) ne propose pas d'identifiants commun mais met à disposition de l'utilisateur un ensemble d'outils permettant d'explorer, analyser et visualiser, à l'aide de MetExploreViz [70], les GSMNs. La multiplication des GSMNs pour un même organisme nécessite de comparer ces modèles afin de pouvoir choisir le modèle le plus adéquat pour répondre à la question posée. Cependant, la validation et la comparaison des GSMNs n'est pas une tache triviale [92–94]. En effet, l'absence de standardisation dans le choix des identifiants pour les gènes, réactions et métabolites rend difficile la comparaison entre les GSMNs. Des initiatives telles que « GEM Comparison, https://metabolicatlas.org/gems/comparison » vont dans ce sens mais ne proposent pas encore d'outil standardisé permettant de comparer d'autres modèles que Human-GEM et ses variantes. A noter que le mode de curation de Human-GEM [95], qui est l'un des réseaux métaboliques à l'échelle du génome pour l'humain le plus récent est plutôt novateur car il reprend des concepts de l'ingénierie logicielle comme l'intégration continue et le versionnage via l'outil Git qui sont couramment utilisés en développement informatique afin de permettre l'amélioration continue du GSMN par la communauté.

4. Modélisation sous-contraintes

4.1. Principe général

La modélisation sous-contraintes du métabolisme cellulaire consiste à appliquer des contraintes sur les valeurs de flux des réactions pour réduire l'espace de solutions possibles. L'objectif de la modélisation sous-contraintes est de déterminer les flux de l'ensemble des réactions du réseau.

Parmi les contraintes qu'il est possible d'utiliser, la contrainte d'état stationnaire définie par $\frac{d\vec{x}}{dt} = S.v = 0$ est quasi indispensable car elle permet d'une part de s'affranchir du manque de connaissances concernant les capacités catalytiques des enzymes du modèle (ce qui est le cas pour les GSMNs humains) et d'autre part de réduire fortement la taille de l'espace de solutions. La contrainte d'état stationnaire signifie que l'on considère les concentrations de métabolites intracellulaires stables [96,97]. Les informations contenues dans la matrice de stœchiométrie pour chacune des réactions ainsi que les bornes de chacun des flux métaboliques (Fig 7.2) permettent de définir deux contraintes supplémentaires et également fondamentales pour la modélisation sous-contraintes : les contraintes stœchiométriques et les contraintes

thermodynamiques (Fig 7.3). Ces contraintes supplémentaires permettent également de contraindre l'espace de solutions en limitant les flux métaboliques ainsi que les proportions de métabolites consommés et produit par chaque réaction du modèle.

Après avoir défini les métabolites et réactions du modèle et modélisé le lien entre les réactions et les métabolites au travers de la matrice de stœchiométrie, il convient de définir les contraintes du modèle. Comme évoqué ci-dessus, il convient de définir la contrainte de l'équilibre stationnaire (Fig 7.3) qui permet de forcer toute solution calculée à respecter cet équilibre mais également des contraintes thermodynamiques telles que sens des réactions (Fig 7.4).

Ces contraintes classiquement posées pour un problème de modélisation sous-contraintes permettent de définir un espace de solutions qui a la forme d'un cône convexe polyédrique, dénoté C sur la Figure 7.5. Ce cône contient l'ensemble des solutions (distribution de flux métaboliques) respectant les contraintes définies pour le modèle (matrice de stœchiométrie, état stationnaire et irréversibilité des réactions). D'un point de vue biologique, une distribution de flux différente correspond à un état métabolique différent et donc un phénotype différent (la magnitude de la différence pouvant varier).

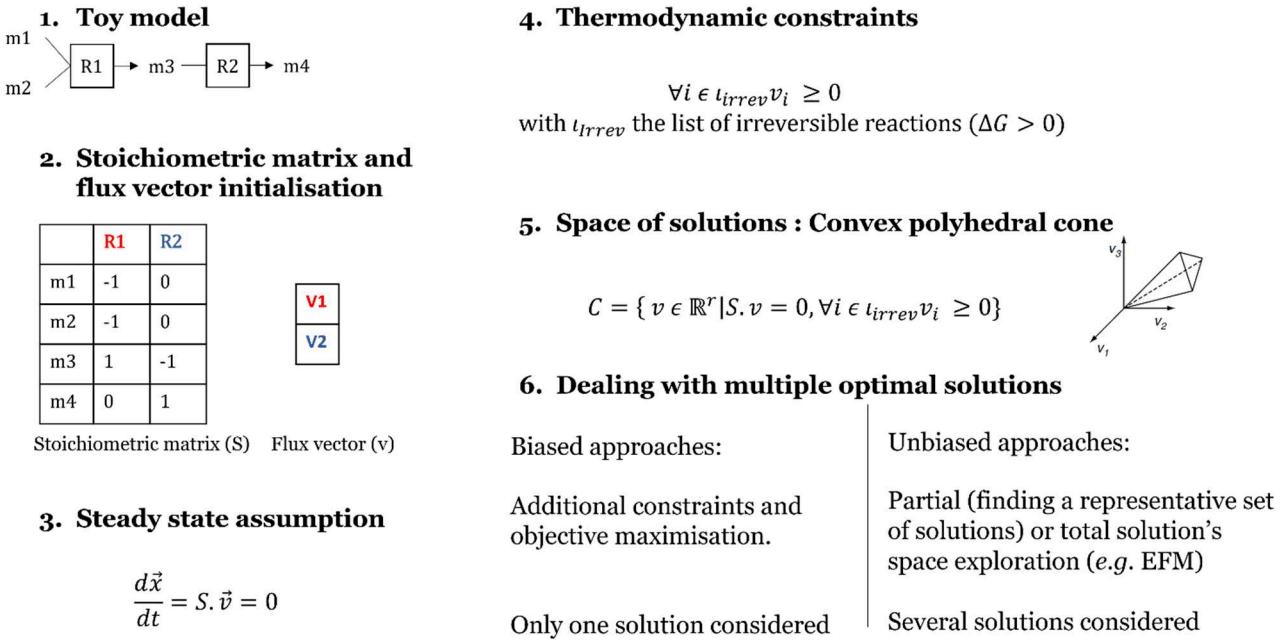

Figure 7: Exemple jouet de la construction d'un modèle sous-contraintes.

Cela signifie également qu'il n'existe pas une seule solution satisfaisant exactement les contraintes définies précédemment. Puisque chacune de ces solutions représente un phénotype différent, l'analyse d'un seul phénotype choisi au hasard parmi tous ceux existant dans l'espace de solutions induirait un biais dans les interprétations réalisées à partir de l'analyse de la distribution de flux correspondant à ce phénotype. Au cours des prochaines

sections, nous allons nous d'abord nous intéresser à ce problème et comment il est possible de réduire la taille de l'espace de solutions par l'ajout de nouvelles contraintes. Nous nous intéresserons ensuite aux différentes approches de modélisation sous-contraintes et la manière dont ces approches traite l'espace de solutions. Nous allons diviser ces méthodes en deux groupes : (1) les approches biaisées et les (2) approches non biaisées

4.2. Espace de solutions et problème insuffisamment contraint

4.2.1. Principe général

L'impossibilité à trouver une solution unique au problème mathématique posé lors de la modélisation sous-contraintes est due au manque de contraintes par rapport à la complexité du réseau métabolique (*i.e.* le nombre de réactions qui définit le nombre de flux métaboliques à estimer et donc le nombre d'inconnues à déterminer) [98]. La taille de l'espace de solutions dépend de la précision ainsi que du nombre de contraintes définies dans le problème mathématique (LP ou MILP selon la méthode choisie) (Fig 8).

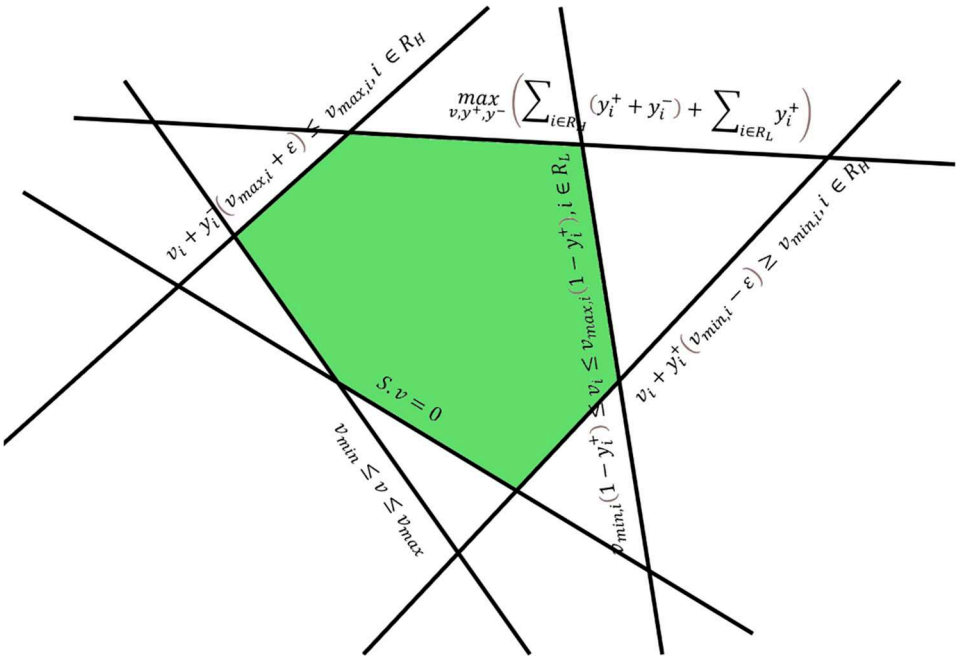

Figure 8 : Schéma de l'espace de solution du problème de modélisation sous-contraintes généré par iMAT*. Les lignes ainsi que les équations et inéquations associées correspondent aux contraintes linéaires définies par les auteurs de* [99,100], la zone en vert correspond à l'espace de solutions possibles.

Le nombre de solutions possibles est souvent très important voire infini, ce qui rend la caractérisation de l'espace de solutions complexe [98]. Comme nous allons l'aborder au cours des prochaines sections, ces solutions alternatives ne sont que rarement prises en compte. Dans un premier temps, il peut être intéressant de réduire la taille de l'espace de solutions (et donc le nombre de solutions alternatives) en ajoutant des contraintes supplémentaires au problème de modélisation sous-contraintes déjà crée.

4.2.2. Réduction de l'espace de solutions par l'ajout de nouvelles contraintes

L'intégration de plusieurs types de données omiques au cours de la modélisation permet de définir de nouvelles contraintes et ainsi de réduire la taille de l'espace de solutions. Différents types de données omiques peuvent être utilisés, idéalement toutes les données omiques correspondant à une condition auront été générées sur les mêmes échantillons afin de limiter les effets lots de cellule et conditions d'expérimentation. Par exemple dans [101] les auteurs ont combiné des données transcriptomiques avec des données de métabolomique en intégrant de nouvelles contraintes à iMAT. iMAT est un algorithme de reconstruction de réseaux condition spécifiques recherchant la meilleure adéquation possible entre des données transcriptomique et la topologie d'un GSMN, nous le décrirons en détail dans une prochaine section. Ces nouvelles contraintes veillent à ce que les solutions obtenues par iMAT soient non seulement en adéquation avec les données transcriptomiques mais également que tous les métabolites détectés puissent être produits par les modèles condition-spécifique simulés. Les auteurs de [102] ont développé une méthode d'intégration de données protéomiques et métabolomiques continues qui cherche à optimiser l'adéquation entre les valeurs mesurées de protéomique et de métabolomique et les flux estimés par le modèle. Enfin, il est également possible de contraindre les flux des réactions d'échange à partir de données d'exométabolomique (*i.e.* des mesures de la concentration extracellulaire de certains métabolites) [103].

Comme décrit précédemment ajouter de nouvelles contraintes au travers de l'intégration de nouvelles données omiques permet de réduire la taille de l'espace de solution et donc réduire la complexité et le risque d'erreur lors de l'analyse (Fig 8). Sur l'exemple jouet présenté en figure 9, les zones contenant des solutions n'étant plus considérées comme optimales après ajout de deux nouvelles contraintes (nommées Metabolomics constraints 1 and 2) ne font plus partie de l'espace de solution, ce qui implique une réduction de ce dernier et ainsi une meilleure caractérisation de l'état du métabolisme cellulaire pour la condition modélisée.

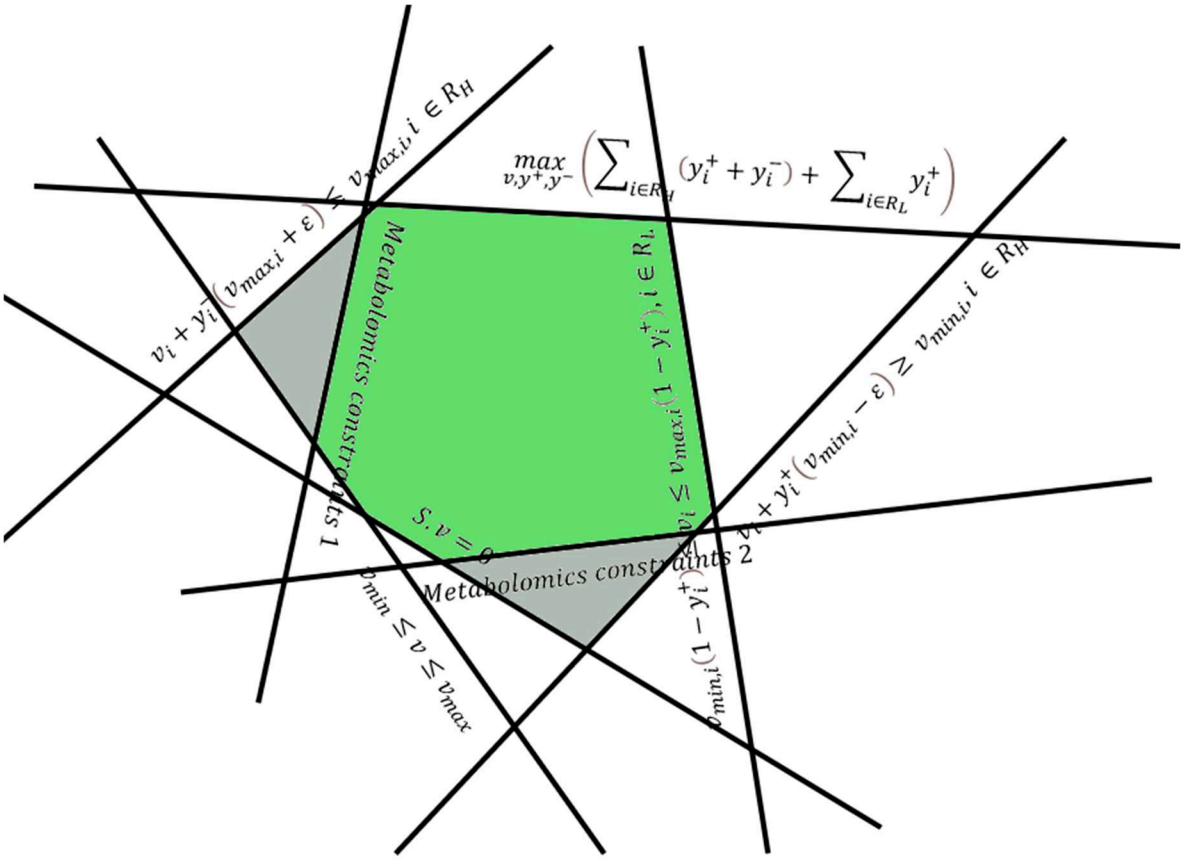

Figure 9 : Schéma de l'espace de solution initial de construit par iMAT mis à jour par l'ajout de deux contraintes d'adéquation avec des données de métabolomique. *La zone en vert correspond à l'espace de solution prenant en compte l'ensemble des contraintes (iMAT + contraintes d'adéquation avec des données de métabolomique) et les zones en gris correspondent à des solutions respectant les contraintes d'iMAT mais pas les contraintes d'iMAT combinées aux contraintes d'adéquation avec les données de métabolomique.*

4.3. Approches biaisées : optimisation d'une fonction objective

Au cours de cette section, nous allons aborder quelques approches de modélisation sous-contraintes traitant le problème des solutions alternatives de manière dites « biaisées » [104]. Ces approches n'explorent pas l'espace de solutions mais vont plutôt chercher une solution maximisant une fonction objective. Elles sont dites biaisées car il peut exister un très grand nombre de solutions alternatives optimales maximisant la fonction objective définie par l'algorithme.

Dans le cadre d'un problème d'optimisation linéaire, la fonction objective correspond à la fonction que le solveur cherchera à maximiser ou minimiser. La définition de cette fonction objective est donc importante puisque c'est cette fonction qui sert de « contrôle qualité » et déterminer si une solution est optimale vis-à-vis des contraintes définies pour le modèle ou non. Au cours des deux prochaines sections, nous allons nous intéresser à deux types de fonction objectives : (1) Les fonctions objectives visant à optimiser un objectif biologique et (2) les fonctions objectives visant à optimiser l'adéquation avec un phénotype représenté par des données omiques.

4.3.1. Optimisation d'un objectif biologique

Le principe général de l'optimisation d'une fonction objective biologique s'appuie sur l'hypothèse qu'une cellule cherche à réaliser un objectif métabolique particulier. Les approches maximisant cette fonction objective partent donc du principe que la cellule cherche à optimiser son métabolisme afin de maximiser la réalisation de cet objectif.

La FBA [105] est la méthode d'optimisation d'une fonction objective la plus utilisée. Le problème défini pour la FBA est le suivant :

$$\max_{v} f(v) = c^t v$$
$$s.t.$$
$$v \in C$$
$$l \leq v \leq u$$

Avec v, un vecteur de flux appartenant à l'espace de solutions C, dont les flux respectent les bornes inférieures et supérieures (l et u, respectivement) définies pour chacune des réactions du modèle et maximisant la fonction linéaire représentée par $f(v) = c^t v$.

Dans le cas des micro-organismes, cet objectif est souvent de maximiser la croissance et la multiplication cellulaire [106]. La fonction linéaire à maximiser sera généralement une fonction de biomasse inclue dans le modèle et représentant cet objectif de maximisation de la croissance de l'organisme. Par exemple, la fonction de biomasse de l'un des modèles les plus complet pour *Escherichia* Coli [107], contient l'ensemble des métabolites nécessaires à la constitution du microorganisme et est directement liée au taux de croissance cellulaire. Maximiser cette réaction revient donc à maximiser la croissance d'E.Coli et donc indirectement la production de biomasse. L'ensemble des contraintes seront donc résolues sous la forme d'un problème linéaire par un solveur mathématique capable de résoudre ce type de problème tel que CPLEX ou Gurobi par exemple. Après résolution du problème linéaire, une distribution de flux est obtenue. Cette distribution de flux correspond à l'ensemble des valeurs de flux prédites pour le modèle et respectant les contraintes établies précédemment. Il est important de noter que cette distribution de flux n'est pas unique et qu'il peut exister un grand nombre de distribution de flux alternatives.

Le choix de la fonction objective à maximiser dépend principalement de l'objectif de l'étude. Il est par exemple possible de chercher la distribution de flux minimisant la production d'ATP (optimisation de l'efficacité énergétique), minimiser la consommation de nutriments (optimiser la consommation du milieu de culture), maximiser la production d'un métabolite d'intérêt (particulièrement intéressant pour l'ingénierie métabolique) ou encore maximiser

deux fonctions objectives telles que la biomasse et la production d'un métabolite d'intérêt [108].

Depuis la popularisation des analyses de flux *in silico* avec la FBA, de nombreuses approches alternatives ont été publiées afin de répondre à certaines limites de la FBA ou étendre ses capacités. Par exemple, la parsimoniousFBA [109] (pFBA) vise à apporter une réponse au problème des solutions optimales alternatives en minimisant la somme des flux alors que la dynamicFBA [109] (dFBA) permet une étude dynamique (*i.e.* au cours du temps) des flux et des métabolites produits/consommés.

Enfin, il existe des approches « hybrides » telles que GIMME [110] qui recherche une distribution de flux maximisant un objectif biologique (*e.g.* la production de biomasse) tout en minimisant l'utilisation de réactions dites inactives. Ces réactions sont définies inactives car leur intensité d'expression est inférieure à un seuil défini par l'utilisateur généralement à partir de données omiques.

4.3.2. Optimisation d'un objectif d'adéquation avec des données omiques

4.3.2.1. Principe général

Définir une fonction objective pour des cellules qui ne sont plus en phase de croissance et par conséquent dont l'objectif n'est pas la création de biomasse peut s'avérer difficile et partiel. En effet, il existe de nombreux types cellulaires ne réalisant pas une unique fonction métabolique que l'on pourrait maximiser/minimiser mais plutôt différentes capacités métaboliques. Par exemple, les hépatocytes sont capables de détoxifier des molécules exogènes présentes dans la circulation sanguine mais jouent également un rôle de régulation des sucres et acides gras en stockant ces molécules et en les libérant sous des formes assimilables lorsque cela est nécessaire [111]. Il est donc difficile d'identifier une fonction à maximiser parmi cet ensemble de capacités métaboliques dont l'activation est spécifique du contexte cellulaire. C'est notamment pour répondre à cette difficulté que des approches permettant de décrire des flux métaboliques spécifiques d'un phénotype sans avoir recours à l'optimisation d'une fonction objective biologique ont été développées. Ces approches permettent notamment de reconstruire des réseaux métaboliques spécifiques d'une condition biologique.

4.3.2.2. Construction de réseaux condition-spécifiques par l'intégration de données omiques

Comme nous avons pu l'aborder précédemment, les GSMNs sont des réseaux métaboliques représentatifs de l'ensemble des capacités métaboliques d'un organisme. Par définition, ces réseaux métaboliques sont donc génériques alors que le métabolisme de chaque type cellulaire est spécifique et capable de réaliser seulement une sous-partie de l'ensemble des fonctions

métaboliques décrites dans le GSMN. La spécificité métabolique peut être encore plus importante lorsqu'il s'agit d'évaluer la réponse d'un type cellulaire particulier (*e.g.* les hépatocytes) lors d'une exposition à un composé chimique précis ou à des conditions environnementales données.

En intégrant des données omiques sous formes de nouvelles contraintes, il est possible de construire des réseaux métaboliques spécifiques de la condition biologique à laquelle ces données omiques ont été générées (type cellulaire, exposition à un xénobiotique, pathologie, etc). Un réseau condition-spécifique est donc un sous-réseau du GSMN qui n'est plus représentatif de l'ensemble des capacités métaboliques connues d'un organisme mais représentatif de l'état du métabolisme cellulaire dans une condition biologique donnée. Il existe plus d'une dizaine d'algorithmes permettant de construire des réseaux condition-spécifiques en intégrant des données omiques sous la forme de contraintes. Parmi ces algorithmes, on distingue différents types d'objectif à optimiser. Certaines méthodes comme MBA [112] , FASTCORE [113] et mCADRE [114] cherchent à minimiser le réseau métabolique final en identifiant des réactions centrales, devant être systématiquement actives et en retirant les autres réactions si possibles [115]. Minimiser le réseau métabolique revient à chercher un réseau contenant le moins possible de réactions tout en respectant les contraintes définies dans le problème à résoudre.

D'autres méthodes comme INIT [116] et iMAT [99,100] recherchent le meilleur consensus entre le fait de retirer les réactions considérées comme étant inactives au regard des données omiques et le fait de conserver les réactions considérées comme actives au regard des données omiques. Pour la suite de cette section, nous allons nous focaliser sur le fonctionnement de iMAT puisque la méthode de modélisation employée au cours de nos travaux (DEXOM) est basée sur iMAT. iMAT est une approche particulièrement adaptée à la construction de réseaux condition-spécifiques à partir de données transcriptomiques [99,117] sans pour autant rechercher un réseau minimal comme cela peut être le cas avec les approches telles que MBA ou FASTCORE.

iMAT pour « integrative Metabolic Analysis Tool », est un algorithme permettant la création de réseaux condition-spécifiques à partir de données transcriptomiques (ou protéomiques) et d'un GSMN initial. iMAT prédit un ensemble de réactions actives (*i.e* ayant un flux métabolique non nul) et inactives (*i.e.* ayant un flux nul ou quasi nul) en maximisant l'adéquation entre les réactions actives/inactives d'après les données omiques et les

réactions/actives d'après la topologie du réseau (qui contient notamment les informations de connexion entre les réactions).

En retirant toutes les réactions prédites comme inactives par iMAT il est possible de reconstruire le réseau métabolique spécifique de la condition biologique modélisée.

Afin de pouvoir intégrer les données transcriptomiques (ou protéomiques) sous la forme de contraintes pour la modélisation, il convient de transformer l'information transcriptomique continue en une liste de réactions actives/inactives (d'après ces données transcriptomiques). Cela peut être réalisé au travers des GPRs. Les GPRs étant des règles booléennes, il est cependant nécessaire de binariser l'information transcriptomique (comme détaillé dans le chapitre 2) afin de pouvoir déterminer si les gènes essentiels à la synthèse de l'enzyme sont actifs et par extension si la réaction est active d'après les données transcriptomiques.

Considérons un modèle jouet (Fig 10) constitué de 9 réactions biochimiques et 10 métabolites ainsi qu'un jeu de données transcriptomiques binarisées (gènes fortement ou faiblement exprimés) constitué de 6 gènes.

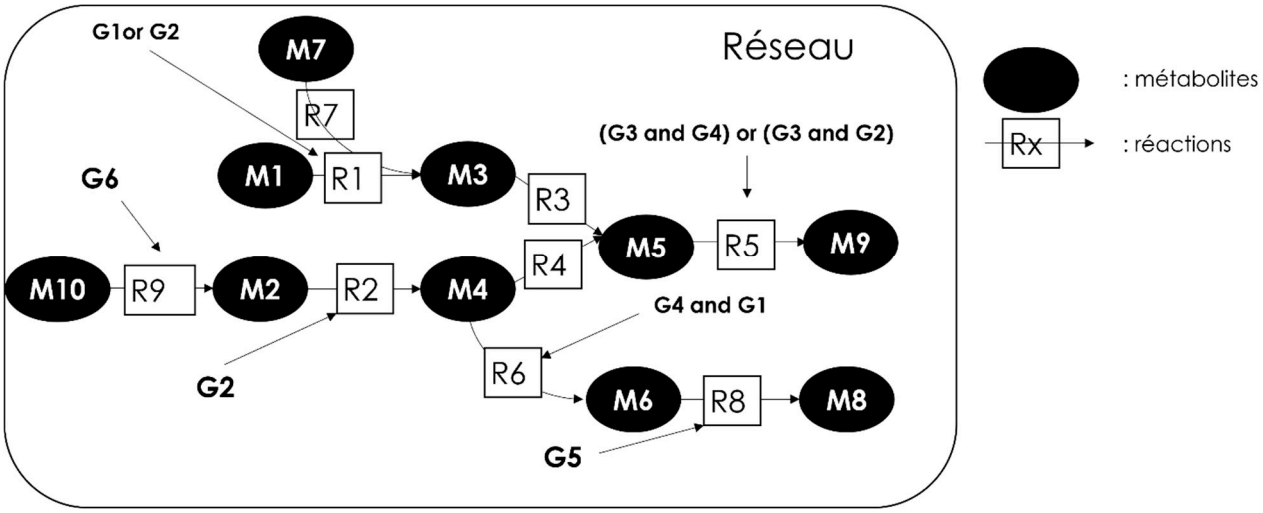

Figure 10 : Schéma du réseau jouet. *Ce réseau est constitué de 9 réactions biochimiques et 10 métabolites. 6 des 9 réactions ont une GPR et peuvent donc être contraintes par les données transcriptomiques. Les métabolites sont représentés par des ellipses et les réactions par des carrés. Le sens des flèches connectant passant par les réactions donnent la direction de la réaction. Chaque GPR est associée à la réaction correspondante par une flèche.*

	Equation	Gène-Protéine-Réaction (GPR)	Statut d'activité selon données transcriptomiques
R1	**M1 -> M3**	**G1 OR G2**	**ACTIVE**
R2	**M2 -> M4**	**G2**	**INACTIVE**
R3	**M3 -> M5**	**Passive**	**INCONNU**
R4	**M4 -> M5**	**Passive**	**INCONNU**
R5	**M5 -> M9**	**(G3 AND G4) OR (G3 AND G2)**	**ACTIVE**
R6	**M4 -> M6**	**G4 AND G1**	**ACTIVE**
R7	**M7 -> M3**	**Passive**	**INCONNU**
R8	**M6 -> M8**	**G5**	**INACTIVE**
R9	**M10 -> M2**	**G6**	**ACTIVE**

Tableau 1. Tableau des équations et GPRs correspondant aux réactions du modèle jouet.

Les gènes G1, G3, G4 et G6 sont des gènes fortement exprimés dans la condition à modéliser alors que les gènes G2 et G5 sont des gènes faiblement exprimés dans la condition à modéliser. En intégrant ces données transcriptomiques à ce réseau métabolique jouet par les GPRs, il est possible d'identifier 4 réactions actives selon les données transcriptomiques (R1, R5, R6 et R9, visibles sur le Tableau 1), 3 réactions pour lesquelles le statut d'activité selon les données transcriptomiques est inconnu (R3, R4 et R7, visibles sur le Tableau 1) car ces réactions n'ont pas de GPRs (il s'agit en général de réactions passives ou spontanées, n'impliquant pas d'enzyme) et 2 réactions inactives selon les données transcriptomiques (R2 et R8, visibles sur le Tableau 1).

La seconde étape consiste à construire le problème linéaire à résoudre. Le problème construit par iMAT est un problème du type MILP (Mixed-Integer Linear Programming). Les problèmes d'optimisation linéaire en nombre entiers (comme le MILP) sont des problèmes décrits par une fonction de cout, de contraintes linéaires et de variables entières. C'est notamment cette nature entière des variables qui distingue le MILP du LP classiquement utilisé par la FBA par

exemple. Dans le cas iMAT, les variables sont effectivement des entiers : 1 si la réaction est active et -1 si elle est inactive.

Le MILP construit par iMAT est le suivant :

$$\max_{v,y^+,y^-} \left(\sum_{i\in R_H}(y_i^+ + y_i^-) + \sum_{i\in R_L} y_i^+\right) \quad (1)$$

$s.t$

$$S.v = 0 \quad (2)$$

$$v_{min} \leq v \leq v_{max} \quad (3)$$

$$v_i + y_i^+\left(v_{min,i} - \varepsilon\right) \geq v_{min,i}, i \in R_H \quad (4)$$

$$v_i + y_i^-\left(v_{max,i} + \varepsilon\right) \leq v_{max,i}, i \in R_H \quad (5)$$

$$v_{min,i}(1 - y_i^+) \leq v_i \leq v_{max,i}(1 - y_i^+), i \in R_L \quad (6)$$

$v \in R^m$

$y_i^+, y_i^- \in [0,1]$

Avec v, le vecteur de flux de chaque réaction et S la matrice de stœchiométrie du modèle. R_H correspond à la liste des réactions identifiées comme actives (*i.e.* ayant un flux) selon les données transcriptomiques et R_L correspond à la liste des réactions identifiées comme inactives (*i.e.* n'ayant pas de flux) selon les données transcriptomiques. y_i^+et y_i^- sont des variables booléennes (*i.e.* ayant deux valeurs possibles, 0 ou 1) indiquant si la réaction est active ou inactive dans chacune des directions. Les équations 2 à 6 correspondent aux contraintes du modèle. L'équation (2) représente la contrainte de conservation des masses à l'état d'équilibre. L'équation (3) contraint la solution à respecter les bornes de flux définies dans le modèle. Ces bornes indiquent les capacités minimales et maximales d'une réaction. Les équations (4) et (5) fixent les contraintes à respecter pour prédire une réaction du groupe R_H comme étant active. Une réaction du groupe R_H sera prédite comme active si elle a un flux positif supérieur au seuil ε ou un flux négatif inférieur à -ε. L'équation (6) fixe les contraintes à respecter pour considérer une réaction du groupe R_L comme étant inactive. Une réaction de ce groupe est considérée comme inactive si elle a flux métabolique nul. A noter que pour l'équation (6), y_i^+ vaut 1 lorsque la réaction est inactive, indiquant une correspondance entre

le statut inactif défini par les données transcriptomiques et le statut inactif prédit par iMAT. Enfin, l'équation (1) correspond à la fonction maximisée lors de la résolution du MILP. Cette fonction cherche à maximiser l'adéquation entre les réactions considérées actives selon les données transcriptomiques R_H et celles considérées actives selon les contraintes du modèles associées au flux et en maximisant l'adéquation entre les réactions considérées inactives selon les données transcriptomiques $R_L$et celles considérées inactives selon les contraintes associées aux flux.

La troisième étape revient à résoudre le MILP créé précédemment. Cette étape est réalisée par un solveur mathématique (*e.g.* CPLEX ou Gurobi) permettant de rechercher une distribution de flux optimale, c'est-à-dire maximisant la fonction objective (ici l'adéquation entre les données transcriptomiques et le réseau métabolique) tout en respectant les contraintes définies dans le MILP.

Pour le modèle jouet présenté en Tableau 1, une solution optimale serait la suivante (Fig 11) :

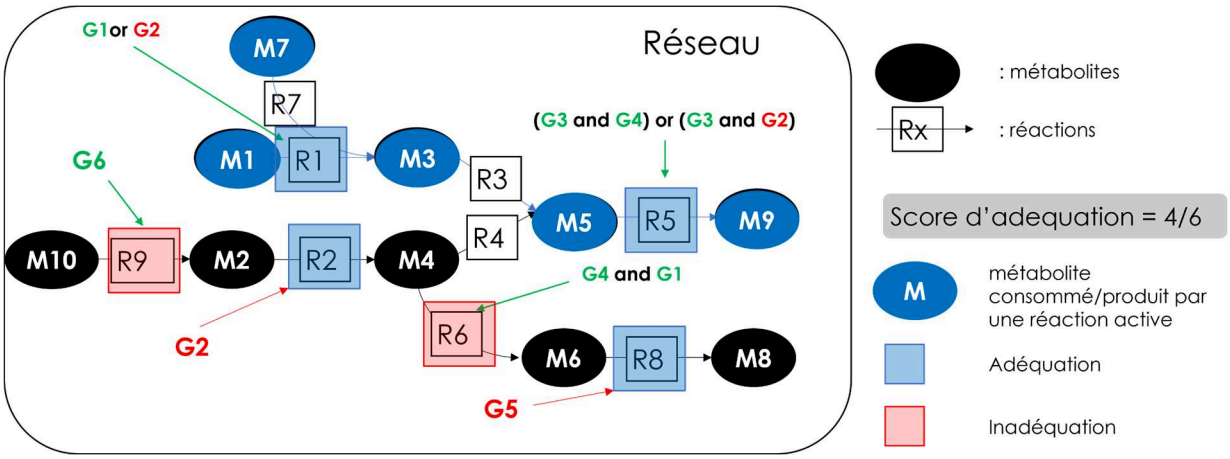

Figure 11: Exemple d'une solution maximisant l'adéquation entre les données transcriptomiques et le réseau dans le cadre de l'exemple jouet*. La solution optimale consiste en l'activation des réactions (flèches bleues) R7, R1, R3 et R5 et en l'inactivation des réactions (flèches noires) R9, R2, R6 et R8. Les métabolites sont représentés par des ellipses et les réactions par des carrés. Le sens des flèches connectant passant par les réactions donnent la direction de la réaction. Chaque GPR est associée à la réaction correspondante par une flèche. Les métabolites colorés en bleu sont des métabolites associés à des réactions actives, donc des métabolites consommés ou produits dans ce sous-réseau. Un cadre bleu sur une réaction signifie que l'activation/inactivation de cette réaction par iMAT est en adéquation avec les données transcriptomiques alors qu'un cadre rouge sur une réaction signifie l'inverse, c'est-à-dire que l'activation/inactivation de cette réaction n'est pas en adéquation avec les données transcriptomiques.*

Une solution du MILP est un résultat possible du problème de modélisation sous-contraintes construit par iMAT, c'est-à-dire une liste de réactions prédites actives, représentant un sous-réseau condition-spécifique avec un score d'adéquation optimal selon les paramètres définis par l'utilisateur.

A la fin de l'exécution d'iMAT, on retire les réactions prédites comme étant inactives (flèches noires sur la Fig 11) pour la solution retournée afin d'avoir le réseau métabolique spécifique de

la condition biologique en adéquation avec les données transcriptomiques ayant servi à contraindre le modèle. Cependant, la solution optimale obtenue par iMAT n'est pas unique. En effet avec les mêmes contraintes et les mêmes données omiques il est possible de trouver une autre solution (Fig 12) (*e.g.* un ensemble de réactions métaboliques prédites actives/inactives satisfaisant les contraintes et atteignant un score optimal).

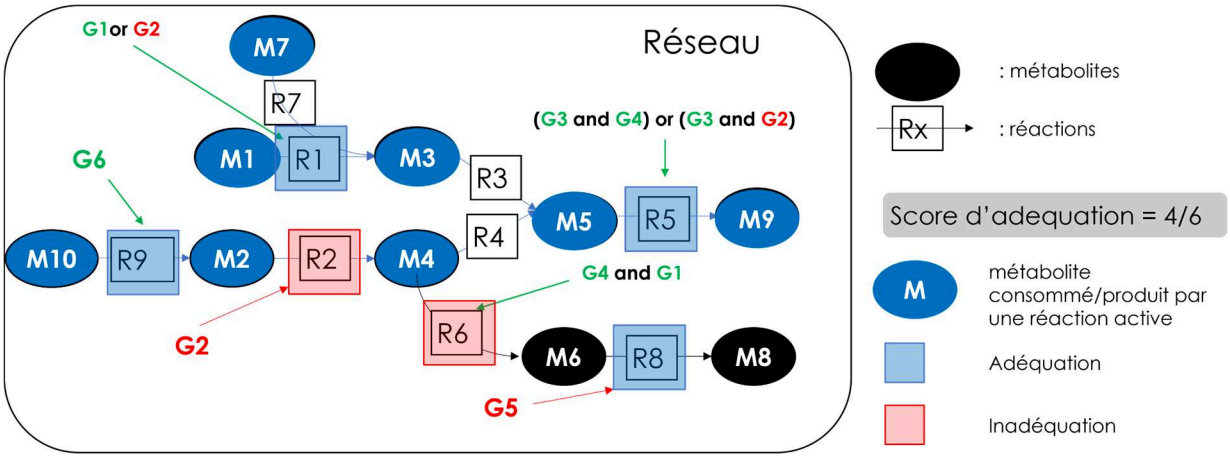

Figure 12 : Exemple d'une solution alternative au MILP construit pour l'exemple jouet. *La solution optimale consiste en l'activation des réactions (flèches bleues) R7, R1, R3, R5, R4, R2 et R9 et en l'inactivation des réactions (flèches noires) R6 et R8. Les métabolites sont représentés par des ellipses et les réactions par des carrés. Le sens des flèches connectant passant par les réactions donnent la direction de la réaction. Chaque GPR est associée à la réaction correspondante par une flèche. Les métabolites colorés en bleu sont des métabolites associés à des réactions actives, donc des métabolites consommés ou produits dans ce sous-réseau. Un cadre bleu sur une réaction signifie que l'activation/inactivation de cette réaction par iMAT est en adéquation avec les données transcriptomiques alors qu'un cadre rouge sur une réaction signifie l'inverse, c'est-à-dire que l'activation/inactivation de cette réaction n'est pas en adéquation avec les données transcriptomiques.*

L'existence de plusieurs solutions optimales (*i.e.* ayant le même score mais différentes réactions actives/inactives) est notamment due à un manque de contraintes lors de la modélisation et a été évoqué par les auteurs de iMAT [100]. Cependant, l'obtention d'un ensemble de solutions alternatives représentatives de la diversité des solutions possibles n'est pas un problème trivial. Nous allons aborder ce problème ainsi que les approches permettant de prendre en compte cette diversité de solutions possibles dans la section suivante.

4.4. Approches non biaisées : exploration de l'espace de solutions

Au cours de cette section, nous allons aborder différentes approches permettant d'explorer l'espace de solutions alternatives existantes (EFM ou énumération via l'approche « Integer-cut ») ainsi que des approches permettant d'énumérer un ensemble de solutions alternatives représentatif de la diversité des solutions existantes dans l'espace de solutions. Ces approches sont dites « non biaisées » car elles prennent en compte le problème des solutions alternatives.

4.4.1. Méthodes d'exploration complète de l'espace de solution

Les méthodes d'exploration complète de l'espace de solutions sont capables, en théorie, de trouver l'ensemble des solutions alternatives possibles. Ces approches n'ont pas été utilisées au cours de nos travaux car elles ne sont pas adaptées à des espaces de solutions de grande taille qui comme nous l'avons mentionné précédemment peuvent donner lieu à un nombre infini de solutions alternatives. Nous allons tout de même décrire l'approche des « Elementary Flux Modes » (EFM) qui est approche permettant de décomposer le réseau métabolique en sous-unité minimales fonctionnelles [118] ainsi que l'approche d'énumération par integer-cut.

Le calcul des EFM permet la décomposition du réseau métabolique en sous-unité minimales fonctionnelles. Décomposer le réseau en EFM permet d'identifier toutes les combinaisons de sous-unité minimales fonctionnelles permettant de réaliser une fonction métabolique et ainsi identifier des points faibles (un EFM retrouvée dans de nombreuses combinaisons, donc jouant un rôle central) ou à l'inverse des points forts dans le réseau. Chaque EFM est dit « non-décomposable », car chaque EFM est un ensemble minimal de voies contenant le moins de réactions possibles pour constituer une unité fonctionnelle. De fait, retirer une réaction d'un EFM lui fait perdre sa nature d'unité fonctionnelle [119]. Ce découpage en EFM permet de représenter chacune des distributions de flux comme une superposition d'EFMs. Chaque réaction dont le flux est à zéro implique que chaque EFM comprenant cette réaction devra également avoir un flux de zéro pour cette réaction. L'analyse des EFMs est un outil couramment utilisé dans le domaine de l'ingénierie métabolique [118]. Par exemple, les EFMs peuvent permettre de déterminer si une cellule est capable de produire un métabolite d'intérêt à partir d'un substrat donné. Il est également possible de calculer l'efficacité de cette réaction et d'identifier le mode ayant la meilleure efficacité. Les EFMs permettent également de retirer rapidement toutes les fonctionnalités (représentées par des EFMs) associées à une réaction et donc en retirant itérativement des réactions, de construire le réseau minimal capable de réaliser une fonction d'intérêt [118]. Il existe divers algorithmes permettant de calculer les EFMs d'un réseau métabolique, tel que Efmtools [120] et Meta-tool [121]. Efmtools a été décrit comme étant l'outil le plus performant pour calculer les EFMs d'un réseau métabolique mais nécessite de très grande quantité de RAM (mémoire vive) : 150 Gb pour le réseau central d'E.Coli [122], constitué d'une centaine de réactions [118]. Le calcul d'EFMs pour des GSMNs constitués de plusieurs milliers de réactions semble donc difficilement réalisable avec ces approches.

Le principe de l'énumération par « Integer-cut » est relativement simple. Cette approche d'énumération résout le problème d'optimisation linaire itérativement en ajoutant une nouvelle contrainte au problème pour chaque solution alternative calculée. Cette contrainte

oblige le solveur à trouver une nouvelle solution alternative qui ne soit pas celle qui vient d'être trouvée. Cette approche ne s'arrêtera donc qu'à partir du moment ou plus aucune solution satisfaisant l'ensemble des contraintes (contraintes initiales + contrainte ajoutées pour chacune des solutions déjà calculées) ne pourra être trouvée. Cependant, lors de ces travaux (de d'une manière générale avec les GSMNs constitués de plusieurs milliers de réactions), l'espace de solution est de très grande taille et le nombre de solutions alternatives également très grand voir infini, il n'est donc pas envisageable d'utiliser cette approche pour explorer l'espace de solutions. Etant données que les méthodes d'exploration complète de l'espace de solutions ne sont pas adaptées lorsque le nombre de solutions alternatives est extrêmement grand voir même infini, nous nous sommes intéressées aux méthodes d'exploration partielle.

4.4.2. Méthode d'exploration partielle de l'espace de solutions

4.4.2.1. Principe général

Les méthodes d'exploration partielle de l'espace de solutions correspondent à des approches permettant d'énumérer un ensemble de solutions alternatives qui, idéalement, sont représentatives de l'ensemble des solutions existantes dans l'espace de solutions.

Parmi ces méthodes, nous allons nous intéresser à deux approches différentes pour répondre au problème de l'énumération. Il est notamment possible d'estimer les variations de flux à partir d'une première distribution de flux prédite par l'algorithme de modélisation sous-contraintes. C'est notamment cette approche qui a été mise en place par les auteurs de iMAT++ [123], une évolution de l'algorithme iMAT introduite précédemment. iMAT++ a pour objectif de maximiser l'adéquation entre des données transcriptomiques et la topologie du réseau tout en corrigeant certaines limites identifiées pour iMAT [123]. A la fin de l'exécution de iMAT++, une distribution de flux optimale est obtenue. Cependant, comme mentionné précédemment, cette distribution de flux n'est pas unique puisque des réactions ne portant pas ou peu de flux dans la distribution de flux optimale retournée par iMAT++ peuvent porter des flux dans une distribution de flux optimale alternative. Afin de prendre en compte cette problématique, les auteurs ont estimé les bornes inférieures et supérieures dans l'espace de solutions pour chacune des réactions à l'aide d'une Flux Variability Analysis (FVA) tout en conservant les contraintes définies au cours des étapes précédentes d'iMAT++. La FVA est une approche de modélisation sous-contraintes dont l'objectif est de successivement maximiser puis minimiser le flux d'une réaction dans le réseau (les contraintes existantes dans le modèle étant conservée lors de la FVA) afin de déterminer la valeur minimale et maximale de flux possible pour chacune des réactions du réseau au regard des contraintes classiquement définies dans un problème de modélisation sous-contraintes (état d'équilibre, stœchiométrie et thermodynamique) et d'éventuelles contraintes supplémentaires (définies par iMAT ++ par

exemple). Ainsi, une réaction n'ayant pas de flux dans la distribution de flux optimale obtenue avec iMAT++, mais ayant une capacité à porter un flux selon la FVA (*e.g.* un écart entre les bornes inférieures et supérieures non nul) sera intégrée à une distribution de flux alternative représentant les réactions pouvant porter un flux non nul dans l'espace de solutions défini par le problème généré par iMAT++. Les auteurs de iMAT++ se servent donc de la FVA pour identifier les réactions dont le statut actif/inactif peut varier au sein de l'espace de solutions. Il existe également des approches dites de « flux sampling » (échantillonnage de flux en français) dont l'objectif est d'explorer l'ensemble des flux possibles en générant des distributions de probabilités de flux pour chacune des réactions du modèle [124]. L'échantillonnage de flux permet donc d'une part d'estimer les bornes minimales et maximales du flux de chaque réaction mais également d'estimer les valeurs de flux possibles entre ces bornes ainsi que leur probabilité. Il existe plusieurs algorithmes permettant d'échantillonner les valeurs de flux dans un modèle : CHRR, ACHR et OPTGP [124]. Etant donné que l'échantillonnage de flux est une approche pouvant être couteuse en temps de calcul, le choix d'un algorithme computationnellement efficace est important [124].

D'autres approches, telles que $RegREx_{AOS}$ [125], explorent l'espace de solutions en générant aléatoirement un vecteur de flux dont les bornes sont définies par une analyse de variabilité des flux et cherchent ensuite à trouver un vecteur de flux appartenant à l'espace de solutions et étant le plus proche possible du vecteur aléatoire. L'exploration de l'espace de solutions s'arrête une fois que le nombre de solutions alternatives à trouver défini par l'utilisateur a été atteint. Cependant, considérer les valeurs quantitatives de flux peut sembler être une hypothèse forte. En effet, bien que les données transcriptomiques puissent servir de proxy pour la reconstruction de réseaux métaboliques condition-spécifiques [126–128], la corrélation entre les intensités d'expression et les flux métaboliques est controversée [129–131]. Afin de limiter de potentielles erreurs d'interprétation, il convient d'avoir une approche plus conservative. Pour cela, les valeurs de flux peuvent être catégorisées afin de limiter le risque d'erreur d'interprétation.

D'autres approches permettent l'exploration de l'espace de solutions en ne considérant que le statut actif/inactif des réactions [101,132]. EXAMO [132], est une approche qui explore l'espace de solutions en bloquant itérativement chacune des réactions du modèle une à une en recherchant à chaque fois s'il existe une solution optimale respectant les contraintes initiales et en ayant la réaction bloquée. Les réactions non réversibles sont également forcées dans le sens direct et les réactions réversibles sont forcées à la fois dans le sens direct et indirect. Après cette phase d'exploration de l'espace de solutions, Rossel *et al.* ont identifié des réactions étant prédites comme actives dans toutes les solutions alternatives ainsi que des réactions prédites comme étant inactives dans toutes les solutions alternatives. Cette identification leur

permettant d'identifier un ensemble de réactions « core » (*i.e.* toujours actives) autour desquelles construire un réseau métabolique consensus minimal à l'aide de l'approche MBA [112]. L'approche développée par [101], utilise une approche similaire à celle développée par Rossel *et al.* (EXAMO) en recherchant pour chaque réaction prédite active, une solution alternative optimale ayant cette réaction inactive et inversement, pour chaque réaction prédite inactive, une solution alternative optimale ayant cette réaction active. A la fin de cette étape, les réactions sont classées en 3 catégories : « requises » lorsque la réaction est active dans toutes les solutions alternatives, « inactives » lorsque la réaction est inactive dans toutes les solutions alternatives et « potentiellement actives » lorsque la réaction est active dans certaines solutions et inactive dans d'autres. Pour les deux approches, l'espace de solutions est exploré afin de trouver des solutions alternatives comportant des variations dans l'ensemble de réactions prédites actives/inactives mais la diversité des solutions alternatives ainsi que la qualité de couverture de l'ensemble de l'espace de solutions n'est pas pris en compte. Afin de proposer une exploration de l'espace de solutions plus complète, les auteurs de [98] ont proposé DEXOM. DEXOM a été développé par le Dr Pablo Rodriguez-Mier lors de son post-doctorat au sein de l'équipe MeX. DEXOM est un algorithme basé sur iMAT qui propose plusieurs stratégies d'énumération partielle dont une (diversity-enum) tirant parti des points forts de plusieurs stratégies d'énumérations afin de couvrir l'ensemble de l'espace de solutions, c'est-à-dire en allant du centre de l'espace à ses extrémités. C'est notamment pour cette capacité à explorer l'ensemble de l'espace de solutions et donc le calcul d'un ensemble de solutions alternatives représentant le plus fidèlement possible l'ensemble des capacités métaboliques du système étudié dans une condition donnée que nous avons choisi DEXOM pour le développement de notre stratégie de modélisation de l'impact métabolique de xénobiotiques. Nous allons donc détailler son fonctionnement au cours des prochaines sections.

4.4.2.2. Exploration de l'espace de solution par DEXOM

Comme nous avons pu le mentionner précédemment, explorer l'espace de solutions est une étape nécessaire pour prendre en compte la diversité des solutions possibles et ainsi éviter l'analyse d'une solution unique qui serait sélectionnée aléatoirement parmi toutes les solutions présentes dans l'espace de solutions. Choisir une solution au hasard risquerait d'impacter les analyses menées sur cette solution. Par exemple, un gène identifié comme essentiel (*i.e.* un gène supposé crucial pour la survie de l'organisme) pour une solution pourrait ne pas être considéré essentiel pour une autre condition. De la même manière, une voie métabolique sur-représentée dans une solution pourrait ne pas l'être dans une autre solution [98], impliquant une différente d'interprétation et donc des conclusions potentiellement différentes d'un point

de vue de la toxicité ou du mMoA de la molécule étudiée. Bien que les méthodes développées par [101,132] soient un premier pas vers la prise en compte de la diversité de solutions dans l'espace de solutions en générant des solutions alternatives avec des modifications dans toutes les voies possibles des réseaux métaboliques et donc une certaine couverture des fonctions biologiques, elles ne garantissent pas une exploration homogène de l'ensemble de l'espace de solution [98]. DEXOM est un algorithme implémentant une approche d'exploration de l'espace de solutions permettant de trouver un ensemble de solutions représentatif de la diversité présente dans l'espace de solutions. DEXOM a été initialement développé en MATLAB afin de s'intégrer à l'écosystème COBRA (principalement développé en MATLAB) puis ensuite en python (https://pypi.org/project/dexom-python/)

Nous avons sélectionné DEXOM pour sa capacité à énumérer un ensemble de solutions alternatives couvrant la diversité de solutions existantes dans l'espace de solutions [98]. Couvrir la diversité de solutions existantes dans l'espace de solutions est important pour être capable de représenter au mieux la capacité métabolique du système dans la condition étudiée (*i.e.* un HPH exposé à un xénobiotique à une dose donnée et une concentration donnée). Représenter au mieux cette capacité métabolique permet d'une part de limiter le risque de se tromper en réalisant une interprétation qui ne représente qu'une sous partie de l'impact métabolique du composé étudié. D'une manière générale, cela permet également de prendre en compte la « plasticité » du métabolisme cellulaire qui se traduit par sa capacité à réaliser certaines fonctions métaboliques de plusieurs manières différentes et peut s'avérer important pour l'étude de certaines pathologies comme le cancer [133] ou certains phénomènes toxiques [134]. L'objectif de ce choix méthodologique est donc *in fine* de maximiser la robustesse des prédictions de l'impact métabolique des composés étudiés.

DEXOM implémente 4 algorithmes d'exploration de l'espace de solutions (Fig 13) : « Reaction-Enum », « Icut-enum », « Maxdist-enum » et « Diversity-enum ». Ces approches complètent le MILP initial de DEXOM (qui est similaire à celui décrit pour l'approche iMAT) en ajoutant des contraintes permettant d'explorer l'espace de solutions.

Reaction-Enum (Fig 13C) :

Cette méthode d'énumération est une re-implémentation de l'approche décrite par [101,132] qui consiste à bloquer successivement le flux de chacune des réactions dans le sens direct et indirect (pour les réactions réversibles). Elle permet d'obtenir des solutions avec des modifications réparties sur l'ensemble des voies métaboliques du réseau. En revanche cette approche ne prend pas en compte des combinaisons de réactions bloquées et génère un grand nombre de solutions dupliquées [98].

Icut-Enum (Fig 13D) :

Cette méthode d'énumération consiste à ajouter une nouvelle contrainte au MILP lorsque l'on trouve une solution alternative. Cette nouvelle contrainte a pour rôle d'empêcher le solveur mathématique de trouver à nouveau la solution alternative venant d'être trouvée lors des prochaines itérations. Théoriquement, cette approche permet d'explorer l'ensemble de l'espace de solutions et de trouver toutes les solutions possibles. Cependant, l'ajout de nombreuses contraintes au cours de l'énumération rend le problème mathématique très difficile à résoudre. Cette approche d'énumération peut notamment être utilisée en combinaison d'autres approches d'énumération (*e.g.* Reaction-Enum ou Maxdist-Enum) afin d'optimiser le temps de calcul en évitant l'énumération de solutions redondantes.

Maxdist-Enum (Fig 13B) :

Cette méthode d'énumération consiste à rechercher les solutions alternatives qui soient les moins similaires possibles. Pour arriver à ce résultat, il faut résoudre deux objectifs de manière simultanée.

Le premier objectif est l'optimisation du problème orginial (*i.e.* maximiser l'adéquation entre les données transcriptomiques et la topologie du réseau métabolique). Le second objectif consiste à maximiser la diversité des solutions énumérées. Pour cela, il faut rechercher une solution qui soit la plus distante possible de la solution de référence (*e.g.* une solution de départ aléatoire). La fonction à minimiser dans ce cas est une fonction mesurant la similarité entre deux solutions. Afin de préserver l'optimalité des solutions alternatives, une nouvelle contrainte est ajoutée au MILP.

Diversity-Enum (Fig 13A) :

Cette méthode d'énumération combine les approches précédentes : « Reaction-Enum », « Icut-Enum » et « Maxdist-Enum » afin de tirer parti des avantages de ces méthodes pour permettre une exploration maximale de l'espace de solutions. La première étape consiste à générer un premier ensemble de solutions avec « Reaction-Enum » combinée à « Icut-Enum » qui permet d'éviter de trouver des solutions alternatives redondantes lors de la phase d'énumération avec « Reaction-Enum ». Les solutions de ce premier ensemble servent ensuite de solutions de départ pour la seconde étape qui est une adaptation de « Maxdist-Enum ». Cette version adaptée de « Maxdist-Enum » recherche graduellement des solutions étant de plus en plus distantes. L'augmentation graduelle de la contrainte de distance requise entre la solution de départ et la solution alternative est contrôlée par un paramètre. Cette étape est répétée en utilisant la solution trouvée précédemment comme solution de départ jusqu'à ce que le nombre de solutions à trouver défini par l'utilisateur soit atteint ou que toutes les solutions possibles aient été trouvées.

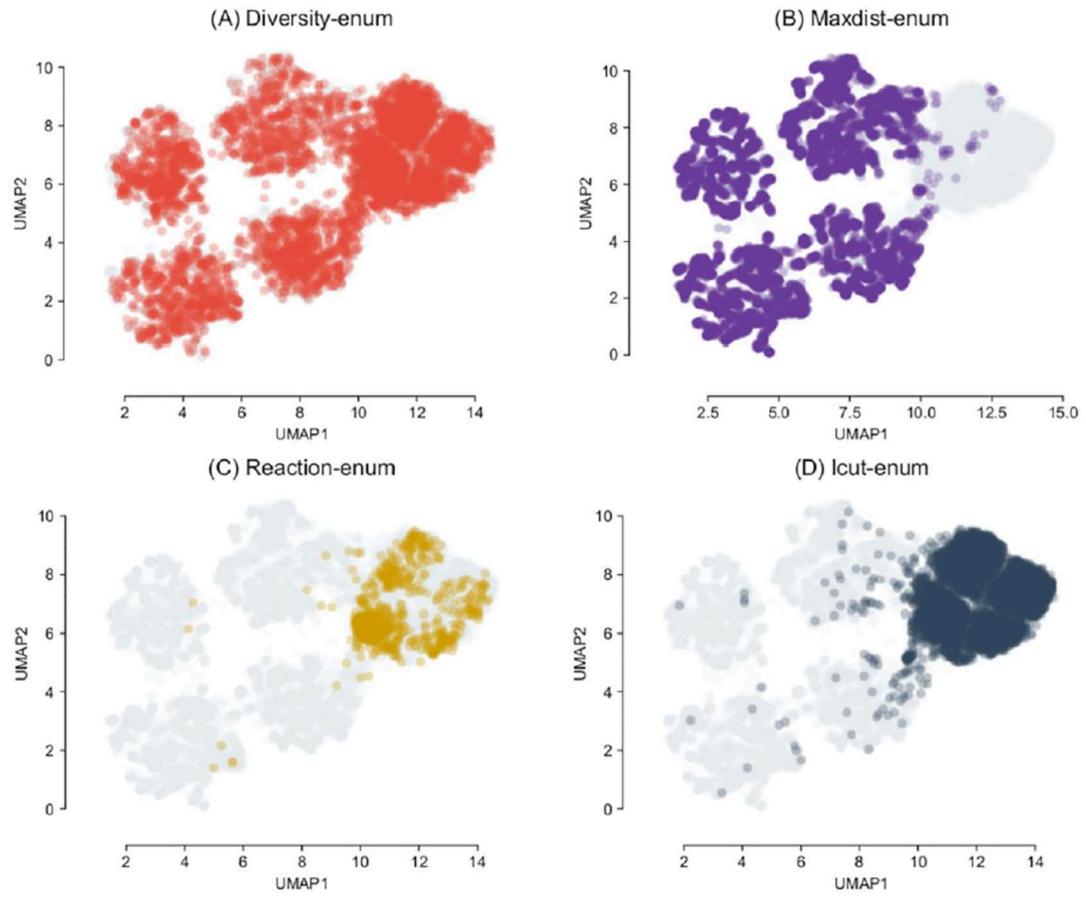

Figure 13 : Ensemble des solutions alternatives trouvées par chacune des méthodes implémentées dans DEXOM dans le cadre de reconstruction condition-spécifique pour une levure*. La capacité de Diversity-Enum à tirer parti des points forts de chacune des 3 autres approches d'énumération est visible sur la Figure 13A puisque Diversity-Enum parvient à trouver les solutions les plus différentes (par rapport aux autres solutions) trouvées par Maxdist-Enum (Figure 13B) ainsi que les solutions plus similaires (les unes par rapport aux autres) trouvées par Icut-Enum (Figure 13D). Figure provenant de* [98]

4.5. Conclusion

Au cours de cette section, nous avons abordé l'un des concepts méthodologiques centraux de cette thèse : la modélisation sous-contraintes. Nous nous sommes d'abord intéressés aux différentes contraintes qu'il était possible de définir puis à la reconstruction de réseaux condition-spécifiques et enfin aux méthodes d'exploration de l'espace de solutions. Nous avons décrit les méthodes « biaisées » et « non biasées » afin d'identifier la méthode la plus adaptée pour modéliser l'impact métabolique d'un xénobiotique avec un coût computationnel maitrisé

et une robustesse de l'interprétation qui soit maximale. Au cours de la prochaine section, nous allons aborder la modélisation du métabolisme par les approches de graphe.

5. Modélisation par les approches de graphe

5.1. Représenter le métabolisme sous forme de graphes

Un graphe est un ensemble de nœuds (ou sommets), reliés entre eux par des arêtes. Le terme de graphe est généralement utilisé pour désigner l'objet mathématique sur lequel un ensemble d'algorithmes et de concepts mathématiques s'applique mais est d'une manière générale interchangeable avec le terme « réseau » qui est plus généralement employé dans un contexte applicatif comme en début de chapitre. Dans cette section, nous utiliserons le terme « graphe » qui se réfère au formalisme mathématique, un graphe G étant définit par un ensemble de nœuds V et un ensemble d'arêtes E reliant des éléments appartenant à V.

Le formalisme du graphe s'adapte particulièrement bien à la représentation du métabolisme. En effet, comme nous avons pu le décrire dans le premier chapitre, le métabolisme est constitué d'un ensemble de réactions et métabolites en interaction. Il est donc possible de représenter le métabolisme sous forme de graphe en représentant les entités impliquées dans le métabolisme (*e.g.* réactions et métabolites) ainsi que leurs interactions au travers de nœuds et d'arêtes.

Il est possible de représenter le métabolisme avec 4 types de graphes différents : le graphe des composés, le graphe des réactions, le graphe métabolique biparti ou l'hypergraphe. Chaque type de graphe ayant ses propriétés topologiques, ses avantages et inconvénients pour la représentation du métabolisme. À noter que le choix des graphes dépend à la fois des données à disposition mais également de l'objectif de l'étude (études des interactions entre les métabolites ou interactions entre les réactions). Au cours des 4 prochaines sections, nous allons représenter un réseau selon les 4 types de graphes énoncés précédemment afin de décrire les propriétés de chacun de ces graphes.

Le réseau que nous allons utiliser au cours des 4 prochaines sections est défini par les équations stœchiométriques suivantes :

M1+M2 -> M3 (R1)
M3 <-> M4 (R2)
M4 -> M5+M6+M8 (R3)
M5+M6 -> M7 (R4)

5.1.1. Graphe des composés

Le graphe des composés est un graphe orienté dans lequel les nœuds représentent des métabolites et les arêtes relient deux nœuds (métabolites) s'il existe une réaction consommant l'un des deux métabolites et produisant l'autre. Chaque arête peut être annotée (Fig 14) avec le nom de la réaction consommant/produisant les deux nœuds (métabolites) qu'elle connecte. A titre d'exemple, définissons le graphe G(V,E) représenté en Figure 14, avec V = {M1,M2,M3,M4,M5,M6,M7,M8} et E → {(u,v) | u,v ∈ V, u≠v}. Sur ce graphe, la direction des réactions est donnée par le sens des flèches (qui représentent les arêtes sur la Figure 14). Plusieurs réactions peuvent consommer/produire les mêmes métabolites, ce qui implique que plusieurs arêtes peuvent connecter les mêmes nœuds dans le graphe des composés. Ces arêtes sont dites « parallèles » et le graphe des composés devient alors un multigraphe des composés avec E un ensemble de triplets (u,v,r), où {u,v} ⊑ V et r ∈ {R1,R2,R3,R4}, l'ensemble des réactions. D'un point de vue pratique, le graphe des composés permet de centrer l'analyse sur les interactions entre les métabolites et peut s'avérer utile pour analyser des jeux de données de métabolomique par exemple.

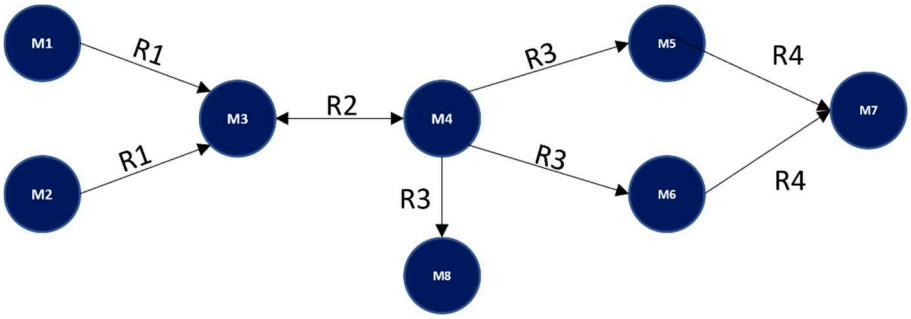

Figure 14: Exemple d'un graphe des composés dirigé. *Les cercles représentent les métabolites, les arêtes représentent les réactions consommant/produisant les métabolites qu'elles connectent. Le sens de la flèche indique le sens de la réaction (quel métabolite est consommé et quel métabolite est produit)*

5.1.2. Graphe des réactions

Le graphe des réactions est un graphe orienté dans lequel les nœuds représentent des réactions et les arêtes relient deux nœuds (réactions) s'il l'une produit un métabolite consommé par l'autre. Chaque arête peut être annotée (Fig 15) avec le nom du métabolite consommé/produit par les deux nœuds qu'elle connecte. A titre d'exemple, définissons le graphe G(V,E) représenté en Figure 15, avec V = {R1,R2,R3,R4} et E un ensemble de triplets (u,v,m), où {u,v} ⊑ V, u≠v et m ∈ {M3 M4, M5, M6}, l'ensemble des métabolites consommés et produits. Sur ce graphe, les flèches (qui représentent les arêtes) pointent vers la réaction consommant le métabolite produit par la réaction à l'autre extrémité. De la même manière que pour le graphe des composés, il est possible d'avoir des arêtes parallèles lorsque plusieurs produits d'une

réaction sont substrats d'une même réaction. Dans ce cas, le graphe des réactions est représenté par un multigraphe des réactions. D'un point de vue pratique, le graphe des réactions permet de focaliser l'attention sur les interactions entre les réactions et peut donc s'avérer utile pour interpréter des perturbations métaboliques dues à des réactions métaboliques plus/moins actives ou l'impact de l'absence d'une enzyme (et donc des réactions associées) par exemple. La topologie du graphe des réactions peut être plus densément connectée que son équivalent graphe des composés, notamment lorsque le graphe contient des réactions réversibles. Il est également intéressant de noter qu'avec ce formalisme, les métabolites source (M1, M2) et puits (M7,M8) du graphe des composés ne sont pas représentés.

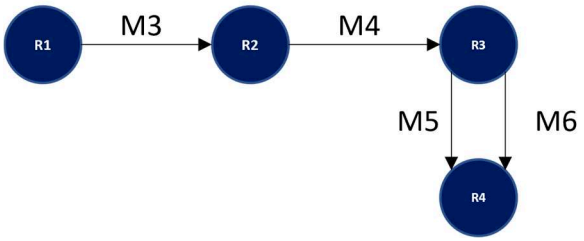

Figure 15: Exemple d'un graphe des réactions dirigé. *Les cercles représentent les réactions, les arêtes représentent les métabolites consommés/produits par les réactions qu'elles connectent. Le sens de la flèche indique le sens de la réaction (par quelle réaction est produit le métabolite ainsi que par quelle réaction le métabolite est consommé)*

5.1.3. Graphe biparti

Le graphe biparti est un graphe dans lequel les nœuds peuvent être séparés en 2 ensembles disjoints tels qu'il n'existe aucune arête entre les nœuds d'un même sous-ensemble. Ainsi, pour un graphe biparti métabolique, les métabolites sont représentés par un premier ensemble de nœuds et les réactions sont représentées par un second ensemble. Il existe une arête entre deux nœuds de deux sous-groupes différents si un nœud du sous-groupe « métabolites » est consommé/produit par un nœud du sous-groupe « réactions ». A titre d'exemple, définissons le graphe G(V,E), V = M∪R représenté en Figure 16, avec M = {M1, M2, M3, M4, M5, M6, M7, M8}, R = {R1,R2,R3,R4} et E = {M1->R1, M2->R1, R1->M3, M3->R2, R2->M3, R2->M4, M4->R2, M4->R3, R3->M8, R3->M5, R3->M6, M5->R4, M6->R4, R4->M7}.

En représentant explicitement toutes les interactions entre les métabolites et réactions du graphe, ce type de graphe permet une compréhension fine des interactions entre métabolites et réactions dans un réseau métabolique. Cependant la présence de groupes de nœuds disjoints implique des algorithmes adaptés capables de prendre en compte cette propriété et peut

également rendre la lecture du graphe plus compliquée à cause du nombre de nœuds plus important par rapport à ses équivalents « graphe des composés » et « graphe des réactions ».

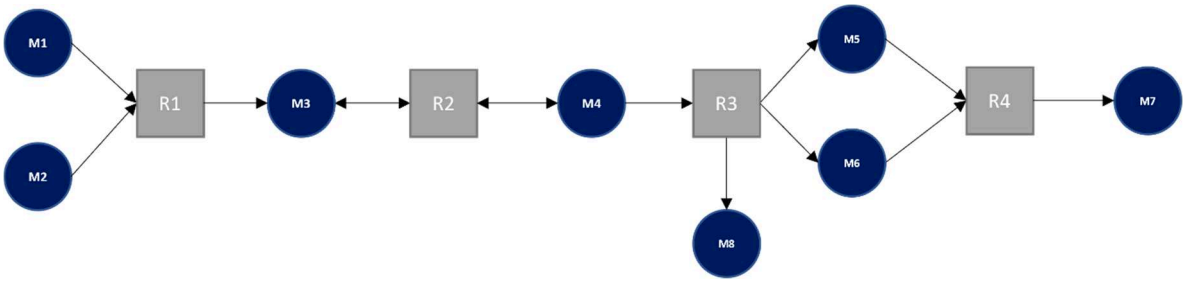

Figure 16 : Exemple d'un graphe biparti dirigé. *Les cercles représentent les métabolites, les carrés représentent les réactions, les arêtes représentent l'interconnexion entre les métabolites et les réactions. Le sens de la flèche indique si la réaction auquel le métabolite est connecté consomme ou produit ce métabolite.*

5.1.4. Hypergraphe

Dans un hypergraphe métabolique, les métabolites sont représentés par des nœuds et les hyper-arêtes représentent les réactions. Une hyper-arête est une arête reliant plus de 2 nœuds [135]. Dans le cas d'un hypergraphe métabolique, une hyper-arête connectera les substrats d'une réaction aux produits de cette même réaction. La différence notable avec un graphe simple tel que le graphe des composés étant la notion d'hyper-arêtes connectant plusieurs nœuds à la fois. D'un point de vue pratique, ce type de représentation est adapté à la visualisation du métabolisme puisque les hyper-arêtes permettent de mieux représenter les interactions entre les métabolites impliqués dans une réaction. En effet, les hyper-arêtes permettent de visualiser le rôle de chacun des métabolites comme un ensemble d'éléments nécessaires à une réaction plutôt que différents éléments séparés comme c'est le cas dans un graphe simple. A titre d'exemple, nous pouvons considérer l'hypergraphe H(V,E) où V = {M1, M2, M3, M4, M5, M6, M7, M8} et E = {R1,R2,R3,R4} = {{M1,M2},{M3,M4},{M4,M5,M6,M8},{M5,M6,M7}} représenté en Figure 17.

Bien que les hypergraphes soient adaptés à la représentation du métabolisme et que certaines approches tirent parti de ce type de représentation [136,137], la majorité des algorithmes fonctionnent sur les graphes simples pouvant être représentés par des matrices d'adjacence en 2 dimensions, tels que le graphe des composés, le graphe des réactions et le graphe biparti [138–140]. L'utilisation d'hypergraphe nécessite donc fréquemment l'usage de méthodes dédiées qui ne bénéficient pas de la même représentativité en termes d'implémentation au sein des bibliothèques logicielles usuelles.

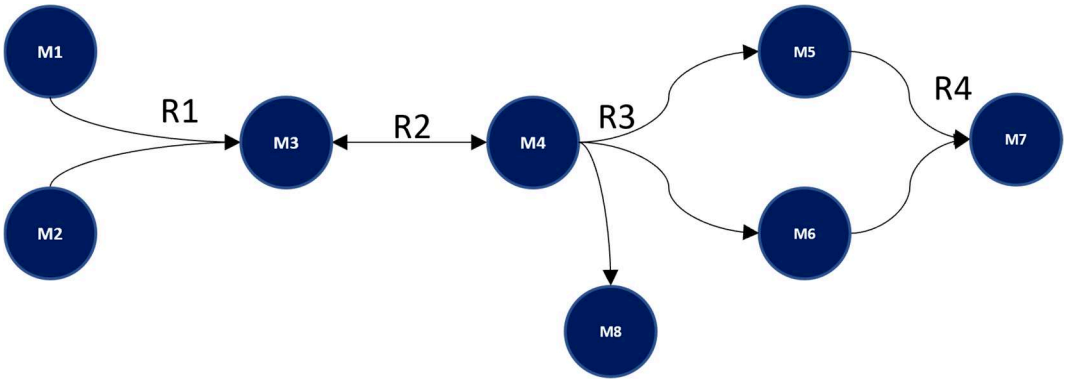

Figure 17 : Exemple d'un hypergraphe dirigé. *Les cercles représentent les métabolites, les arêtes représentent les réactions consommant/produisant les métabolites qu'elles connectent. Une même arête peut connecter plusieurs nœuds (e.g. L'arête R3 qui connecte les nœuds M5, M6 et M8, signifiant que la réaction R3 produit ces trois métabolites). Le sens de la flèche indique le sens de la réaction (quel métabolite est consommé et quel métabolite est produit)*

5.1.5. Conclusion

Comme nous avons pu le discuter au cours des sections précédentes, le formalisme des graphes permet de représenter les interactions entre les réactions et métabolites. Cette capacité à représenter le métabolisme comme un ensemble d'éléments connectés permet d'une part la visualisation de ces interactions et d'autre part l'utilisation d'algorithmes sur ces graphes métaboliques. Nous avons également montré qu'il existait plusieurs types de graphes métaboliques. Le choix du type de graphe métabolique dépend de la pertinence d'un centrage sur les métabolites ou les réactions (en accord avec le type de données utilisé), ou, alternativement, d'une représentation conjointe. Il dépend également du type d'algorithme que l'on souhaite utiliser sur le graphe métabolique puisque comme nous avons pu l'évoquer, la majorité des algorithmes de théorie des graphes ont été développés pour des graphes simples tels que le graphe des composés ou le graphe des réactions. Le formalisme des graphes étant utilisé dans un très grand nombre de domaines (réseaux moléculaires, réseaux de transport, réseaux sociaux, …), nous allons décrire au cours de la prochaine section uniquement les classes d'algorithmes adaptées aux graphes métaboliques simples ainsi qu'à l'objet de notre étude.

5.2. Algorithmique appliquée à la théorie des graphes

Comme nous avons pu le décrire au cours de la section précédente, le formalisme des graphes permet de représenter un grand nombre d'éléments en interaction de manière structurée. Ces deux propriétés sont clés pour représenter le métabolisme. Afin de pouvoir comparer et interpréter ces graphes, différentes méthodes algorithmiques ont été mises au point. La première utilisation des concepts liés à la théorie des graphes correspond aux travaux de Leonhard Euler qui a proposé le problème des « sept ponts de Königsberg » au début du 18ème siècle (Fig 18). Ce problème consistait à trouver un chemin à travers la ville, en traversant chacun des sept ponts une seule fois. Afin de répondre à ce problème, Euler a travaillé avec un

schéma simplifié de la ville de Königsberg en la découpant en 4 régions, régions reliées entre elles par les 7 ponts de la ville. Cette représentation schématique de la ville correspond en fait à une représentation sous forme de graphe de la ville où les régions correspondent à des nœuds et les ponts à des arêtes. En calculant le degré (*i.e.* le nombre de ponts connectés) de chacune des régions, Euler a prouvé qu'un tel chemin dans un graphe n'es possible que si le graphe est connexe et s'il existe exactement zéro ou deux nœuds de degré impair : chaque entrée dans une région nécessitant une sortie (degré pair), sauf pour le point de départ et d'arrivée, s'ils sont distincts.

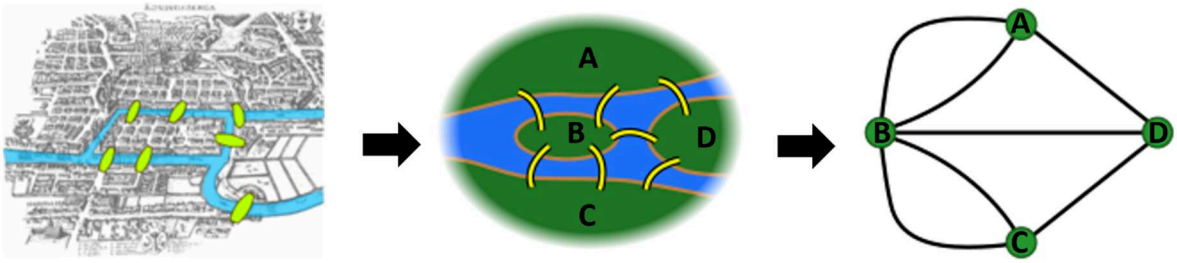

Figure 18: Formalisation du problème des sept ponts de Königsberg. *Schéma adapté de Wikipédia (https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg)*

En ne considérant pas le détail de chacune des régions (sous-parties de la ville constituées de maison, rues et de surface variable) mais plutôt les interconnexions (représentées par les ponts) entre ces régions, Euler a montré qu'une telle abstraction pouvait permettre une résolution simple et généralisable d'un tel problème. Depuis, la théorie des graphes a continué d'évoluer tant d'un point de vue représentation que d'un point de vue algorithmique afin de permettre la résolution de problèmes de plus en plus difficiles à résoudre. Il existe de nombreux algorithmes appliqués à la théorie des graphes, nous allons nous intéresser à deux catégories qui nous serons utiles pour la suite de ce manuscrit. Nous commencerons par décrire des algorithmes permettant l'analyse des propriétés topologiques des graphes pour ensuite décrire des algorithmes permettant d'extraire des sous-graphes.

5.2.1. Algorithmes pour l'analyse topologique des graphes : recherche de chemins

Le problème de la recherche du plus court chemin consiste à trouver un chemin entre deux nœuds d'un même graphe qui soit le chemin de coût minimal. Dans le cas d'un graphe simple non pondéré, ce coût représente la longueur totale du chemin, c'est-à-dire son nombre d'arêtes. Ce problème peut être généralisé aux cas des graphes pondérés, où une fonction de valuation adjointe au graphe défini un coût pour chaque nœud et/ou arêtes, qui peut varier entre les éléments.

Le coût minimal d'un chemin dans un graphe pondéré (un graphe dont les étiquettes (ou label) de sont des nombres positifs) correspond alors à la somme des poids des arêtes ou des nœuds par lequel le chemin passe.

L'un des algorithmes les plus connus pour la recherche de plus court chemin est l'algorithme de Dijkstra [141]. Cet algorithme est adapté aux graphes pondérés par des poids non négatifs. L'algorithme de Dijkstra recherche le plus court chemin entre un nœud source et un nœud cible. Par exemple, définissons un graphe G(V,E) avec V = {A, B, C, D, E, F, G} et E = {A-B, B-E, A-D, D-F, E-F, A-C, C-D, C-G, F-G} représenté en Figure 19. L'algorithme de Dijkstra recherche le chemin le plus court entre le nœud source (A, en bleu sur la figure 19) et tous les autres nœuds du graph G, jusqu'à atteindre le nœud cible (G, en orange sur la figure 19). A noter qu'il est également possible de continuer l'exploration du graphe pour calculer le plus court chemin entre le nœud source et tous les nœuds du graphe, comme représenté sur la figure 19. Sur la figure 19, le chemin le plus court entre le nœud source (A) et le nœud cible (G) est le chemin en rouge de forte épaisseur passant par A->D->C->G. Pour arriver à ce résultat, l'algorithme de Dijkstra calcul progressivement le plus court chemin entre le nœud source et le nœud cible en itérant sur les nœuds voisins du nœud source jusqu'à atteindre le nœud cible. Par exemple, sur l'exemple figure 19, l'algorithme de Dijkstra a calculé le plus court chemin de A vers B, D et C puis vers G et C puis vers H le nœud cible. A noter qu'il est possible de ne trouver aucun chemin entre un nœud source et un nœud cible. Ce cas de figure arrive notamment lorsque le graphe étudié n'est pas connexe, ce qui est peut-être le cas pour les graphes métaboliques. Un graphe connexe étant un graphe pour lequel quels que soient les nœuds d'un graphe, il existe un chemin entre ces nœuds

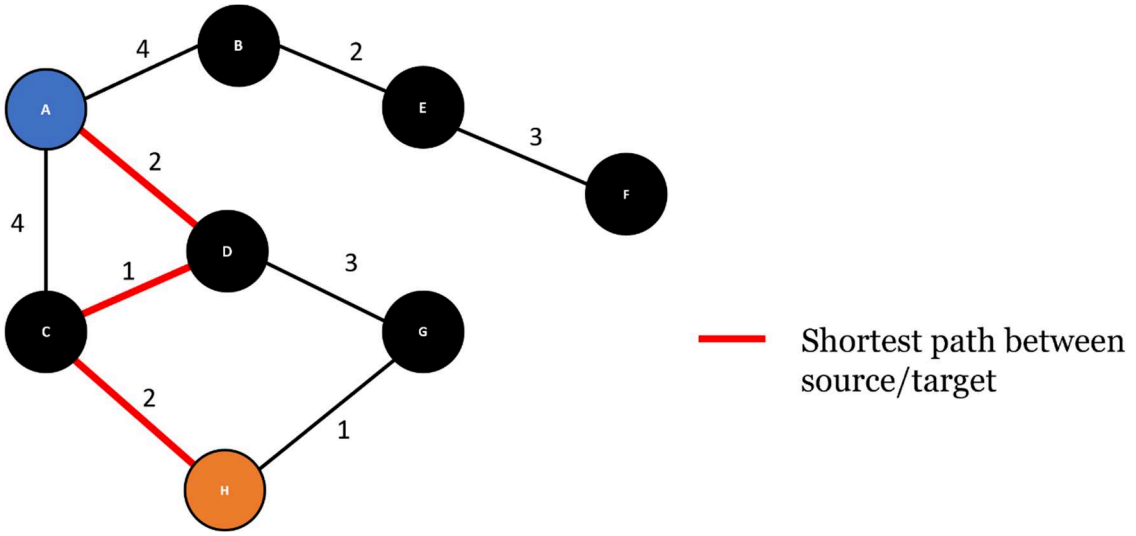

Figure 19 : Exemple du fonctionnement de l'algorithme de Dijkstra pour la recherche du plus court chemin entre deux nœuds. *Les chiffres indiquent le poids attribué à chacune des arêtes, les arêtes colorées en rouge et plus épaisses sont les arêtes faisant partie du plus court chemin entre le nœud source et le nœud cible.*

Il existe d'autre algorithmes permettant de calculer des plus courts chemins dans un graphe. Par exemple, l'algorithme de Bellman-Ford [142] est capable de calculer les plus courts chemins entre un nœud source et tous les autres nœuds d'un graphe dont les arêtes sont pondérées négativement et/ou positivement. L'algorithme de Bellman-Ford est également capable d'identifier des chemins de poids strictement négatifs entre le nœud source et les autres nœuds du graphe.

L'algorithme de Floyd-Warshall [143] permet de rechercher le plus court chemin entre toutes les paires de nœuds existantes dans un graphe.

La recherche de plus court chemin est un problème qui peut s'avérer couteux en temps de calcul si le nombre de plus courts chemins à calculer est important (*i.e.* nombreuses paires de nœuds source/cible) et que l'on travaille sur de très grands graphes (*e.g.* les réseaux sociaux numériques ou les réseaux routiers). De plus, plusieurs « plus courts » chemins peuvent exister entre deux nœuds, ce qui n'aura pas d'impact si l'on s'intéresse uniquement à la distance entre deux nœuds mais risque de biaiser l'interprétation si l'on s'intéresse au contenu des chemins (*i.e.* les réactions et métabolites impliqués).

5.2.2. Algorithmes pour l'analyse topologique des graphes : partitionnement de graphes

Le partitionnement de graphes permet d'identifier des groupes de nœuds localisés à proximité ou densément connectés les uns des autres dans le graphe. Cette mesure de la proximité dans le graphe peut notamment être obtenue en calculant les plus courts chemins entre tous ces nœuds. Identifier des partitions dans un graphe métabolique permet notamment d'identifier des groupes de nœuds représentant des éléments que l'on peut supposer être en interaction dans le système modélisé par le graphe étudié et donc permettre une interprétation plus « fonctionnelle » du graphe.

Ces groupes de nœuds peuvent par exemple représenter des communautés ou des modules fonctionnels, informant ainsi sur la structure du graphe et du système qui est modélisé. Dans un graphe les nœuds sont connectés entre eux par des arêtes et il faut donc un critère ou une fonction objective permettant de déterminer la séparation optimale permettant de constituer des groupes de nœuds proches. Il existe plusieurs approches permettant de partitionner des graphes.

Certaines de ces approches utilisent une fonction qualitative afin de maximiser la qualité des sous-graphes qui ont été calculés. La modularité est certainement l'une des fonctions qualitatives les plus populaires [144]. La modularité mesure la qualité des partitions trouvées en comparant la densité des arêtes connectant chacun des nœuds à celle que l'on trouverait si les nœuds de cette partition étaient connectés de manière aléatoire. Une valeur de modularité

importante signifie que le partitionnement trouvé est de qualité. Les algorithmes tels que l'algorithme de détection de communauté de Louvain ou son évolution, l'algorithme de Leiden [144], permettent d'utiliser la modularité pour trouver de manière itérative le meilleur partitionnement. D'autres fonctions qualitatives peuvent être utilisées pour évaluer la pertinence du partitionnement, certaines partagent des similitudes avec la modularité en évaluant la pertinence des partitions via des mesures de densité (*e.g*. Graph Density ou Partition Density) ou de connectivité au sein des partitions (*e.g*. Normalized Cut ou Pairwise Connectivity Index) alors que d'autres approches évaluent la distance entre les nœuds au sein d'une même partition ou entre deux partitions.

Il est également possible de calculer différentes mesures topologiques à partir d'un graphe (distances, degrés, …) et de les stocker dans des matrices. Par exemple, la matrice de distance ou de similarité (e.g « common neighors ») entre toutes les paires de nœuds d'un graphe permet d'appliquer un algorithme de partitionnement hiérarchique sur cette matrice et d'identifier des groupes de nœuds à partir de l'observation du dendrogramme obtenu. Le partitionnement spectral cherche également à représenter un graphe sous forme de matrice en construisant la matrice laplacienne du graphe et d'appliquer une méthode de classification non supervisée telles que les « k-means » sur les valeurs propres de cette matrice. La matrice laplacienne d'un graphe est construite en soustrayant la matrice d'adjacence par la matrice des degrés [145]. La matrice d'adjacence étant une matrice de taille *noeuds x noeuds* représentant le nombre d'arêtes reliant chacune des paires de nœuds du graph et la matrice des degrés étant une matrice qui contient sur sa diagonale, le degré de chacun des sommets du graphe.

Ces deux approches nécessitent de définir un nombre de partition attendue *a priori* ce qui peut s'avérer difficile à réaliser de manière objective. En effet, la définition du nombre de partition attendue sera propre à chaque utilisateur et à son sujet d'étude.

Il existe un grand nombre d'algorithmes et d'approches différentes permettant de partitionner un graphe. Dans le cadre des graphes métaboliques, déterminer ce qu'est un bon partitionnement peut être une tache compliquée et plutôt subjective. Il est par exemple possible de comparer le partitionnement obtenu aux différentes voies métaboliques existantes. On peut également évaluer la qualité de chacune des partitions par des métriques topologiques telles que la modularité ou la distance moyenne entre les nœuds de la partition.

Le choix de la méthode de partitionnement se fera donc d'une part par rapport aux propriétés topologiques du graphe à partitionner et d'autre part en adéquation avec les contraintes d'analyse.

5.2.3. Algorithmes pour l'extraction de sous-graphes

Les graphes représentant des systèmes comme le métabolisme peuvent être de grande taille et densément connectés (10 000 nœuds et plusieurs dizaines de milliers d'arêtes). De fait, l'analyse visuelle de ces graphes peut être difficile sans des méthodes permettant de se focaliser sur certaines sous-parties d'intérêt [146]. Ces sous-parties d'intérêt peuvent alors être représentées sous la forme de sous-graphes du graphe principal. Un sous-graphe G'(V,E) du graphe G(V,E) est un graphe G' tel que $V(G') \in V(G)$ *et* $E(G') \in E(G)$ [147,148]. Il existe plusieurs algorithmes permettant d'extraire des sous-graphes. Certains permettent d'extraire un sous-graphe à partir d'une liste de nœuds définie *a priori* alors que d'autres algorithmes permettent d'extraire des motifs répétés ou des sous-graphes communs entre deux ou plusieurs sous-graphes.

L'une des approches permettant d'extraire un sous-graphe à partir d'une liste de nœuds d'intérêt (appelés « seeds ») est l'extraction de l'arbre couvrant de poids minimal. Un arbre est un graphe connexe (il existe un chemin entre tous les nœuds du graphe) et acyclique (il n'existe pas de chemin permettant de partir d'un nœud et y revenir sans repasser par des nœuds déjà visités). L'arbre couvrant d'un graphe non dirigé G est un arbre incluant tous les nœuds du graphe G. Enfin, un arbre couvrant de poids minimal est un arbre couvrant dont la somme des poids (des arêtes ou des nœuds selon l'approche) n'est pas plus importante que celle des autres arbres couvrant du graphe G. L'extraction d'un arbre couvrant de poids minimum est utilisée dans de nombreux domaines tels que l'optimisation de réseaux de télécommunications, routiers ou dans les réseaux biologiques [149]. La notion de poids peut être portée par les arêtes ou les nœuds (par les arêtes dans les cas développés au cours de ces travaux) fait référence à une métrique topologique (*i.e.* la distance métabolique, représentée par le nombre d'intermédiaires connectant deux nœuds). En calculant l'arbre minimal couvrant une liste de nœuds fourni par l'utilisateur, cette approche permet de faciliter l'interprétation des graphes de grande taille (plusieurs milliers de nœuds) en permettant de focaliser l'étude d'une sous-partie du graphe que l'on considère d'intérêt. La taille du sous-graphe est impactée par le nombre de nœuds mais également par la distance entre ces nœuds. Cependant, il peut exister plusieurs arbres couvrant de poids minimal pour un même graphe et un même ensemble de « seeds ». L'analyse d'un seul arbre revient donc à ignorer les arbres alternatifs. Ces arbres alternatifs contiennent systématiquement les « seeds » passées en paramètre mais varient par rapport aux nœuds ajoutés pour la connectivité de l'arbre couvrant. Ignorer ces arbres alternatifs revient donc à ignorer ces différentes « options » possibles entre les nœuds d'intérêt, ce qui dans le cas d'un graphe métabolique peut impacter l'interprétation qui sera faite du sous-graphe extrait. Bien que le formalisme des arbres (qui sont des graphes acycliques et connexes) soit intéressant pour obtenir un sous-graphe de petite taille (relativement au

nombre de nœuds « seeds »), il peut s'avérer restrictif pour la représentation de sous-graphes métaboliques qui ne sont pas forcément connexes ou acycliques.

Il peut donc être intéressant de se tourner vers d'autres algorithmes d'extraction de sous-graphes permettant d'une part d'apporter une solution au problème de non-unicité des arbres couvrants de poids minimal et d'autre part de ne pas être limité par les propriétés des arbres.

Par exemple, il est possible de considérer les plus courts chemins alternatifs avec l'algorithme des « k-shortest path » [150,151]. L'algorithme du « k-shortest path » consiste à rechercher un nombre k de plus courts chemins entre deux nœuds. Par exemple, en utilisant itérativement cet algorithme sur toutes les paires de nœuds ayant servi de « seed » pour l'approche d'extraction de sous-graphe par le calcul de l'arbre couvrant de poids minimal, il est possible d'obtenir un sous-graphe contenant des chemins non représentés dans l'arbre de Steiner, et pouvant contenir des cycles. Cependant, s'il existe de nombreux chemins alternatifs la taille du sous-graphe représentant tous ces chemins alternatifs risque d'augmenter très rapidement et de rendre ce sous-graphe très difficile à analyser par des approches visuelles.

Enfin, une autre approche consiste à rechercher le sous-réseau qui capture au mieux les relations d'interdépendance entre des nœuds d'intérêt en évaluant l'importance de chacune des arêtes dans le graphe via un algorithme de marche aléatoire. Chaque arête porte une valeur relative à la probabilité d'être empruntée lors d'une marche aléatoire partant d'une des « seeds ». Un sous-graphe peut être extrait par ajout itératif des arêtes de plus grande valeur puis élagage. Ce type d'approche développée par [152] peut soit explorer l'ensemble des chemins possible entre des nœuds d'intérêt (*i.e.* algorithme « K-walks ») ou ne considérer que les chemins d'une longueur inférieure à un entier passé en paramètre (*i.e.* algorithme « Limited k-walks »).

Bien qu'au cours de ces travaux de thèses nous ayons choisi de porter notre attention sur des approches permettant de focaliser notre attention sur le mMoA via des approches permettant d'extraire un sous-graphe contenant une liste de nœuds d'intérêt (« seeds »), il existe des approches permettant d'extraire des sous-graphes sans définir de « seeds ». Ces approches s'appuient principalement sur la topologie du graphe et sont généralement utilisées pour comparer des graphes [147] ou calculer des propriétés topologiques.

Une des approches les plus basiques consiste à extraire les composantes connexes d'un graphe. La composante connexe d'un graphe est un sous-graphe dans lequel il existe un chemin entre toute paire de nœuds. Dans un graphe non connexe, il existe généralement plusieurs composantes connexes d'une taille pouvant aller de 1 nœud à la quasi-totalité des nœuds du graphe (Fig 20). L'intersection de deux composantes connexes est toujours nulle car s'il existait

une arête entre deux nœuds appartenant chacun à une composante différente alors ces deux composantes n'en formeraient qu'une seule.

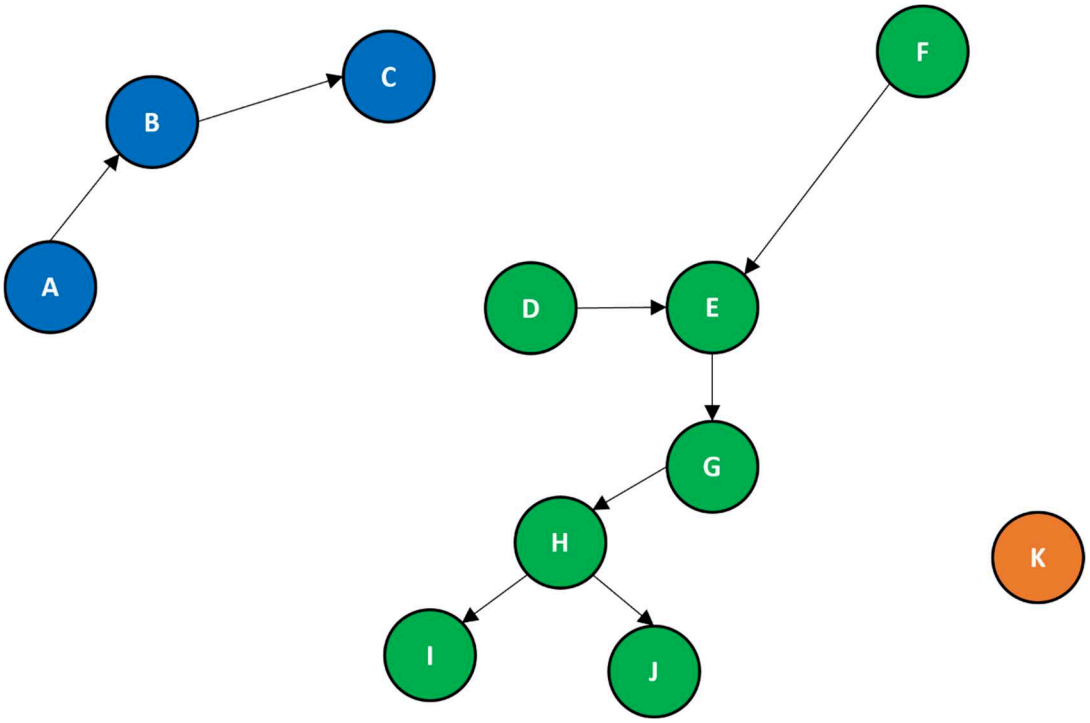

Figure 20 : Exemple de composantes connexes*. Ce graph est constitué de 11 nœuds et 8 arêtes. Trois composantes connexes ont été identifiés : la composante bleue (A, B, C), la composante verte (D, E, F, G, H, I, J) et la composante orange constituée seulement du nœud K.*

Dans le cadre des graphes métaboliques, extraire les composantes connexes peut permettre d'identifier des zones/fonctions métaboliques n'étant pas connectées et donc n'interagissant pas avec le reste du métabolisme (dans le cas où on considère le réseau métabolique parfaitement reconstruit). Travailler avec la composante connexe géante (ou principale) peut également permettre l'utilisation de certains algorithmes ne fonctionnant que sur un graphe connexe comme par exemple, le clustering hiérarchique qui n'est pas applicable lorsqu'il n'existe pas de chemin entre au moins deux réactions (distance infinie).

Les approches d'extraction du sous-graphe commun maximal (Maximum Common Subgraph en anglais), permettent notamment de comparer deux ou plusieurs graphes. Leur objectif est d'identifier un sous-graphe isomorphique commun aux deux graphes et qui soit le plus grand possible. A noter que deux graphes sont dits isomorphiques s'ils possèdent la même topologie [147]. Ces approches sont fréquemment utilisées en chemoinformatique afin de comparer des structures chimiques [153], explorer les relations structure-activité [154] mais également dans d'autres disciplines telles que la cybersécurité ou la reconnaissance de motifs [147].

Enfin il existe des méthodes d'extraction de sous-graphes dont l'objectif est d'identifier des « graphlets » ou des structures répétées. Ces sous-graphes peuvent alors être utilisés comme des descripteurs topologiques de graphes [155,156] ou des variables pour l'utilisation

d'algorithmes d'apprentissage machine [157]. Les « graphlets » sont des ensembles de sous-graphes induits d'un graphe. Un sous-graphe induit de G(V,E) étant défini par G'=(S,E(S)) avec S un sous-ensemble de nœuds appartenant à V et E(S), les arrêtes connectées à l'ensemble de de nœud S d'après les arêtes du graphe G. Contrairement aux motifs répétés, les motifs représentés par les « graphlets » ne sont pas significativement plus/moins présents dans le graphe principal par rapport à un graphe aléatoire (un graphe dont les nœuds sont connectés de manière aléatoire) [158]. On peut considérer les motifs répétés (Fig 21) comme un cas particulier des « graphlets » qui représentent l'ensemble des combinaisons de sous-graphes possibles avec 1, 2, 3, n nœuds et sont couramment utilisés afin de comparer des graphes [159].

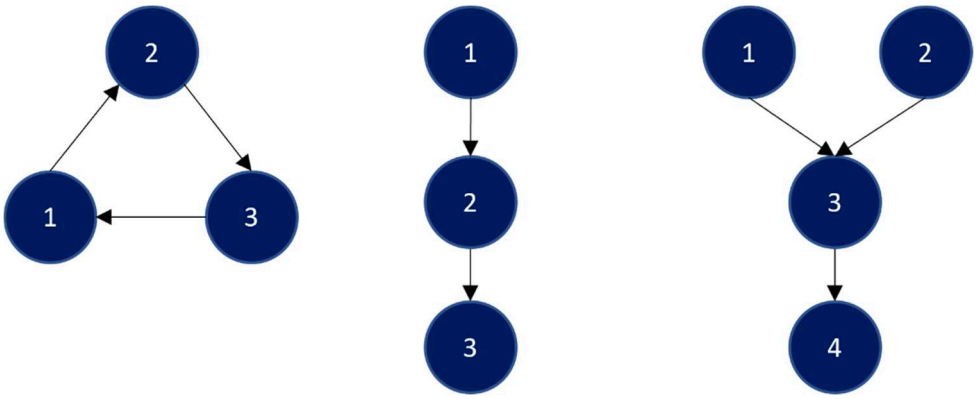

Figure 21 : Exemple de motifs répétés. *Les motifs répétés sont des groupes de nœuds interconnectés reproduisant une structure significativement plus fréquente dans le graph d'intérêt par rapport à un graph aléatoire.*

Dans ce chapitre, nous avons abordé une fraction de la grande diversité des algorithmes ayant été développés au fil du temps pour tirer parti du formalisme des graphes. Représenter le métabolisme par des graphes permet donc de tirer profit de ces développements réalisés dans la grande variété des champs d'application de la théorie des graphs et donc de bénéficier de méthodes performantes pour étudier les graphs métaboliques. Comme nous avons pu le constater la recherche de chemin, le partitionnement ainsi que l'extraction de sous-graphes sont des champs d'application très pertinents pour l'étude des graphes métaboliques en permettant une précision et une simplicité accrue de l'analyse.

6. Objectifs de la thèse

En introduction générale, nous avons pu aborder l'importance du développement de nouvelles approches *in silico* pour répondre aux nombreux défis posés par l'interdiction complète de l'expérimentation animale dans le secteur cosmétique et plus largement dans le cadre du respect des 3Rs (Réduire, Remplacer, Raffiner). La recherche de nouvelles méthodes alternatives pour l'amélioration de l'évaluation de la toxicité est un domaine très dynamique à l'heure actuelle. Ces méthodes s'appuient généralement sur des données structurales ou omiques. Cependant, les données structurales sont peu ou pas disponibles pour les nouvelles

classes de matière première cosmétiques (produit naturel et/ou mélanges complexes), de nouvelles méthodes se basant sur les données omiques et notamment les données de transcriptomique sont développées. L'exploitation des données transcriptomiques passe en général par l'identification de gènes significativement différentiellement exprimés suivi par une analyse fonctionnelle via une analyse de sur-représentation. Cependant, ces approches couramment utilisées pour caractériser de potentielles perturbations suite à une exposition à un xénobiotique ne prennent pas directement en compte le métabolisme et considèrent les gènes comme indépendants les uns des autres. L'objectif principal des travaux décrits dans cette thèse est donc d'améliorer l'analyse de l'impact métabolique suite à l'exposition à un xénobiotique au travers de l'intégration de données transcriptomiques issues d'expérimentation *in vitro* au réseau métabolique à l'échelle du génome humain par la modélisation sous-contraintes et l'extraction de sous-graphes. Pour répondre à cet objectif, nous proposons une stratégie permettant de modéliser l'impact métabolique d'un xénobiotique en calculant un ensemble de réseaux condition-spécifique pour chaque condition modélisée (Fig 22A). Ces réseaux condition-spécifiques sont ensuite exploités pour calculer des DARs (Fig 22B) dont l'interprétation mécanistique sera ensuite réalisée à l'aide d'un ensemble de méthodes de théorie des graphes (Fig 22C) permettant une meilleure compréhension du mMoA.

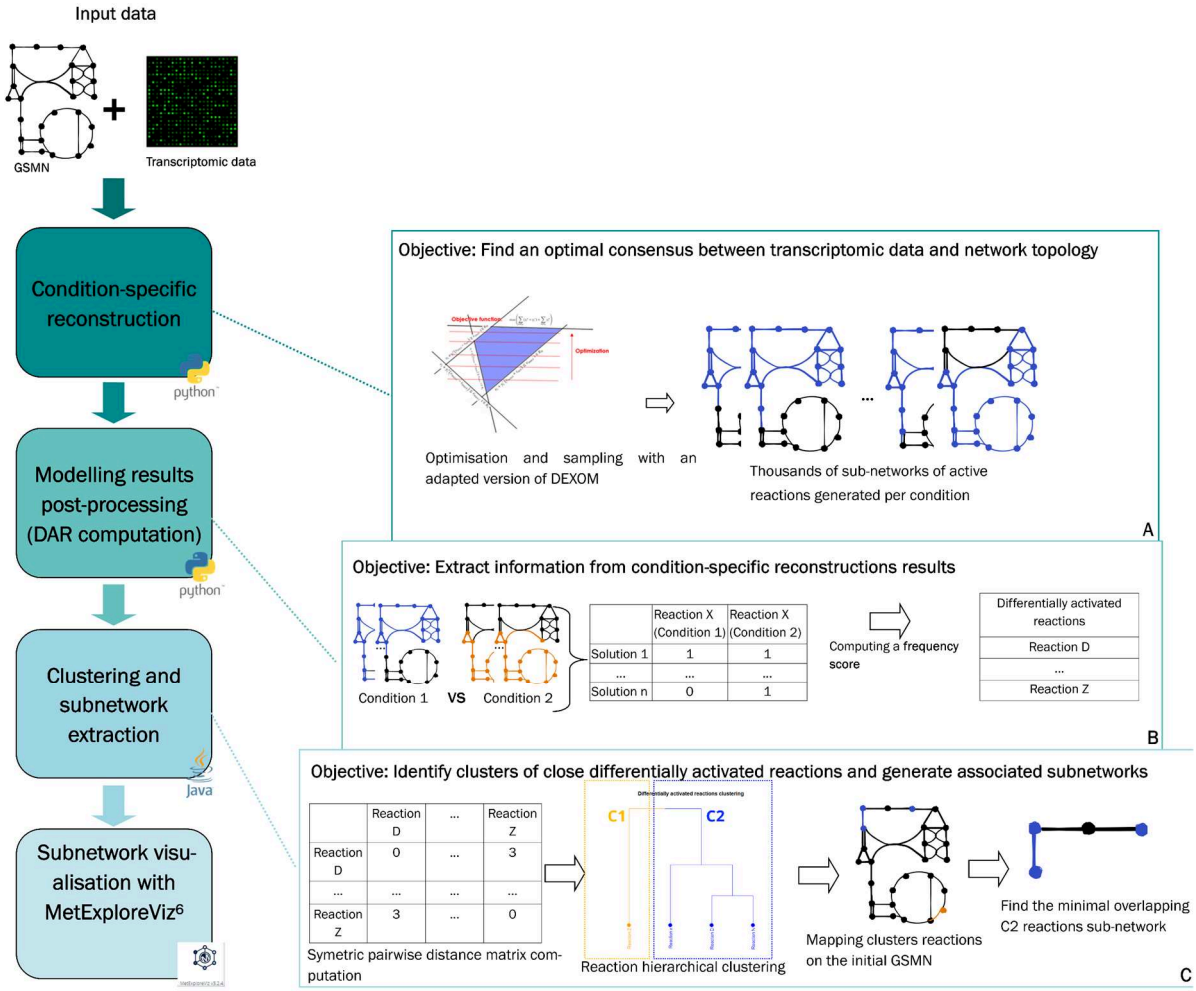

Figure 22 : Schéma de la stratégie de modélisation de l'impact métabolique de xénobiotiques. *La première étape (A) de la stratégie correspond à l'intégration des données transcriptomiques sous la forme de contraintes lors de la modélisation avec énumération partielle réalisée par une version adaptée de DEXOM. Dans la seconde étape (B), des réactions différentiellement activées (DARs) sont calculées à partir des dizaines de milliers de réseaux métaboliques (solutions alternatives) énumérés pour chaque condition par notre version adaptée de DEXOM. Enfin, la troisième étape (C) fait appel à des approches d'analyse basée sur les graphes pour visualiser la relation entre les DARs et améliorer notre compréhension du mécanisme d'action métabolique (mMoA) prédit pour chaque composé testé.*

Nous aborderons donc dans la prochaine partie comment nous avons analysé et traité les données transcriptomiques provenant de la base de données publique Open TG-GATEs avant de pouvoir les intégrer au réseau métabolique. Dans une troisième partie, nous aborderons en détail comment nous avons adapté DEXOM pour calculer des ensembles de réseaux condition-spécifiques. Nous détaillerons également la méthode de calcul des DARs à partir de ces ensembles de réseaux condition-spécifiques. Une application sur les 8 molécules hépatotoxiques sera également réalisée. Enfin, la quatrième partie sera consacrée aux approches de théorie des graphes ayant été implémentée afin de mettre en évidence et visualiser le mécanisme d'action métabolique à partir des DARs prédites à l'étape précédente. Une application sur 2 des 8 molécules hépatotoxiques sera présentée.

Chapitre 2 : Obtention, Exploration et Préparation des données pour la reconstruction condition-spécifique

1. Open TG-GATEs : une base de données transcriptomiques d'exposition à des molécules pharmaceutiques

Les bases de données publiques sont des ressources d'une très grande valeur scientifique mises à la disposition de la communauté. Mettre ces bases de données en libre accès permet de maximiser l'intérêt des données ayant été générées initialement. Ces bases peuvent alors servir pour les études de read-across biologique (qui nécessite de grandes quantités de données homogènes), l'entrainement d'algorithmes d'apprentissage machine ou encore le développement de nouvelles approches comme cela a été le cas au cours de cette thèse. Cependant, avant d'utiliser tout ou partie des données issues d'une base, il convient d'explorer en détail les données et métadonnées qu'elle contient afin de comprendre sa structure, la complétion des données ainsi que les potentiels biais. Comme nous avons pu le voir en introduction, les bases de données omiques et plus précisément les bases de données transcriptomiques peuvent permettre d'apporter de nouvelles contraintes pour la modélisation sous-contraintes. C'est dans ce cadre que nous avons utilisé les données transcriptomiques d'exposition à des composés pharmaceutiques disponibles dans la base de données Open TG-GATEs afin d'étudier l'impact de ces composés sur le métabolisme cellulaire par des approches de modélisation sous-contrainte. Nous avons donc dans un premier temps réalisé une étude détaillée des caractéristiques de la base de données Open TG-GATEs [160], qui est décrite dans les sections suivantes.

1.1 Caractéristiques générales

Open TG-GATEs est une base de données transcriptomiques. Lors de sa création en 2002 dans le cadre du projet Toxicogenomics Project One (TGP1), l'objectif principal des auteurs était de créer une base de données d'empreintes géniques d'exposition à des concentrations faiblement cytotoxiques (*i.e.* cytotoxicité inférieure à 20%) pour 150 molécules. Les composés ayant été sélectionnés sont principalement des molécules pharmaceutiques. Ce projet s'est déroulé sur une période de 10 ans et a impliqué des acteurs académiques et industriels japonais de premier plan tels que le National Institute of Health Sciences (NIHS), le National Institute of Biomedical Innovation (NIBIO) et plusieurs partenaires. Les données ont été générées sur des puces à ADN (généralement appelées microarray) Affymetrix HGU133Plus2. Le grand nombre de molécules testées ainsi que la forte proportion de molécules pharmaceutiques représentaient, au moment de la création de la base, environ 10% des molécules autorisées sur le marché japonais du médicament. Les auteurs de la base de données Open TG-GATEs ont également cherché à couvrir la majorité des catégories thérapeutiques disponibles sur le marché japonais lors de la création du projet (Fig. 23, adaptée de [161]) bien que la classification en classe thérapeutiques soit subjective. Il est également intéressant de noter la

présence sur la figure 23 de deux classes n'étant pas des catégories thérapeutiques : « Toxicants » et « Preclinical ». La classe « Toxicants » fait référence à des composés non pharmaceutiques mais connus pour leur hépatotoxicité tels que des solvants ou des produits phytosanitaires. La classe « Preclinical » fait quant à elle référence à des composés développés par les entreprises partenaires du projet mais dont le développement a été arrêté après la découverte d'effets hépatotoxiques ou néphrotoxiques.

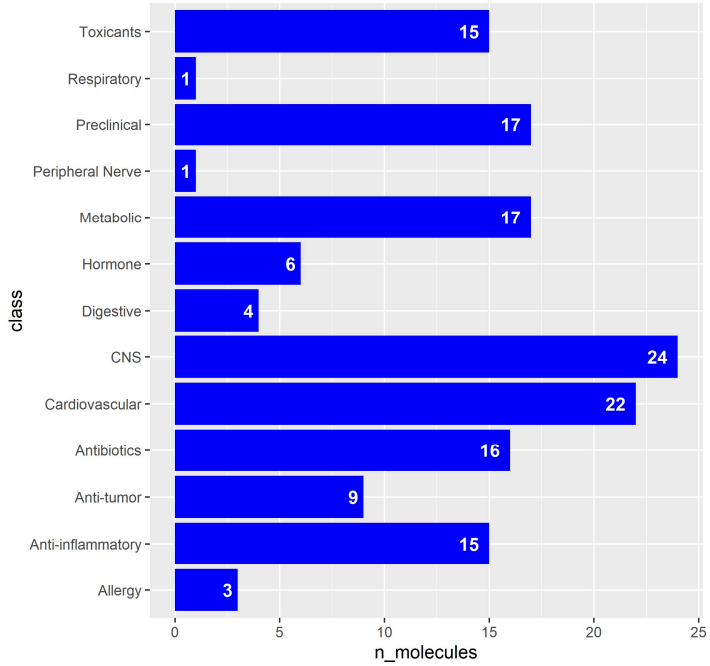

Figure 23 : Répartition des 150 molécules sélectionnées par le projet TGP1 au sein de 13 classes pharmacologiques.

Afin de compléter la base de données Open TG-GATEs, le Toxicogenomics Project Two (TGP2) a été mis en place de 2010 à 2011. Ce projet a notamment permis d'étendre le nombre de molécules testées de 150 à 170 molécules.

Des données ont été générées pour deux modèles : le rat et l'humain. 158 molécules ont été testées pour le modèle humain sur des cultures d'HPH. 170 molécules ont été testées *in vivo* sur le modèle rat pour lequel des données d'expression génique ont été mesurées

Modèle	Rat					Humain
Tissu	Foie			Rein		Foie
Type de modèle	*In vitro*	*In vivo*		*In vivo*		In vitro
Nombre de répétition	Concentration unique	Concentration unique	Concentration répétée	Concentration unique	Concentration répétée	Concentration unique
Nombre de molécules	146	159	144	42	42	158

Tableau 2 : Tableau récapitulatif du nombre de molécules testées pour chaque modèle, tissu et type d'expérimentation.

Les données ayant été générées sur le modèle rat sont celles comprenant le plus grand nombre de conditions testées. En effet, des données *in vivo* ont été générées pour des concentrations aigues à 3, 6, 9 ou 24 heures mais également à des concentrations journalières répétées pendant 4, 8, 15 ou 29 jours. Pour les études *in vivo*, la plus forte concentration correspond à la concentration minimale induisant un effet toxique sur une durée d'exposition de 4 semaines, la concentration moyenne correspond à la concentration forte divisée par 3 et la concentration faible correspond à la concentration forte divisée par 10. Pour les études *in vitro,* la forte concentration correspond à la concentration induisant entre 10 et 20% de cytotoxicité (mort cellulaire) ou la concentration maximale soluble si cette concentration est inférieure à celle induisant le degré de cytotoxicité attendu. La concentration moyenne correspond à la forte concentration divisée par 5 et la concentration faible correspond à la forte concentration divisée par 25.

Bien que les données générées sur le modèle rat soient plus complètes au travers de données transcriptomiques multi organe (foie et rein), à concentration unique et répétées mais également au travers de données biologiques (poids des organes, marqueurs sanguins, …) et histopathologiques, nous nous sommes focalisés sur les données générées *in vitro* chez l'humain afin de développer des méthodes et modèles qui pourront être appliquées aux futures données générées pour l'étude de la sécurité des produits cosmétiques, données qui par définition ne seront pas générées sur des modèles animaux.

1.2 Pré-traitement des données brutes, identification et correction des effets lot d'hépatocyte

Nous avons obtenu les données brutes stockées sous la forme d'archives pour chacune des molécules testées sur les HPH à partir du lien suivant : https://dbarchive.biosciencedbc.jp/data/open-tggates/LATEST/Human/in_vitro/

Les métadonnées associées à la base de données Open TG-GATEs ont été obtenues via le lien suivant :

https://dbarchive.biosciencedbc.jp/en/open-tggates/data-13.html

Ces métadonnées sont stockées dans un tableau Excel contenant les métadonnées disponibles pour chacun des 33566 échantillons dont 24023 pour lesquels des données microarray ont été générées. Dans un premier temps, nous avons donc filtré ce tableau de métadonnées afin de ne conserver uniquement les métadonnées des échantillons correspondant aux expériences *in vitro* chez l'humain.

Après avoir téléchargé l'ensemble des données brutes générées sur HPH ainsi que les métadonnées associées, nous avons utilisé le package R « affy » [162] qui est un package permettant de lire les données brutes contenues dans les fichiers CEL. Les fichiers CEL sont les fichiers permettant de stocker les données lues par le logiciel « Affymetrix DNA microarray image analysis » à partir des puces à ADN Affymetrix. Une fois les données brutes lues et stockées en mémoire, nous avons choisi de les normaliser avec la méthode de normalisation RMA (pour Robust Multi-array Average). Cette méthode est très largement utilisée par la communauté depuis sa publication en 2003 [163] et permet notamment de normaliser les intensités d'expression en évitant le biais associé aux fortes intensités d'expression de certaines sondes PM (Partial Match) décrit pour une autre méthode de normalisation couramment utilisée, la normalisation MAS5.

Cela représente au total 2605 échantillons répartis sur 158 molécules, 3 temps d'exposition, 3 concentrations et les contrôles correspondants. Le calcul du nombre théorique maximal d'échantillons par molécule est le suivant : 3 concentrations sont testées à 3 temps d'exposition soit un total de 9 échantillons. Sont ajoutés à ces 9 échantillons, 3 échantillons contrôle correspondant au solvant de la molécule testé aux 3 temps d'exposition, ce qui porte le nombre d'échantillon à 12 par réplicas. Etant donné qu'il y a deux réplicas par molécule, il y a au maximum 2*12 échantillons, soit un total maximal théorique de 24 échantillons par molécule.

Ces 2605 échantillons n'étant pas répartis de manière uniforme selon les années de génération des données, nous avons calculé un ratio d'intégrité des données afin d'estimer la complétude de la base de données année après année. Le ratio d'intégrité des données fait référence au nombre d'échantillons disponible par rapport au nombre d'échantillons théorique maximal,

soit 24 échantillons par molécule pour les données *in vitro* humaines de la base de données Open TG-GATEs.

Ces données ayant été générées sur HPH, il est important de prendre en considération les caractéristiques propres à ce type cellulaire lors du traitement des données transcriptomiques. Comme cela a pu être abordé en introduction, les HPH sont considérés comme le meilleur modèle cellulaire pour l'étude des phénomènes d'hépatotoxicité *in vitro* chez l'Homme [164]. Ce modèle dispose de capacités fonctionnelles très proches de celles d'un hépatocyte fonctionnel dans l'organe et permet donc une meilleure transposition des résultats obtenus [164]. Cependant, les différences inter-individuelles ainsi que les altérations cellulaires lors du prélèvement et de la procédure d'isolation des cellules induisent indéniablement des variations dans les résultats.

Cependant, dans le cadre de la construction de grandes bases de données telles qu'Open TG-GATEs, l'utilisation de différents lots d'hépatocytes rendra difficile la comparaison entre les échantillons au sein même de la base car ces échantillons seront susceptibles d'être générés sur des cellules différentes présentant de forts effets lots. Le risque est notamment que l'effet lot d'hépatocyte soit l'effet mesuré majoritaire et que les molécules ne soient pas classées en fonction de variables d'intérêt telles que le type d'hépatotoxicité, le degré d'hépatotoxicité ou encore le mécanisme d'action mais plutôt en fonction du lot d'hépatocyte sur lequel l'expérimentation a été réalisée.

Ces informations concernant le lot d'hépatocyte utilisé pour chaque échantillon n'étant pas disponibles dans l'article ou en ligne, nous avons sollicité les auteurs principaux de la base de données Open TG-GATEs : Dr Tetsuro Urushidani, Dr Yoshinobu Igarashi et Dr Hiroshi Yamada. Nous avons alors identifié que 6 lots d'HPH différents avaient été utilisés pour générer l'ensemble de la base de données. Grâce à l'analyse en composantes principales (ACP) présentée en figure 24, nous avons constaté que l'effet « lot d'hépatocyte » était le facteur de variabilité majoritaire de la base de données. Un effet « lot d'hépatocyte » important comme celui observé sur la figure 24 complique fortement l'étude de la base de données Open TG-GATEs dans son ensemble sans corriger cette variabilité. En effet, on distingue une séparation claire entre les lots CELL030, CELL080 et les lots CELL0020, CELL0040, CELL0050 et CELL0060. Un tel résultat indique que l'effet lot est plus important que les effets concentrations et temps d'exposition alors que l'on pourrait s'attendre à ce que ces deux variables soient les vecteurs principaux de la variabilité observée dans la base. Cela signifie que ce biais technique engendre une variabilité importante des résultats qui n'est pas lié à l'exposition des HPH aux différentes molécules testées et qui peut par conséquent masquer l'effet de la molécule (qui est celui qui nous intéresse).

L'indisponibilité de ces métadonnées en ligne est potentiellement à l'origine de plusieurs publications utilisant cette base de données sans faire mention de cet effet et sans le corriger [165–167].

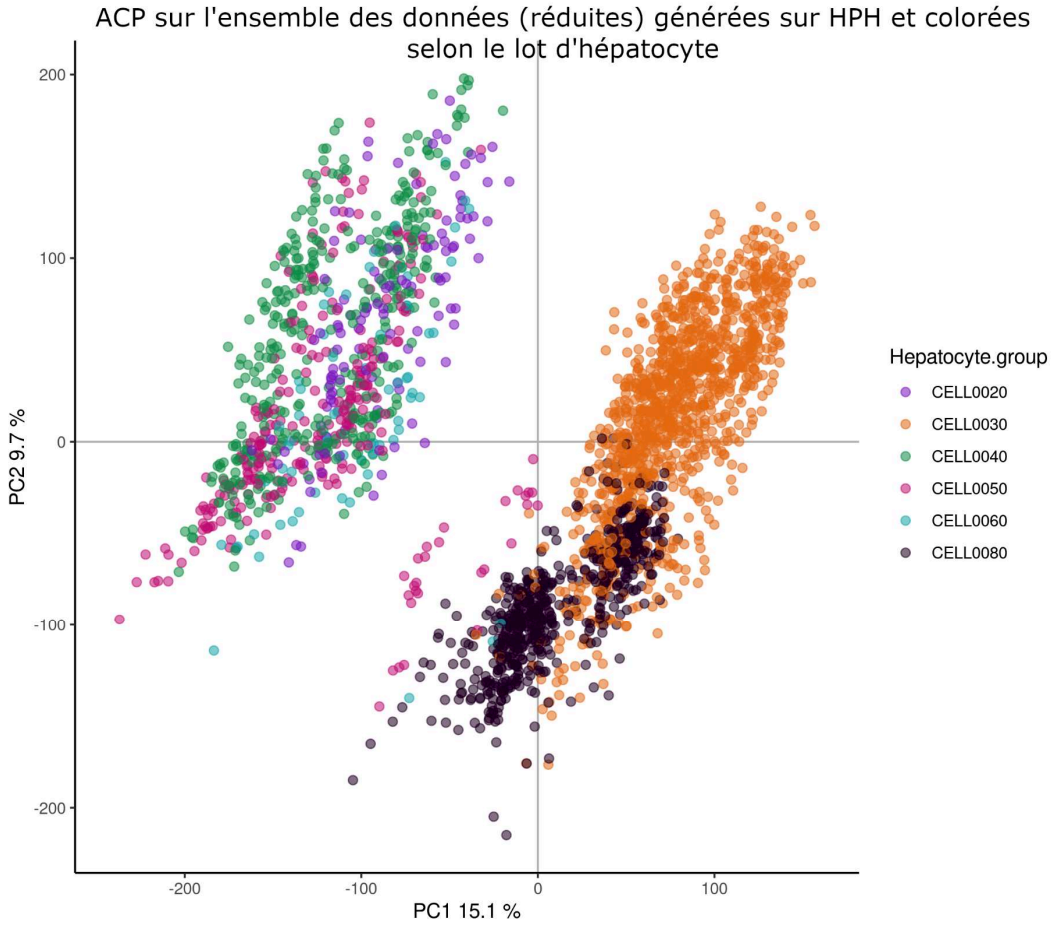

Figure 24 : Analyse en Composantes Principales (ACP) sur l'ensemble des échantillons des données transcriptomiques générées sur HPH dans la base Open TG-GATEs*. Chaque échantillon est coloré selon le lot d'hépatocyte auquel il appartient.*

Dans le tableau 3, on peut s'apercevoir, comme attendu, que la répartition males/femelles est très dépendante du lot d'hépatocyte puisque chaque lot d'hépatocyte correspond à un donneur particulier. Il est également intéressant de remarquer qu'un seul des 6 lots d'hépatocytes à un ratio de complétion de données de 1, c'est-à-dire des données d'expression pour deux réplicas pour toutes les conditions (control, low, middle, high et 2hr, 8hr, 24hr). Le calcul de ce ratio d'intégrité par année nous renseigne d'une part sur certains choix effectués au cours des 10 années pendant lesquelles la base a été générée. D'autre part, cette information est importante à prendre en compte lorsque l'on souhaite utiliser la base de données dans son ensemble via des approches d'apprentissage machine, comme ce fut le cas en début de projet, et qui sont sensible à ces déséquilibres de disponibilité de données.

*Lot de HPH**	Date de réalisation	Nombre d'échantillons	Nombre de composés testés	Ratio d'intégrité des données	Répartition Males/Femelles
CELL0020	2006	120	10	0.5	120/0
CELL0030	2004	1152	48	1	0/1152
CELL0040	2006	384	32	0.5	0/384
CELL0050	2006	276	23	0.5	0/276
CELL0060	2006	72	6	0.5	0/72
CELL0080	2011	601	39	0.64	601/0
TOTAL		2605	158	0.69	721/1884

Tableau 3 : Répartition des échantillons, composés, ratio d'intégrité des données et ratio males/femelles pour chacun des lots d'hépatocytes primaires humain identifiés dans la base de données Open TG-Gates. * Lot de HPH obtenus après avoir contacté les auteurs principaux de la base de données Open TG-GATEs.

Après l'identification et la visualisation de cet effet lot de cellule, nous avons réalisé une correction statistique afin d'en limiter les effets sur notre analyse. Nous avons réalisé cette correction avec le package R « sva » [168] et plus spécifiquement la fonction « ComBat » qui permet de corriger les effets lot connus. La fonction « ComBat » prend en entrée l'ensemble des données transcriptomiques normalisées (par RMA dans notre cas) générées sur HPH de la base de données Open TG-GATEs, ainsi que le lot d'hépatocyte auquel chaque échantillon appartient. Après correction des données par « ComBat », nous pouvons observer (Fig 25) que les échantillons ne sont plus répartis selon le lot d'hépatocyte sur lequel les données ont été générées, ce qui montre que l'effet lot d'hépatocyte a été correctement corrigé. Nous distinguons dorénavant deux groupes séparés sur les deux premières composantes de l'ACP.

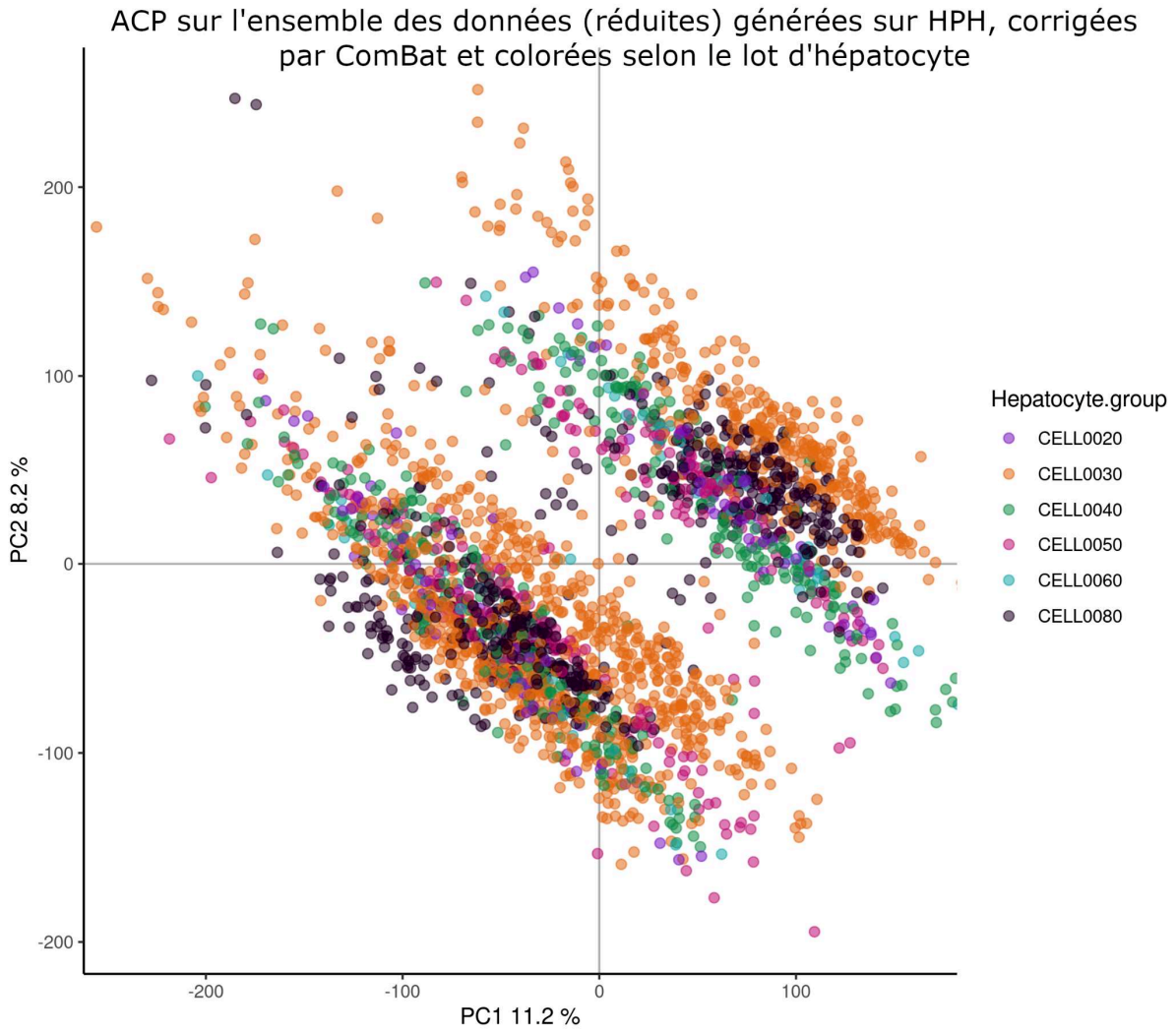

Figure 25: Analyse en Composantes Principales (ACP) sur l'ensemble des données transcriptomiques générées sur HPH et corrigées par l'approche ComBat dans la base de données Open TG-Gates. *Chaque échantillon est coloré selon le lot d'hépatocyte auquel il appartient.*

En colorant l'ACP selon le temps d'exposition associés à chaque échantillon, on s'aperçoit que l'on peut en réalité séparer les échantillons selon 2 groupes : d'un côté les échantillons associés à un temps d'exposition de 2 heures (en violet sur la Figure 26) et 8 heures (en orange sur la Figure 26), ayant des profils transcriptomiques plus proches que les échantillons associés à un temps d'exposition de 24 heures (en vert sur la Figure 26). Après correction de l'effet lot, l'observation des données via une ACP suggère que l'effet temps d'exposition est l'effet le plus important pour expliquer la variabilité dans les données d'expression génique. Les ACPs présentées en Figure 25 et 26 suggèrent que ComBat a effectivement corrigé l'effet lot d'hépatocytes et que des comparaisons entre molécules testées sur différents lots d'hépatocytes peuvent maintenant être réalisées.

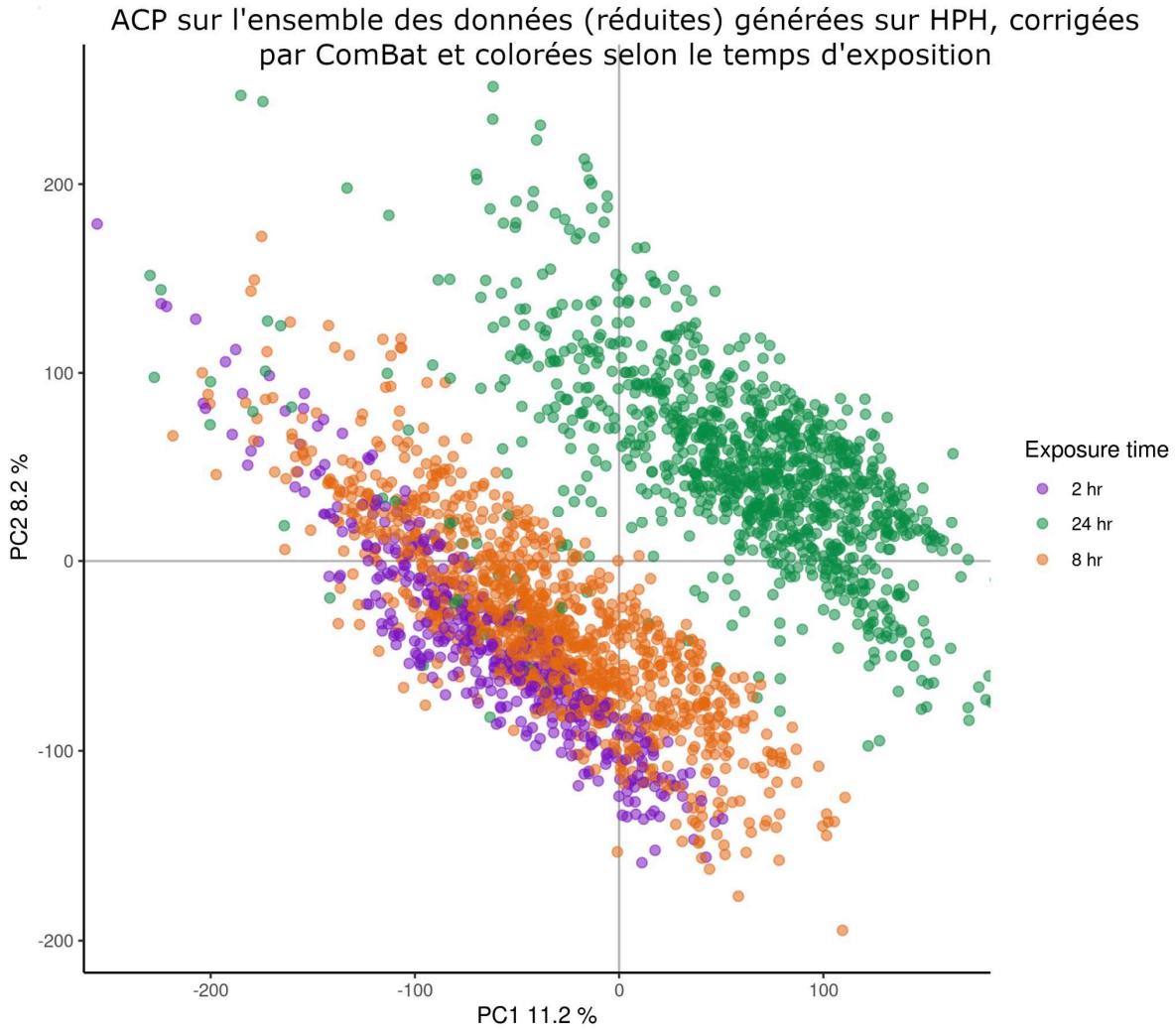

Figure 26 : Analyse en Composantes Principales (ACP) sur l'ensemble des données transcriptomiques générées sur HPH et corrigées par l'approche ComBat dans la base de données Open TG-Gates. *Chaque échantillon est coloré selon le temps d'exposition.*

1.3 Classification des molécules selon leur potentiel hépatotoxique selon la FDA

Pour continuer dans notre exploration de la base et l'accumulation de connaissances tant au niveau de la structure de la base que des composés qui ont été testés, nous avons recueilli un ensemble de métadonnées indicatives du niveau d'hépatotoxicité des molécules d'Open TG-Gates.

La classification en « Most-DILI-concern », « Less-DILI-concern », « No-DILI-concern » et « Ambiguous-DILI-concern » provient du jeu de données DILIrank [169] faisant partie de la base de données « Liver Toxicity Knowledge Base », maintenue par la FDA. Nous avons été en mesure d'obtenir la classe « DILI » de 105 molécules sur les 158 molécules pour lesquelles des données *in vitro* sur HPH sont disponibles. Les 53 molécules n'étant pas dans le jeu de données « DILIrank » appartiennent à des classes de molécules telles que les solvants, les produits phytosanitaires ou des chimiokines et ne sont donc pas des molécules pharmacologiques.

Selon Thakkar *et al* [169,170], une molécule classée comme « Most-DILI-concern » est une molécule pour laquelle un lien de causalité avec des phénomènes hépatotoxiques a été établi et qui est classée dans l'une des catégories suivantes par la FDA : « withdrawn drugs due to hepatotoxicity », « Black-Box warning for hepatotoxity » and « high severity for hepatotoxicity in the warnings and precautions section » (Tableau 4). Une molécule classée comme « Less-DILI-concern » est une molécule appartenant à la catégorie « low DILI severity for hepatotoxicity in the warnings and precautions section » ou une molécule faisant partie de la catégorie « Adverse Reactions » (Tableau 4). Une molécule classée comme « Ambiguous-DILI-concern » est une molécule associée à des phénomènes hépatotoxiques mais pour laquelle le lien de causalité reste à établir. Enfin une molécule classée comme « no-DILI-concern » est une molécule pour laquelle aucune association avec des phénomènes hépatotoxiques n'a été retrouvée. La classification DILIrank fournie par la FDA est une classification intéressante. Bien que peu précise, elle permet d'identifier rapidement et facilement le potentiel hépatotoxique d'une base de données et donc de comprendre le profil des molécules concernant leur potentiel hépatotoxique. Elle nous a permis d'une part de mieux connaitre le profil toxicologique de la base de données.

Classe DILI	Classe FDA
Most-DILI-concern	Withdrawn drugs due to hepatotoxicity
	Black-Box warning for hepatotoxicity
	High severity for hepatotoxicity in the warnings and precautions section
Less-DILI-concern	Low DILI severity for hepatotoxicity in the warnings and precautions section
	Adverse Reactions
Ambiguous-DILI-concern	Pas de classe attribuée par la FDA mais associée à des phénomènes hépatotoxiques dont la causalité reste à déterminer
No-DILI-concern	Aucune association avec des phénomènes hépatotoxiques retrouvée

Tableau 4 : Répartition des classes de DILI (Drug-Induced Liver Injury) pour les molécules avec des données *in vitro* chez l'humain.

1.4 Annotation toxicologique à partir de la littérature

Une annotation simplifiée telle que le « DILIrank » est adaptée à une compréhension rapide du risque associé à une molécule. Cependant, le manque de couverture (*i.e.* 105 molécules sur 158) ainsi que la diversité des mécanismes pouvant engendrer des dommages hépatiques nous a poussé à construire une classification plus précise des différents types d'atteintes hépatiques associées aux molécules de la base de données Open TG-Gates. Afin de construire cette classification, nous avons fait appel à l'expertise du Dr Bernard Fromenty au travers de sa collaboration avec l'équipe de biologie systémique de L'Oréal. Dans un premier temps, nous avons recherché un maximum d'informations au travers de bases de données telles que LiverTox [171] et Hepatox [172,173]. Cela nous a permis notamment de construire un premier niveau de classification à partir des classes définies par Biour *et al :*

- BIOL (Biologique)
- AIGUE (Aiguë)
- CYTOL (Aiguë cytolytique)
- CHOLE (Aiguë cholestatique)
- MASS (Aiguë massive)
- GRAN (Granulomateuse)
- CHRON (Chronique)
- CIRRH (Cirrhose)
- STEAT (Stéatose)
- VASC (Vasculaire)
- TBEN (Tumeur bénigne)
- TMAL (Tumeur maligne)

Cette classification a pu être complétée grâce à deux informations complémentaires disponibles dans la base de données Hépatox. Pour chaque molécule, nous avons recueilli le mécanisme supposé de l'atteinte hépatique ainsi que le nombre de publications ayant permis l'annotation du degré d'hépatotoxicité. Ces deux informations permettent de mieux interpréter la classe associée à chaque molécule tant d'un point de vue mécanistique que de la robustesse de l'annotation.

La base de données Hépatox ainsi que les publications qui y sont associées nous ont permis d'annoter 117 des 158 molécules de la base de données Open TG-Gates pour lesquelles des données transcriptomiques *in vitro* chez l'Homme ont été générées. Nous avons pu obtenir des informations concernant le potentiel hépatotoxique de 3 autres molécules à l'aide d'une recherche dans la base de données LiverTox. Les 38 molécules sans annotations ont fait l'objet

d'une recherche bibliographique par le Dr Bernard Fromenty. Afin d'optimiser cette recherche et de regrouper certaines classes de Biour *et al*, nous avons défini 5 grandes classes d'atteinte hépatique regroupant les classes de Biour *et al* :

- Anomalies biologiques modérées ou franches (BIOL, AIGUE)
- Atteintes hépatiques aiguës cytolytiques (CYTOL)
- Atteintes hépatiques aiguës cholestatiques (CHOLE)
- Hépatites ou atteintes hépatiques chroniques (CHRON, CIRRH, GRAN, TMAL, TBEN)
- Stéatoses (STEAT)

Ces 5 grandes classes devraient nous permettre de classer les molécules par grands types d'hépatotoxicité.

D'une manière générale, une des limites à cette méthodologie de recherche bibliographique est la disparité dans la quantité d'information disponible pour chaque molécule. En effet, certaines molécules telles que l'amiodarone sont très bien représentées dans la littérature avec 192 articles ayant permis de caractériser les différents types d'atteinte hépatique engendrés par cette molécule. A l'inverse, pour certaines molécules pour lesquelles les phénomènes hépatotoxiques sont peu étudiés telles que l'interleukine 1β (1 article) ou l'éthynylestradiol (2 articles), la caractérisation des potentielles atteintes hépatiques repose sur très peu d'information. Il serait donc intéressant de prendre en compte ce déséquilibre dans la littérature dans l'interprétation de la caractérisation de l'hépatotoxicité. De plus, dans la littérature biomédicale portant sur la toxicologie, ne sont généralement décrits que les effets toxiques : il est plus courant de trouver les atteintes hépatiques engendrées par la molécule plutôt que l'absence d'atteinte hépatique (l'absence d'effet n'étant souvent pas publié). Nous avons donc choisi de classer les molécules dans ces 5 classes en précisant si : Oui, ce type d'anomalie a été retrouvé dans la littérature ou Non, ce type d'anomalie n'a pas été retrouvé dans la littérature. Une annotation binaire telle que « Oui » ou « Non retrouvée » pour chacune des classes est notamment adapté à l'entrainement de modèles prédictifs de chacune des classes, ce qui était l'un des objectifs à long terme que nous avions au moment de l'annotation de la base de données. D'une manière générale, cette annotation en 5 grandes classes d'hépatotoxicité plus orientés « mécanismes » que les classes définies par la FDA, nous a permis une meilleure connaissance de la répartition des différents types et mécanismes d'hépatotoxicité au sein de la base de données Open TG-GATEs et donc renforcer notre connaissance de la base de données et le choix des molécules pour les futurs développements réalisés.

1.5 Prise en compte des limites des données transcriptomiques générées sur puces à ADN

Les puces à ADN permettent de mesurer l'intensité d'expression de milliers de gènes voir de plusieurs dizaines de milliers de gènes comme c'est le cas pour les puces Affymetrix HGU133plus2 utilisées pour générer la base de données Open TG-GATEs. Ces puces à ADN permettent de mesurer l'intensité d'expression d'une liste de gène fixe et déterminée à l'avance. Les puces Affymetrix HGU133plus2 permettent de mesurer l'intensité d'expression d'une large partie des gènes humains connus. Cependant, la génération de données d'expression génique sur ces puces souffre de quelques limites telles que l'impossibilité de mesurer des gènes n'ayant pas été intégrés à la puce par le fabriquant, l'occurrence d'hybridations non-spécifiques [174] sur les sondes engendrant du bruit pouvant induire en erreur certaines méthodes de normalisation et la présence de plusieurs sondes par gène donnant lieu à de multiples intensité d'expression associées à un seul gène.

Les données transcriptomiques provenant de la base de données Open TG-Gates ont été générées sur des puces à ADN affymetrix HGU133plus2. Comme expliqué ci-dessus, cette puce constituée de 54 220 sondes soit 22187 gènes après annotation permet une couverture très importante du génome humain. Cependant, le manque de spécificité d'un nombre important de sondes (Fig 27) implique que certains gènes sont associés à plusieurs sondes avec des intensités d'expression différentes. En analysant la distribution du nombre de sondes par gènes (Fig 27), on peut constater que plus de la moitié (12074 sur un total de 22187) des gènes identifiés sont associés à au moins 2 sondes. Etant donné que l'intégration des données transcriptomiques lors de la modélisation sous-contraintes se fait à l'échelle du gène et requiert de n'avoir qu'une seule intensité d'expression par gène, nous avons dû associer une seule sonde par gène lors de l'étape d'annotation du la puce. Il existe plusieurs approches permettant de faciliter l'association sonde/gène et de prendre en compte les cas où plusieurs sondes peuvent être associés au même gène. Il existe des approches dites naïves telles que la sélection de la sonde ayant la meilleure p-value lors de l'analyse d'expression différentielle ou la sonde avec l'intensité d'expression la plus proche de la moyenne des intensités d'expression des autres sondes. Certaines approches plus élaborées font appel à des alignements de séquence et/ou des mesure de similarités entre les sondes [175] pour tendre vers une assignation gène-sonde plus fine. Dans notre cas, nous avons opté pour une approche « naïve » qui consiste à calculer l'écart-type des intensités d'expression propre à chaque sonde pour l'ensemble des échantillons. Lorsqu'un gène est associé à plusieurs sondes, c'est l'intensité d'expression de la sonde avec le plus grand écart-type parmi toutes les sondes associées au gène qui est attribuée au gène en question. Cette approche est similaire à l'approche utilisée par les auteurs de [176] et part du principe que bien que plusieurs sondes soient en mesure de s'hybrider avec le brin

d'ARN messager correspondant au transcrit du gène cible, une seule sonde aura une intensité d'expression très différente de celle mesurée pour les autres sondes. Cependant dans le cas d'un gène associé à seulement deux sondes, cette approche revient à choisir aléatoirement l'une des deux sondes, ce qui n'est pas une approche idéale. Il serait intéressant d'évaluer à quel point des approches plus élaborées permettent une meilleure association transcrit/gène.

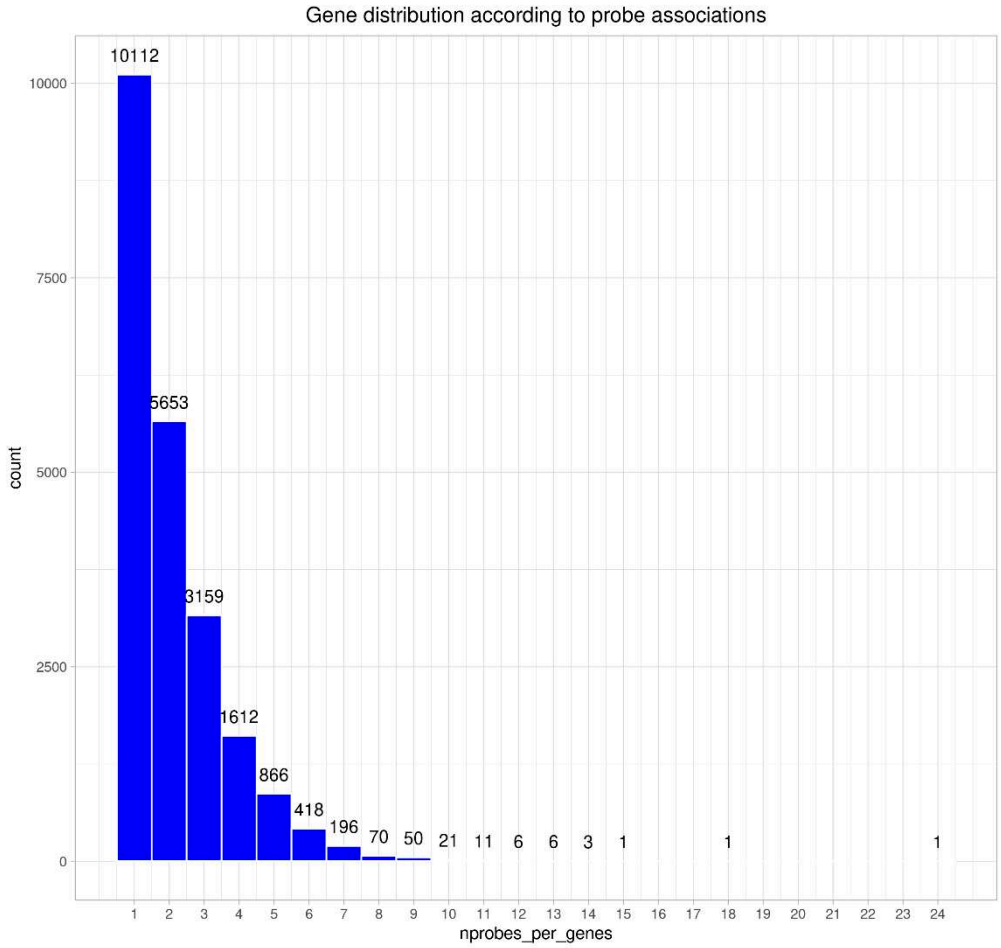

Figure 27 : Distribution du nombre de gènes selon le nombre de sondes auxquelles ils sont associés à l'issue de l'étape d'annotation. *L'annotation a été réalisée à l'aide du package R « AnnotationDbi » et de la base de données d'annotation également sous la forme d'un package R : « hgu133plus2.db ».*

2. Binarisation des données transcriptomiques et préparation du réseau métabolique

La binarisation des données transcriptomiques consiste à convertir les valeurs continues d'intensité d'expression en information binaire (ou catégorielle) du type : « gène fortement exprimé » ou « gène faiblement exprimé ». Bien que la binarisation des données transcriptomiques entraine la perte de la nature quantitative de ces données, elle peut permettre d'éliminer la variabilité technique introduite par les choix algorithmiques réalisés lors du traitement des données transcriptomiques [177]. Dans notre cas, la binarisation des données transcriptomiques est un prérequis à l'utilisation de DEXOM qui comme décrit

précédemment, intègre les données transcriptomiques au réseau métabolique à l'échelle du génome au travers des règles booléennes gène-protéine-réaction et qui nécessite donc une information transcriptomique binarisée.

2.1 Choix de la méthode de binarisation

De multiples approches de binarisation ont été publiées, certaines sont basées sur la définition de seuils [178] et d'autres sur la construction de modèles statistiques [179] ou le calcul de distances [180,181]. Parmi les approches de binarisation par seuil, on distingue deux types de seuils : global et local.

La définition d'un seuil global revient à définir un seuil d'intensité d'expression à partir duquel le gène sera considéré comme fortement exprimé qui est le même pour tous les gènes. Ce type d'approche est généralement appliquée lorsqu'aucune information concernant la distribution des niveaux d'expression de ce gène pour une puce à ADN donnée n'est disponible dans la littérature et que l'on ne dispose pas d'un nombre d'échantillons et de conditions suffisant pour définir des seuils plus précis. Le seuil est alors défini en utilisant la distribution des intensités d'expression pour tous les gènes et en considérant tous les gènes appartenant à un quantile supérieur au quantile seuil comme fortement exprimés. Par exemple, si le quantile seuil est le $75^{ème}$ quantile, alors tous les gènes dont l'intensité d'expression est supérieure à l'intensité d'expression du $75^{ème}$ quantile seront considérés comme fortement exprimés et tous ceux dont l'intensité d'expression est inférieure seront considérés comme faiblement exprimés [182]. Une adaptation couramment utilisée est l'utilisation de deux seuils globaux, l'un en dessous duquel tous les gènes seront considérés comme faiblement exprimés (seuil généralement placé au 25^{th} percentile) et l'autre au-dessus duquel tous les gènes seront considérés comme fortement exprimés (seuil généralement placé au 75^{th} percentile).

La définition d'un seuil local consiste à définir un seuil propre à chaque gène. Pour pouvoir appliquer ce type d'approche, il faut en général disposer d'un grand nombre d'échantillons et de conditions différentes afin d'estimer au mieux la distribution des intensités d'expression de chaque gène ou suffisamment d'information disponible dans la littérature afin de permettre de définir un seuil par gène comme réalisé par les auteurs de Barcode [183–185].

La définition d'un seuil global identique à tous les gènes ne prend pas en compte l'importante variabilité des intensités d'expression existante entre tous les gènes connus chez l'humain. Cela implique que des variations importantes retrouvées pour des gènes ayant systématiquement (*i.e.* indépendamment de son expression ou non dans la cellule) une faible intensité d'expression ne soient pas pris en compte car le seuil général considérera ce gène comme faiblement exprimé dans toutes les conditions. A l'inverse, un gène ayant systématiquement

une forte intensité d'expression sera toujours considéré comme fortement exprimé même si son intensité d'expression est plus faible que celle retrouvée dans d'autres conditions.

Afin de nous affranchir des limites inhérentes à l'utilisation d'un seuil global, nous avons donc choisi d'utiliser la méthode BARCODE qui est basée sur l'utilisation d'un seuil local. BARCODE est une approche qui permet de définir quels gènes peuvent être considérés exprimés/non exprimés à partir de données transcriptomiques générées sur une puce à ADN Affymetrix. Dans un premier temps, McCall et *al,* ont obtenu les intensités d'expression d'un grand nombre d'expériences réalisées sur des puces affymetrix HGU133plus2 (le même travail a également été réalisé pour d'autres puces affymetrix) afin d'obtenir une distribution des intensités d'expression de chaque gène dans un grand nombre de conditions différentes. Pour chacune de ces distributions, la distribution des Z-scores correspondante a été calculée afin d'obtenir des distributions d'intensité d'expression pour chaque gène avec un mode (*i.e.* valeur la plus fréquente dans la distribution) similaire. Le Z-score est une mesure de l'écart d'une valeur donnée par rapport à la moyenne de la population (ici l'ensemble des autres intensités d'expression mesurées pour un gène). Le mode de la distribution des Z-scores correspond aux intensités d'expression que l'on peut qualifier de bruit basal car retrouvées dans la majorité des conditions alors que les valeurs supérieures au mode (situées dans la queue de la distribution) correspondent à des intensités d'expression plus importante que le bruit et signifie que ce gène est plus exprimé qu'habituellement. Comme on peut l'observer sur la figure 28B, la transformation des distributions d'intensité d'expression en distributions de Z-scores a pour avantage d'harmoniser le mode des distributions de tous les gènes et de pouvoir utiliser un seuil global en s'affranchissant des limites évoquées précédemment pour ce type de seuil.

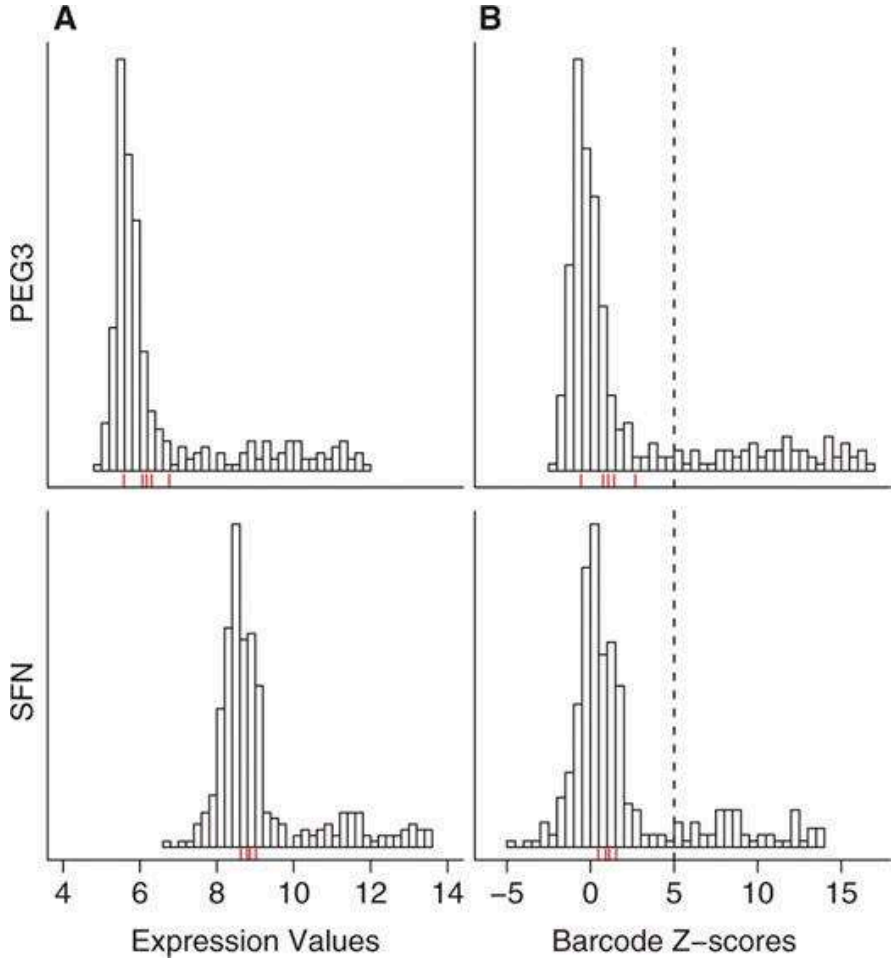

Figure 28 : Distribution des intensités d'expression pour les gènes PEG3 et SFN (A) et distribution des intensités d'expression corrigées par l'approche Barcode pour les gènes PEG3 et SFN (B). *Figure de (McCall et al., 2011).*

Nous avons donc utilisé l'approche BARCODE implémentée dans le package R « frma » pour calculer le Z-score de l'intensité d'expression de chaque gène de chaque échantillon de la base Open TG-GATEs (humain) au regard de des distributions d'intensité d'expression compilées dans BARCODE. Afin d'identifier une liste de gènes fortement et faiblement exprimés pour chaque échantillon, nous avons utilisé deux seuils : les gènes ayant un z-score inférieur au 25ème quantile de la distribution des z-scores de la puce sont considérés comme faiblement exprimés alors que les gènes ayant un z-score supérieur au 75ème quantile de la distribution des z-scores de la puce sont considérés comme fortement exprimés. Les gènes dont la valeur de Z-score est comprise entre le 25ème quantile et le 75ème quantile ne sont donc pas considérés fortement/faiblement exprimés par rapport à leur intensité d'expression retrouvée dans la littérature.

2.2 Limites de la binarisation des données transcriptomiques

Richelle et *al.* [178] a évalué l'influence de chaque choix méthodologique (mapping des gènes, méthode de binarisation, seuil, …) prise lors du traitement des données transcriptomiques préalablement à leur intégration dans les réseaux métaboliques afin de servir de contraintes lors de reconstructions condition-spécifiques par modélisation sous-contraintes. Leurs conclusions suggèrent que c'est la méthode de binarisation choisie par l'utilisateur qui influe le plus sur les résultats. Qui plus est, nous nous sommes aperçus que la binarisation en elle-même pouvait avoir tendance à maximiser des diférences relativement modestes. En effet, les gènes ayant une intensité d'expression très légèrement inférieure au seuil dans une condition et très légèrement supérieure dans une autre impliqueront d'avoir un gène considéré comme ni faiblement exprimé ni fortement exprimé dans un cas et un gène considéré comme fortement exprimé dans l'autre cas alors que quantitativement, la différence d'intensité d'expression est relativement faible. Par exemple, ce cas de figure se présente pour le gène FASN lorsque l'on compare la distribution des z-scores des intensités d'expression de HPH exposées ou non à 7µM d'amiodarone pendant 24h (Fig 29). En considérant les gènes dont le z-score est inférieur au 25ème percentile comme faiblement exprimés et les gènes dont le z-score est supérieur au 75ème percentile comme fortement exprimés, on peut noter que le gène FASN n'est pas considéré comme fortement exprimé en condition contrôle (Fig 29A) alors qu'il est considéré comme fortement exprimé en condition traitée (Fig 29B). Le z-score de FASN pour la condition contrôle est de 3,65 alors qu'en condition traitée, FASN a un z-score de 4,63. La différence entre les deux conditions n'est quantitativement pas très importante mais est donne lieu à une différence importante après la binarisation car elle suffit à dépasser le seuil du 75ème percentile pour la condition traitée. Si l'on avait choisi des seuils différents alors les résultats de la binarisation aurait été différents pour ce gène et donc les contraintes appliquées lors de la modélisation auraient également été différentes.

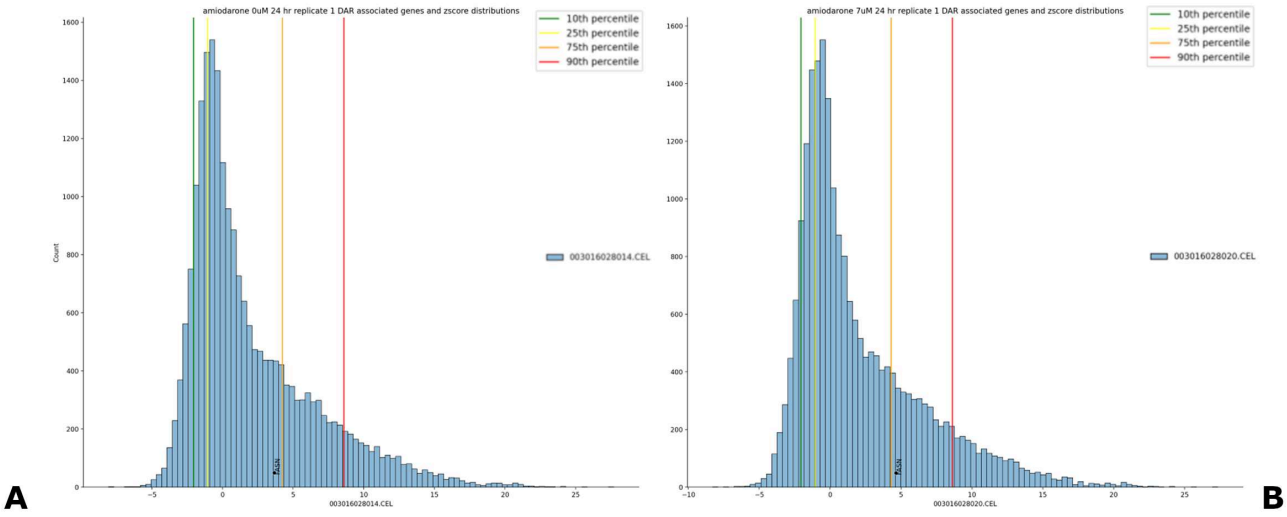

Figure 29 : Comparaison de l'effet du choix du seuil de binarisation à partir de la distribution de z-scores calculés par Barcode pour les intensités d'expression de deux échantillons. *La distribution de gauche (A) correspond à la condition contrôle à 24h et la distribution de droite (B) correspond à la condition exposée à de l'amiodarone à la concentration la plus forte pendant 24h.*

L'impact de la définition d'un seuil est malheureusement un problème assez fréquent. Par example, considérer un gène différentiellement exprimé dès lors que sont intensité d'expression est signficativement différente de celle du contrôle au seuil de p-valeur de 0,05 engendre des biais de seuil similaires à ceux décrits en Figure 29. Améliorer la binarisation des données transcriptomiques pourrait être un axe d'amélioration intéressant à explorer lors de futurs travaux. Il pourrait notamment être intéressant de pondérer les gènes lors de la modélisation sous-contraintes en fonction de leur distance au mode. Par exemple, on pourrait attribuer un poids plus important à un gène fortement exprimé ayant un z-score largement supérieur au seuil à partir duquel un gène est considéré fortement exprimé (*e.g.* dans le 95ème percentile de la distribution) et attribuer un poids plus faible à un gène fortement exprimé mais ayant un z-score tout juste supérieur au seuil à partir duquel un gène est considéré comme fortement exprimé (*e.g.* 76ème percentile de la distribution). Enfin, il pourrait également être intéressant d'intégrer les données transcriptomiques de manière continue à condition de prendre en compte la diversité d'intensité d'expression basale décrite en début de section. Cela pourrait être réalisé en calculant des z-scores sur les intensités d'expression d'Open TG-GATEs comme ceux calculés précédemment avec l'approche BARCODE mais sans appliquer de seuil à ses distributions de zscores, donc sans binariser l'information transcriptomique.

2.3 Préparation du réseau métabolique

Comme discuté en introduction, il existe un grand nombre de réseau métaboliques à l'échelle du génome pour un grand nombre d'espèces différentes. Ici, nous avons choisi d'utiliser Recon2.2 qui est l'un des réseaux métaboliques les plus utilisés pour l'Homme et pour lequel nous disposions d'un bon recul dès le début du projet. Recon2.2 est constitué de 7785 réactions métaboliques, 5323 métabolites et 1675 gènes métaboliques. Dans un premier temps, nous avons obtenu le réseau au format sbml à partir de la base de données Biomodels [186,187] (https://www.ebi.ac.uk/biomodels/MODEL1603150001). A l'aide du module python cobrapy [188], nous avons mis à jour les GPRs de trois réactions contenant des erreurs (*i.e. OIVD1m, OIVD2m, OIVD3m)*. La réaction de production de biomasse de Recon2.2 est construite de manière à représenter la consommation de métabolites nécessaires à la prolifération cellulaire. Cependant, les HPH sont des cellules différenciées et ne sont pas capables de proliférer en conditions de culture normales [17]. Nous avons donc modifié les coefficients stœchiométriques de la réaction de production de biomasse de Recon2.2 de sorte à ce qu'elle représente la maintenance cellulaire et non la prolifération cellulaire. Nous avons fixé la borne inférieure de la réaction de biomasse modifiée à 1 au lieu de sa valeur initiale de 0 de manière à s'assurer qu'elle soit active (avec un flux non nul) et que le modèle prédit soit effectivement capable de produire les métabolites nécessaires à la maintenance cellulaire.

La réaction de production de biomasse de Recon2.2 est la suivante :

$$0{,}706\ M\_Biomass_{protein} + 0{,}014\ M\_Biomass_{DNA} + 0{,}058\ M\_Biomass_{RNA} + 0{,}071\ M\_Biomass_{carbohydrat} + 0{,}097\ M\_Biomass_{lipid} + 0{,}054\ M\_Biomass_{other} \rightarrow 1\ M\_Biomass$$

La maintenance cellulaire ne nécessite pas de duplication de l'ADN, nous avons donc défini le coefficient stoichiomètrique de $M\_Biomass_{DNA}$ à 0 et mis à jour les autres coefficients de la réaction afin de conserver l'équilibre. La réaction de biomasse de Recon2.2 mise à jour afin de représenter la maintenance cellulaire est la suivante :

$$0{,}7160\ M\_Biomass_{protein} + 0{,}0588\ M\_Biomass_{RNA} + 0{,}07201\ M\_Biomass_{carbohydrat} + 0{,}0984\ M\_Biomass_{lipid} + 0{,}05477\ M\_Biomass_{othe} \rightarrow 1\ M\_Biomass$$

Ce modèle Recon2.2 mis à jour en été utilisé pour la suite de nos travaux et notamment pour les reconstructions conditions-spécifiques dont nous discuterons dans une seconde partie. Ces modifications assurent notamment que les modèles développés permettent la production des molécules nécessaires à la composition et donc à la maintenance des cellules.

Chapitre 3 : Développement d'une approche basée sur la modélisation condition-spécifique pour l'étude de l'impact métabolique de xénobiotiques

L'impact d'une exposition à des xénobiotiques sur le métabolisme cellulaire est un élément important à prendre en compte à la fois d'un point de vue santé publique mais également pour l'évaluation de la sécurité des matières premières cosmétiques. En effet, un grand nombre de molécules a la capacité d'affecter le métabolisme à l'échelle de la cellule ou du tissu, ce qui peut engendrer des maladies telles que le diabète, l'obésité ou des dysfonctionnements à l'échelle de l'organe [189–193]. Comme nous avons pu le discuter précédemment, les approches de modélisation sous-contraintes, avec l'intégration de données omiques, peuvent permettre de prendre en compte ces impacts sur le métabolisme cellulaire. Cependant, ces approches présentent des défis à la fois computationnels et analytiques auxquels nous allons tenter d'apporter des solutions au cours de ce chapitre. Dans un premier temps, nous décrirons les développements méthodologiques réalisés pour calculer un ensemble de réseaux métaboliques condition-spécifiques pour chaque condition étudiée. Dans un second temps, nous proposerons une approche permettant de calculer des réactions différentiellement activées entre deux conditions à partir de ces ensembles de réseaux condition-spécifiques.

1. Calcul d'un ensemble de réseaux condition-spécifiques représentatif de l'état métabolique d'une condition étudiée

Au cours de cette section, nous allons détailler les développements méthodologiques réalisées afin de répondre aux défis computationnels que présente l'exploration de l'espace de solutions du problème de modélisation sous-contraintes. Nous allons d'abord décrire les xénobiotiques sélectionnés pour le développement de notre stratégie avant de décrire l'adaptation de la phase d'énumération de DEXOM que nous avons réalisée.

1.1. Choix des molécules et conditions pour l'étude de l'hépatotoxicité

Comme décrit précédemment, nous avons choisi d'utiliser la base de données Open TG-GATEs car il s'agit d'une base de données mettant à disposition une grande quantité de données transcriptomiques d'exposition à un large panel de molécules pharmaceutiques. Il s'agit donc d'une base de données adéquate pour le développement d'approches visant à évaluer l'impact métabolique de molécules cosmétiques et pharmaceutiques. Comme nous avons pu l'aborder en introduction, l'hépatotoxicité peut être utilisée comme un proxy à l'évaluation de la toxicité systémique. Parmi les 158 molécules dont l'exposition a été étudiée sur HPH, nous avons sélectionné 8 molécules connues pour engendrer des effets hépatotoxiques et qui sont bien décrites dans la littérature.

Ces 8 molécules ont également été sélectionnées grâce à l'expertise du Dr Bernard Fromenty (UMR NuMeCan, INSERM) qui a notamment pu nous renseigner sur la littérature disponible concernant les mécanismes d'action à l'origine des phénomènes hépatotoxiques identifiés pour

ces molécules. Nous avons sélectionné l'éthanol, l'acide valproïque, l'indométacine, l'amiodarone, l'allopurinol, la rifampicine, le sulindac et la tetracycline. Les conditions d'exposition disponibles pour ces 8 molécules sont reportées dans le tableau5. 6 de ces 8 molécules disposent de données pour toutes les conditions alors que l'éthanol et le sulindac n'ont pas de données pour le plus court temps d'exposition (2hr) et la plus faible concentration testée (« Low »).

Molécule	**Temps**			**Concentration**			
	2hr	8hr	24hr	Ctrl	Low	Middle	High
Ethanol	**NON**	**OUI**	**OUI**	**OUI**	**NON**	**OUI**	**OUI**
Acide Valproïque	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**
Indométacine	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**
Amiodarone	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**
Allopurinol	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**
Rifampicine	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**
Sulindac	**NON**	**OUI**	**OUI**	**OUI**	**NON**	**OUI**	**OUI**
Tetracycline	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**	**OUI**

Tableau 5 : Résumé des données disponibles pour les 8 molécules sélectionnées. Uniquement le sulindac ainsi que l'éthanol ont des données manquantes pour le plus court temps d'exposition (2hr) et la plus faible dose testée (« Low »).

Etant donné que les HPH ont été exposés sur une durée maximale de 24 heures, les phénomènes toxiques susceptibles d'être identifiés seront des effets de toxicité aigüe. En sachant que les concentrations sélectionnées par les auteurs d'Open TG-GATEs sont des concentrations faiblement cytotoxiques (*i.e.* induisant moins de 20% de cytotoxicité) voire non cytotoxiques pour la majorité des conditions testées (selon les métadonnées), nous avons focalisé notre analyse sur les doses les plus fortes (« High ») et avec la plus longue durée d'exposition (24hr). La prochaine section traitera donc de l'intégration des données transcriptomiques représentant l'exposition de HPH pendant 24h à la plus forte dose n'induisant pas ou peu de cytotoxicité pour les 8 molécules sélectionnées.

1.2. Intégration des données transcriptomiques à Recon2.2

Après avoir normalisé, corrigé et binarisé les données transcriptomiques de la base de données Open TG-GATEs comme décrit dans le chapitre 2, nous avons intégré les données transcriptomiques au réseau métabolique humain (dans notre cas Recon2.2). L'intégration des données transcriptomiques consiste à transformer l'information transcriptomique binarisée en réactions biochimiques actives/inactives en utilisant les GPR associées aux réactions (cf. section 3 du chapitre 1 pour la description des GPRs). Pour cette étape, nous avons utilisé une approche classique selon laquelle : (1) lorsqu'une réaction contient un « AND » dans sa GPR, cette réaction est considérée active si la valeur minimale de l'expression transcriptomique binarisée des gènes de la GPR est égale à 1 (*i.e.* tous les gènes de la GPR sont fortement

exprimés) ; (2) lorsqu'une réaction contient un OR dans sa GPR, cette réaction est considérée active si la valeur maximale de l'expression transcriptomique binarisée des gènes de la GPR est égale à 1 (*i.e.* au moins un gène de la GPR est fortement exprimé).

Par exemple, si l'on considère les gènes suivants dont l'intensité d'expression a été binarisée :

$$Gene1 = 1, Gene2 = -1, Gene3 = 1, Gene4 = -1$$

Avec 1 correspondant à un gène fortement exprimé et -1 à un gène faiblement exprimé.

$$Gene1\ AND\ Gene2 = \min(Gene1, Gene2) = \min(1, -1) = -1 \quad (1)$$

Dans ce cas, la réaction associée à la GPR (1) sera considérée inactive selon les données transcriptomiques car le gène 2 est faiblement exprimé et que l'on a une GPR de type « AND ».

$$Gene1\ OR\ Gene2 = \max(Gene1, Gene2) = \max(1, -1) = 1 \quad (2)$$

Dans ce cas, la réaction associée à la GPR (2) sera considérée active selon les données transcriptomiques car le gène 1 est fortement exprimé et que l'on a une GPR de type « OR »

Ce type de raisonnement s'applique également à des GPRs plus complexes :

$$\begin{aligned}\big((Gene1\ OR\ Gene2)\ AND\ (Gene3\ OR\ Gene4)\big) &= \min(\max(Gene1, Gene2), \max(Gene3, Gene4)) \\ &= \min(\max(1, -1), \max(1, -1)) \\ &= \min(1,1) = 1\end{aligned}$$

Par cette méthode et à partir des données de transcriptomique, nous avons pu identifier pour chaque échantillon une liste de réactions *a priori «* actives » et une liste de réactions *a priori* « inactives ». Ces deux listes de réactions seront utilisées pour contraindre le réseau métabolique Recon2.2 de manière à représenter l'état du métabolisme pour chaque échantillon de chaque condition étudiée.

1.3. Enumération d'un ensemble de sous-réseaux représentatifs d'une condition : adaptation de la méthode d'énumération DEXOM

L'énumération de solutions alternatives pour la modélisation sous-contraintes permet de mieux prendre en compte l'impact d'un composé sur le métabolisme cellulaire en identifiant le maximum de configurations possibles du réseau métabolique (réaction actives / inactives) correspondant aux données de transcriptomique obtenues dans la condition étudiée. Dans notre stratégie, nous avons réalisé l'énumération de solutions alternatives en appliquant et adaptant la méthode DEXOM. Dans son implémentation initiale, DEXOM propose 4 approches d'énumération de solutions différentes.

L'approche d'énumération que nous avons choisi d'utiliser repose sur deux d'entre elles. Tout d'abord, un premier ensemble de solutions alternatives est énuméré avec l'approche de « Reaction-Enum ». Ces solutions servent donc de solutions de départ pour la seconde phase d'énumération appelée « Diversity-Enum » qui recherche graduellement des solutions les plus différentes les unes des autres. Cette approche permet d'énumérer un ensemble de solutions couvrant une large partie de la diversité de solutions existantes dans l'espace de solutions de chacune des conditions modélisées.
Cependant, cette approche d'énumération nécessite la résolution de plusieurs dizaines voire centaines de milliers de MILP ce qui demande d'importantes ressources computationnelles pour des GSMNs de plusieurs milliers de réactions, comme c'est le cas de Recon2.2 (plusieurs centaines de cœurs de calcul et plusieurs jours).
Afin de réduire le coût computationnel de l'énumération et pouvoir ainsi l'appliquer pour l'étude d'un ensemble plus large de molécules, nous avons donc adapté l'approche d'énumération de DEXOM. Nous allons détailler à la fois cette adaptation ainsi que les paramètres permettant de réduire le coup computationnel de l'énumération.

1.3.1. Adaptation de la stratégie d'énumération de DEXOM pour réduire le coût computationnel

Le coût computationnel de DEXOM est lié d'une part au temps nécessaire pour résoudre un MILP (dépendant de la taille du GSMN ainsi que des contraintes) et d'autre part au nombre de MILP qui seront résolus pendant la phase d'énumération (dépendant de la méthode d'énumération choisie) [98]. Le réseau métabolique humain étant par définition de grande taille (*e.g.* Recon2.2 contient 7785 réactions et 5323 métabolites), le coût computationnel lié à la modélisation sous-contraintes et l'exploration de l'espace de solutions associé sera important.

Une façon de réduire le temps de calcul pourrait être de modifier les seuils de binarisation utilisés pour les données de transcriptomique. Par exemple, nous aurions pu choisir deux seuils de binarisation plus contraignants (*e.g.* 10^{th} et 90^{th} percentiles au lieu des seuils actuels de 25^{th} et 75^{th} percentiles) ce qui aurait eu pour effet de réduire le nombre de gènes considérés comme faiblement ou fortement exprimés et par extension le nombre de contraintes imposées par les données transcriptomiques lors de la modélisation. Cela aurait donc permis de trouver plus facilement des solutions au risque d'obtenir des réseaux conditions-spécifiques moins représentatifs de la condition biologique modélisée.

Bien que la définition d'un seuil soit souvent subjective, nous avons choisi de ne pas relâcher les contraintes mais de plutôt adapter la manière d'énumérer les solutions alternatives ainsi que d'adapter les paramètres de DEXOM afin de limiter les temps de calculs. En effet, sans

énumération partielle un seul MILP par condition est résolu par le solveur soit environ 25 secondes avec les contraintes et les paramètres décrits précédemment et l'utilisation de CPLEX, l'un des solveurs les plus performants, pour un ordinateur classique disposant d'un processeur avec 4 cœurs physiques. Faire de l'énumération partielle implique de résoudre plusieurs milliers de MILP par condition ce qui signifie multiplier ce temps de calcul. Augmenter la puissance de calcul en utilisant des centres de calcul disposant de centaines de cœurs est une solution lorsque le nombre de conditions à modéliser est faible. Cependant dès lors qu'il est nécessaire de modéliser plusieurs dizaines de conditions il devient impératif de réduire le coût computationnel. C'est pourquoi nous avons adapté DEXOM afin de réduire le coût computationnel de l'énumération.

Selon Rodriguez *et al.* Reaction-Enum est l'approche d'énumération la plus rapide et la seconde plus performante en termes de diversité de solutions, derrière Diversity-Enum. Nous avons donc opté dans un premier temps pour une énumération avec Reaction-Enum (en bloquant successivement l'ensemble des réactions du modèle).

Diversity-Enum étant la phase d'énumération la plus chronophage, nous avons limité le nombre de solutions alternatives recherchées par Diversity-Enum en utilisant seulement 1% des solutions énumérées par Reaction-Enum comme solutions de départ. Ces solutions ont été sélectionnées par une stratégie d'échantillonnage systématique décrite en Figure 30.

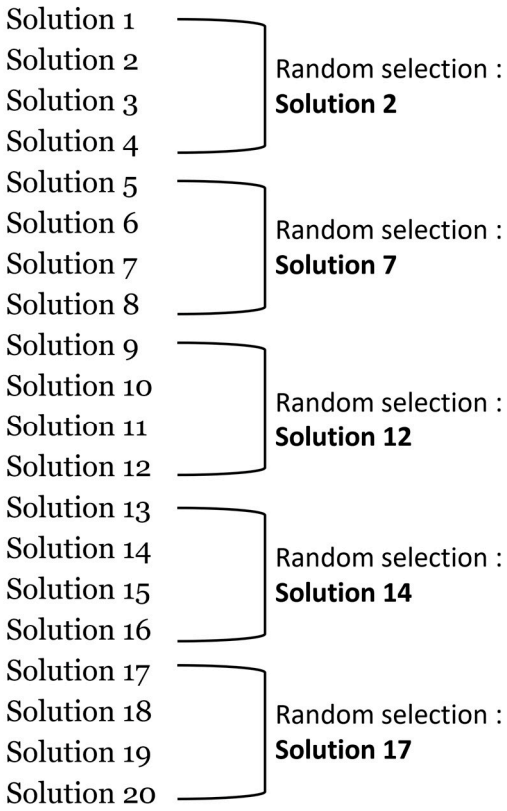

Figure 30 : Schéma du fonctionnement de notre adaptation de l'échantillonnage systématique sur un ensemble jouet de 20 solutions. *Cet exemple jouet est constitué de 20 solutions, divisées en 5 intervalles lots de solutions. Pour chaque intervalle, une solution sera sélectionnée aléatoirement.*

L'échantillonnage systématique est une approche d'échantillonnage dont l'objectif est de sélectionner chaque élément à intervalle régulier. Nous avons adapté cette approche pour sélectionner une solution au hasard dans un intervalle régulier (Fig 30). Cette méthode de d'échantillonnage a été choisie car Reaction-Enum bloque successivement chaque réaction (et calcule une solution alternative si possible) en itérant sur la liste de réactions du modèle, ordonnée alphabétiquement. Les solutions alternatives calculées sont stockées au fur et à mesure de l'itération, ce qui implique que l'ordre des solutions alternatives suit également un ordre alphabétique. L'identifiant des réactions étant souvent informatif de la fonction de la réaction, l'ordre dans lequel les réactions sont stockées (alphabétique) n'est pas aléatoire mais indirectement lié à la fonction métabolique. De fait il est important de prendre en compte cet ordre lors de la sélection des solutions de départ pour Diversity-Enum afin de maximiser la diversité fonctionnelle des solutions de départ. Une approche alternative pourrait être de mélanger aléatoirement la liste de réactions à bloquer par Reaction-Enum avant de sélectionner aléatoirement un ensemble de solutions parmi toutes les solutions.

Après avoir identifié un ensemble de solutions correspondant à 1% des solutions calculées par Reaction-Enum, nous avons utilisé ces solutions de départ pour la phase d'énumération par Diversity-Enum (Fig 31). Ainsi, pour chaque solution de départ de Reaction-Enum

sélectionnée, 100 solutions alternatives obtenues par Diversity-Enum ont été calculées. Ces solutions alternatives supplémentaires calculées par l'approche Diversity-Enum permettent une meilleure couverture de la diversité existante dans l'espace de solutions en maximisant graduellement la distance entre les solutions.

La combinaison de ces deux approches d'énumération, permet notamment d'énumérer des solutions alternatives de plus en plus différentes et donc d'obtenir une meilleure couverture de la diversité de l'espace de solutions.

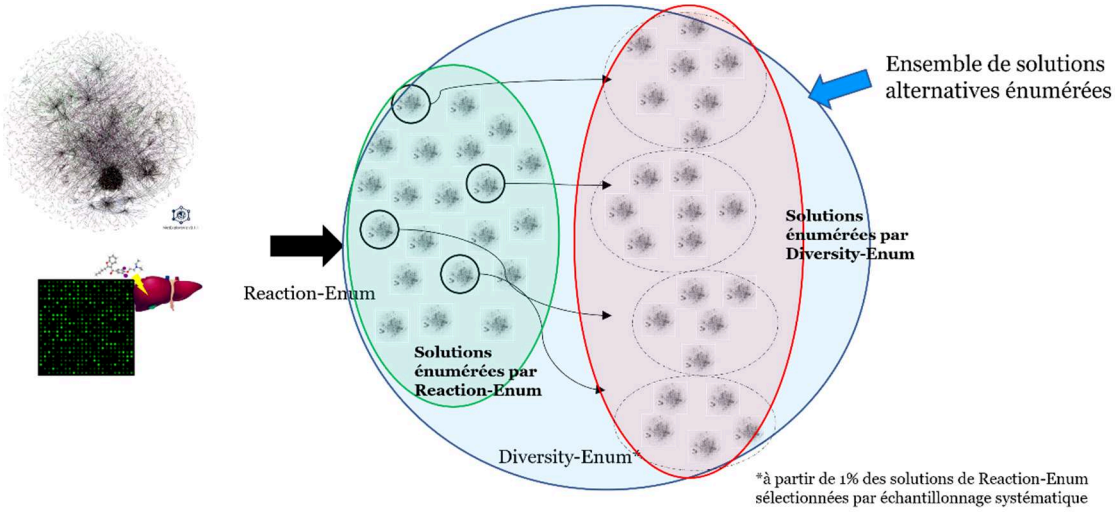

Figure 31 : Schéma de l'adaptation de l'approche d'énumération partielle par DEXOM. *A partir d'un GSMN et de données transcriptomiques binarisées, DEXOM est utilisé pour calculer un premier ensemble de solutions (en vert sur la Figure 31) via l'approche de Reaction-Enum. Une approche d'échantillonnage systématique est ensuite utilisée pour sélectionner 1% des solutions calculées par Reaction-Enum qui servent ensuite de solutions de départ pour l'approche de Diversity-Enum. L'approche de Diversity-Enum calcule par défaut 100 solutions alternatives graduellement plus distantes les unes des autres. Sur cette figure les solutions sont représentées par des sous-réseaux. Les sous-réseaux faisant partie de l'ellipse verte on été énumérés avec l'approche de Reaction-Enum et les sous-réseaux faisant partie de l'ellipse rouge ont été énumérés avec l'approche de Diversity-Enum en partant d'un sous-ensemble de solutions de départ prise dans les solutions calculées par Reaction-Enum. L'union des solutions de Reaction-Enum et Diversity-Enum est représentée par l'ellipse bleue.*

1.3.2. Choix des paramètres de modélisation pour DEXOM

Le choix des paramètres est basé, au même titre que le choix du protocole d'exploration de l'espace de solutions, sur la recherche d'un compromis entre optimalité de la solution trouvée par rapport aux données expérimentales et temps nécessaire au solveur pour trouver une solution optimale. Lors de la recherche d'une solution, le solveur cherche à maximiser la fonction objective définie pour le MILP. Cette fonction objective à une valeur maximale théorique qui correspond à l'adéquation parfaite entre les réactions actives/inactives selon les données transcriptomiques et les réactions actives/inactives selon la topologie du réseau métabolique.

Une solution sera considérée comme optimale si la différence entre son score d'optimalité et le maximum théorique est inférieur à un paramètre nommé « mipgaptol ». Pour ce paramètre

nous avons défini une valeur de 10^{-3} soit une différence devant être inférieure à 0,1% de la valeur optimale théorique. Par exemple, si l'optimum théorique est de 3000, une valeur de tolérance « mipgaptol » à 0,1% correspond à une tolérance de 3 réactions. Cela signifie qu'une solution sera considérée comme optimale si au moins l'activité de 2997 réactions (sur 3000) est en adéquation entre les données transcriptomiques et la topologie du réseau.
Être plus permissif sur le paramètre « mipgaptol » permet de trouver des solutions plus rapidement (division du temps de calcul par deux ou plus avec un mipgaptol à 10^{-2}) mais les solutions trouvées seront alors plus éloignées de la valeur optimale théorique et donc plus loin de la condition à modéliser.
Etant donné que certains MILP peuvent être très difficiles à résoudre, les auteurs de DEXOM ont défini un temps limite pour la résolution du MILP. Le paramètre « tlim » défini donc le temps maximal autorisé pour résoudre un MILP. La recherche de solutions sera interrompue si le seuil est dépassé. Nous avons conservé la valeur par défaut qui est égale à 600 secondes (10 minutes).

Après avoir réalisé les adaptations décrites ci-dessus, la phase d'énumération calcule environ 10 000 solutions alternatives par échantillon. Bien que ce nombre de solutions alternatives soit moins important que ce qu'il aurait été avec une utilisation classique de DEXOM (Reaction-Enum + Diversity-Enum), l'analyse de ces milliers de réseaux métaboliques condition-spécifiques n'est par trivial et nécessite le développement de nouvelles méthodes d'analyse que nous allons aborder dans la section suivante traitant de la caractérisation de la perturbation métabolique.

1.3.3. Conclusion

Notre adaptation consiste donc à réduire la phase de Diversity-Enum afin de réduire le nombre de MILP à résoudre qui est très important pendant cette phase d'énumération. Cependant, réduire le nombre de MILP implique forcément une réduction du nombre de solution alternatives énumérées. Afin d'évaluer l'éventuelle perte en terme de qualité induite par cette exploration partielle de l'espace de solutions, il serait intéressant de reproduire les résultats de l'article initial de DEXOM publié par Rodriguez *et al.*. En effet, nous pouvons supposer qu'en ayant limité la recherche de solutions distantes par Diversity-enum, la diversité de solutions alternatives prédite par notre version modifiée de DEXOM sera diminuée.

2. Caractérisation de la perturbation métabolique par l'identification de réactions différentiellement activées

Bien que nécessaire, l'énumération de solutions alternatives pose un défi pour l'analyse des résultats. En effet, avec une approche de modélisation sous-contraintes classique telle que iMAT, l'état une métabolique est représenté par un seul réseau métabolique condition-spécifique. Cependant, en réalisant une phase d'énumération partielle des solutions alternatives avec DEXOM, une dizaine de milliers de réseaux condition-spécifiques est calculée pour chaque solution étudiée. De tels résultats sont donc difficilement interprétables en l'état à cause de leur grande dimensionnalité et de leur quantité. Puisque l'objectif est d'étudier l'impact métabolique de xénobiotiques, nous allons nous intéresser à l'étude des réactions biochimiques dont l'activité a été perturbée et qui seront donc des marqueurs de changements métaboliques induits par ces composés hépatotoxiques. Cependant, passer de dizaines de milliers de réseaux constitués de plusieurs milliers de réactions à une interprétation à l'échelle de la réaction métabolique représente un réel défi en termes de réduction de dimensionnalité. Une telle dimensionnalité existe également, dans une moindre mesure, dans d'autres types de données omiques. Il pourrait par exemple être intéressant de s'intéresser aux approches de réduction de dimensionnalité utilisées pour traiter des données transcriptomiques et notamment les approches permettant d'identifier des gènes différentiellement exprimés entre deux conditions. Ces approches permettent notamment de traiter des données de grande dimensionnalité d'un point de vue du nombre de variables mesurées mais présentent cependant des limites statistiques associées aux jeux de données contenant un très grand nombre d'échantillons comme nous allons le discuter au cours des prochaines sections. Cependant, nous allons tout de même emprunter cette philosophie d'identification de perturbations entre deux conditions par la mise au point d'une approche d'identification de réactions différentiellement activées (DARs) entre une condition traitée et sa condition contrôle qui soit adaptée à la grande dimensionnalité des résultats de modélisation condition-spécifique avec énumération partielle.

2.1. Recherche d'une métrique robuste pour l'identification de réactions perturbées

L'identification d'une métrique robuste est un élément clé pour l'identification des réactions perturbées par l'exposition à un xénobiotique. La robustesse de cette métrique aura un impact important sur les analyses car si des réactions sont considérées significativement perturbées entre deux conditions alors qu'elles ne le sont pas réellement, cela pourrait induire des biais dans les analyses mécanistiques qui suivent. L'identification de perturbations entre deux conditions est une problématique connue pour plusieurs types de données omiques. L'exemple

le plus connu est l'utilisation de tests statistiques afin d'identifier des gènes significativement dérégulés entre deux conditions. Dans ce cas, l'objectif est de déterminer pour chaque gène si l'on peut rejeter l'hypothèse nulle H0 « Il n'y a pas de différence d'expression génique entre les deux échantillons » avec un risque de se tromper (conclure à une différence d'expression génique alors qu'il n'y en a pas) inférieur à 5%.

Cependant, l'utilisation de tests statistiques « paramétriques » nécessite au préalable de vérifier que les données permettant de réaliser le test suivent la distribution attendue par le test statistique utilisé (*e.g.* une loi normale, négative binomiale, …), ce qui n'est pas toujours le cas.

Dans la prochaine section, nous allons discuter de l'utilisation de tests statistiques pour l'identification de réactions perturbées à partir des résultats d'énumération partielle et notamment de l'identification de certains biais les rendant difficilement applicable aux données d'énumération partielles.

2.1.1. Méthodes statistiques : biais des p-valeurs et limites du calcul de rapport des cotes

La première étape consiste à choisir le test adapté aux données. Les données issues de l'énumération partielle sont des données qualitatives (*i.e.* la valeur indique l'appartenance à une catégorie plutôt qu'une quantité), avec un très grand nombre de solutions qui ne sont pas indépendantes les unes des autres car obtenues avec les mêmes contraintes biologiques. Cependant les échantillons de deux conditions différentes peuvent être considérés comme indépendants les uns des autres car il s'agit d'échantillons biologiques différents (pour rappel nous avons 2 échantillons par condition). Les tests les plus utilisés pour comparer deux groupes de variables qualitatives indépendantes sont le test du Chi^2 [194] et le test exact de Fisher [195]. Ces deux tests permettent l'analyse de tableaux de contingence. Les tableaux de contingence sont des tableaux permettant de représenter les données catégorielles sous la forme de « comptage ». Par exemple, dans le tableau de contingence ci-dessous (Tableau 6), la réaction est active dans 700 solutions et inactive dans 300 solutions pour la condition traitée alors que cette même réaction est active dans seulement 108 solutions et inactive dans 898 solutions de la condition contrôle. Ce tableau permet d'estimer la dépendance ou l'indépendance statistique (est-ce que le traitement influe sur le statut actif/inactif de la réaction) entre deux conditions (ici la condition traitée et la condition contrôle).

	Traitement	***Contrôle***
Active	700	108
Inactive	300	898

Tableau 6. Exemple d'un tableau de contingence. *Cet exemple correspond à une réaction prédite active dans 700 solutions sur 1000 pour la condition traitée et prédite active dans 108 solutions sur 1000 pour la condition contrôle.*

Le test Exact de Fisher permet alors de calculer la probabilité que l'hypothèse nulle H0 : « Les deux groupes sont équivalents », soit vraie. Comme pour la majorité des tests statistiques, on admet un risque alpha de 5%, indiquant que l'on rejette l'hypothèse H0 (*i.e.* en rejetant H0, on considère que les deux groupes ne sont pas équivalents) au risque de se tromper de 5%, soit une p-valeur inférieure ou égale à 0,05.

Le test exact de Fisher permet donc théoriquement de comparer l'activité d'une réaction entre deux ensembles de solutions correspondant à deux conditions différentes. Cependant, à cause du nombre important de solutions énumérées par condition (*i.e.* environ 20 000 solutions par condition), une des limites du calcul des p-valeurs a été atteinte. En effet, le calcul de la p-valeur est sensible à la taille des échantillons [196–198]. Cela signifie que plus les échantillons sont de grande taille et plus la p-valeur tendra vers 0 sans que cela ne soit observé quantitativement sur la différence entre les deux échantillons (mesurée par le rapport des cotes, également appelé Odds-Ratio).

Le rapport des cotes se définit comme le rapport de la cote qu'un évènement arrive à un groupe par rapport à la cote que ce même événement arrive à un autre groupe. La cote correspond au ratio entre la probabilité qu'un évènement se produise et la probabilité qu'il ne se produise pas. Par exemple si un cheval à 1 chance sur 4 de gagner alors il aura 3 chances sur 4 de perdre ce qui revient à la cote suivante : $\frac{1/4}{3/4} = \frac{1}{3}$ que l'on ramène généralement à 1 en parlant de cote à 3 contre 1.

A partir d'un tableau de contingence, le calcul du rapport des cotes est :

$$RC = \frac{cote\ de\ survenue\ de\ l'évènement\ dans\ la\ condition\ 1}{cote\ de\ survenue\ de\ l'évènementdans\ la\ condition\ 2}$$

Par exemple, dans le cas de la comparaison de l'activation d'une réaction dans l'ensemble de solutions d'une condition par rapport à une autre (Tableau 6), le calcul du rapport de cotes est :

$$RC = \frac{700/300}{108/898} = 19{,}4$$

Ce qui signifie que la cote correspondant à l'activation de la réaction dans la condition traitée est plus de 19 fois plus élevée que la cote correspondant à l'activation de cette même réaction dans la condition contrôle.

Concernant le test de Fisher, si l'on prend plusieurs exemples en faisant varier la taille des échantillons sans faire varier la proportion de réactions actives/inactives entre les échantillons (tableaux 7, 8 et 9), nous observons que la p-valeur, issue du test de Fisher, tend effectivement vers 0 au fur et à mesure de l'augmentation de la taille des échantillons sans que le rapport de côte, qui est une mesure de la différence entre les deux échantillons ne change réellement (Fig 32).

	Traitement	***Contrôle***
Active	14	13
Inactive	6	7

Tableau 7. Exemple d'un tableau de contingence avec 20 échantillons pour chaque condition. *Cet exemple correspond à une réaction prédite active dans 14 solutions sur 20 pour la condition traitée et prédite active dans 13 solutions sur 20 pour la condition contrôle.*

Pour ce tableau de contingence (Tableau 7), le résultat du test exact de Fisher est le suivant :

p-valeur = 1, rapport des cotes (Odds-Ratio) = 1.25

	Traitement	***Contrôle***
Active	140	130
Inactive	60	70

Tableau 8. Exemple d'un tableau de contingence avec 200 échantillons pour chaque condition. *Cet exemple correspond à une réaction prédite active dans 140 solutions sur 200 pour la condition traitée et prédite active dans 103 solutions sur 200 pour la condition contrôle.*

Pour ce tableau de contingence (Tableau 8), le résultat du test exact de Fisher est le suivant :

p-valeur = 0.34, rapport des cotes (Odds-Ratio) = 1.26

	Traitement	***Contrôle***
Active	14000	13000
Inactive	6000	7000

Tableau 9. Exemple d'un tableau de contingence avec un nombre d'échantillons équivalent à celui obtenu en pratique avec l'énumération partielle adaptée de DEXOM. *Cet exemple correspond à une réaction prédite active dans 14000 solutions sur 20000 pour la condition traitée et prédite active dans 13000 solutions sur 20000 pour la condition contrôle.*

Pour ce tableau de contingence (Tableau 9), le résultat du test exact de Fisher est le suivant **:**

p-valeur = $1.43*10^{-26}$, rapport des cotes (Odds-Ratio) = 1.26

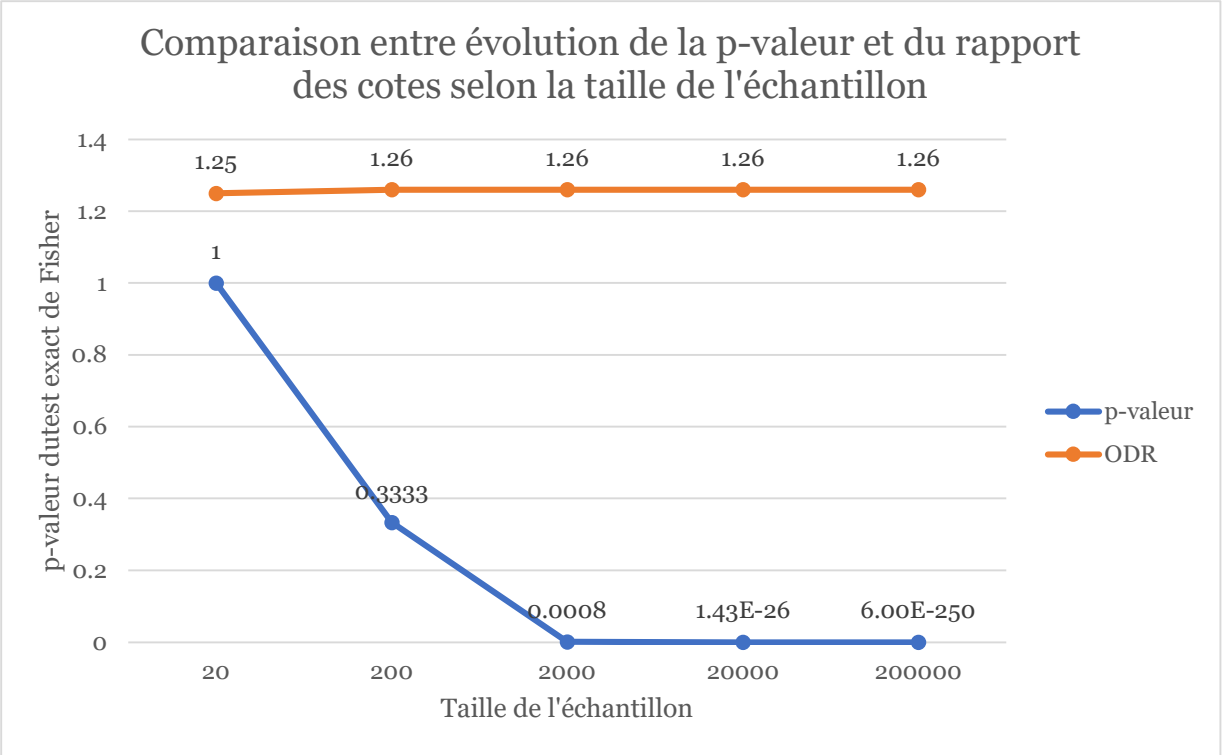

Figure 32: Comparaison entre évolution de la p-valeur et du rapport des cotes (noté ODR sur la figure 32) selon le nombre d'échantillons (sans variation de la proportion de différence entre les deux échantillons).

Contrairement à la p-valeur, le rapport des cotes n'est pas sensible au nombre d'échantillons dans chaque condition. Nous avons donc évalué la pertinence de cette métrique pour identifier des réactions perturbées entre deux ensembles de réseaux condition-spécifiques.

Nous avons envisagé d'utiliser cette métrique afin d'identifier des réactions perturbées en fixant un seuil minimal à partir duquel on considère que la différence entre les deux distributions représentées par le rapport des cotes est significative. Cependant, le calcul du rapport des cotes est sensible aux événements rares [199].

Prenons par exemple deux tableaux de contingence ayant un rapport des cotes similaires (plus ou moins égal à 3) :

	Traitement	***Contrôle***
Active	19000	16000
Inactive	1000	2500

Tableau 10. Exemple d'un tableau de contingence avec deux conditions ne contenant pas d'évènement rare et dont le rapport des cotes est égal à 2.97. *Cet exemple correspond à une réaction prédite active dans 19000 solutions sur 20000 pour la condition traitée et prédite active dans 16000 solutions sur 18500 pour la condition contrôle.*

Pour ce tableau de contingence (Tableau 10), le rapport des cotes est de 2.97**.**

	Traitement	***Contrôle***
Active	19999	19997
Inactive	1	3

Tableau 11. Exemple d'un tableau de contingence avec des conditions contenant des évènements rares et dont le rapport des cotes est égal à 3. *Cet exemple correspond à une réaction prédite active dans 19999 solutions sur 20000 pour la condition traitée et prédite active dans 19997 solutions sur 20000 pour la condition contrôle.*

Pour ce tableau de contingence (Tableau 11), le rapport des cotes est de 3.

On s'aperçoit donc que la présence d'évènements rares génère un rapport des cotes important alors que les deux distributions du tableau de contingence (Tableau 11) sont quasiment identiques. Pratiquement, cela signifie que l'inactivation de la réaction considérée dans seulement une ou deux solutions parmi l'ensemble des solutions alternatives énumérées suffit pour considérer cette réaction comme significativement perturbée dans une condition par rapport à une autre alors qu'un seul cas (ou trois cas) d'inactivation sur 20 000 suggère plutôt que cette réaction est active à la fois dans la condition « traitement » et dans la condition « contrôle ». Etant donné que ce cas de figure peut exister dans les résultats d'énumération partielle, nous avons choisi de ne pas utiliser le rapport des cotes comme métrique pour identifier les réactions différentiellement activées entre deux conditions. Les prochaines sections présentent des alternatives aux approches statistiques discutées ci-dessus et s'appuyant sur le calcul de fréquences d'activation.

2.1.2. Fréquences d'activation des réactions : calcul et comparaison

Comme décrit précédemment, les résultats d'une énumération partielle réalisée par DEXOM (ou son adaptation) sont stockés sous la forme d'un tableau de vecteur binaires. Les lignes correspondent aux solutions alternatives et les colonnes correspondent aux réactions de Recon2.2. On note la fréquence d'activation d'une réaction f_{act} qui correspond au nombre de solutions dans laquelle cette réaction est prédite active ($n_{Ractives}$) divisé par le nombre total de solutions alternatives énumérées pour cette condition (*i.e.* équivalent à $n_{Ractives} + n_{Rinactives}$) n_{RTotal}.

$$\boldsymbol{f_{act} = \frac{n_{Ractives}}{n_{RTotal}}}$$

En partant de cette métrique plutôt simple mais adaptée à la nature ensembliste des résultats issus de l'énumération partielle, nous avons cherché des métriques permettant de comparer ces fréquences d'activations.

2.1.2.1. Comparaison de fréquences d'activation : R2

Calculer une fréquence d'activation pour une réaction peut nous donner une idée de l'importance d'une réaction dans une condition puisqu'une réaction essentielle à la survie de la cellule sera systématiquement active (f_{act} = 1) alors que des réactions qui ne sont pas essentielles (*i.e.* dont les fonctions peuvent être remplacées par d'autres réactions ou qui sont moins ou pas utiles dans la condition donnée) seront peu ou pas actives, ou de manière aléatoire en fonction des solutions. Cependant, pour identifier une différence entre deux conditions il est nécessaire de pouvoir comparer les fréquences d'activation calculées pour chacune des réactions à partir de l'ensemble de solutions de chacune des deux conditions. Pour se faire et en collaboration avec l'équipe de calcul scientifique de L'Oréal, nous avons définis deux métriques permettant d'identifier des réactions différentiellement activées entre deux conditions à partir des fréquences d'activation.

La première métrique, qui est la plus simple, consiste à calculer le carré des différences des fréquences d'activation :

$$R2 = \left(\frac{nActive_{ctrl}}{nTotal_{ctrl}} - \frac{nActive_{trt}}{nTotal_{trt}}\right)^2 = (f_{ctrl} - f_{trt})^2$$

Mettre la différence des fréquences d'activation au carré pénalise les faibles différences sans impacter les différences importantes, ce qui permet d'avoir une métrique plus conservative. Une valeur de R2 égale à 0 correspond à une absence totale de différence entre la fréquence d'activation de la réaction pour la condition contrôle et pour la condition traitée. A l'inverse, une valeur de R2 égale à 1 correspond à une différence totale entre la fréquence d'activation de la réaction pour la condition contrôle et pour la condition traitée. Une différence de 50% correspond à un R2 de 0,25.

La figure 33 permet de visualiser le comportement de la métrique R2 en fonction des fréquences dans les 2 conditions. A noter que tous les points situés sur le tracé d'une droite (Fig 33) ont la même valeur de R2, ce qui permet de visualiser l'évolution de la métrique selon les combinaisons de fréquence d'activation dans la condition traitée et la condition contrôle.

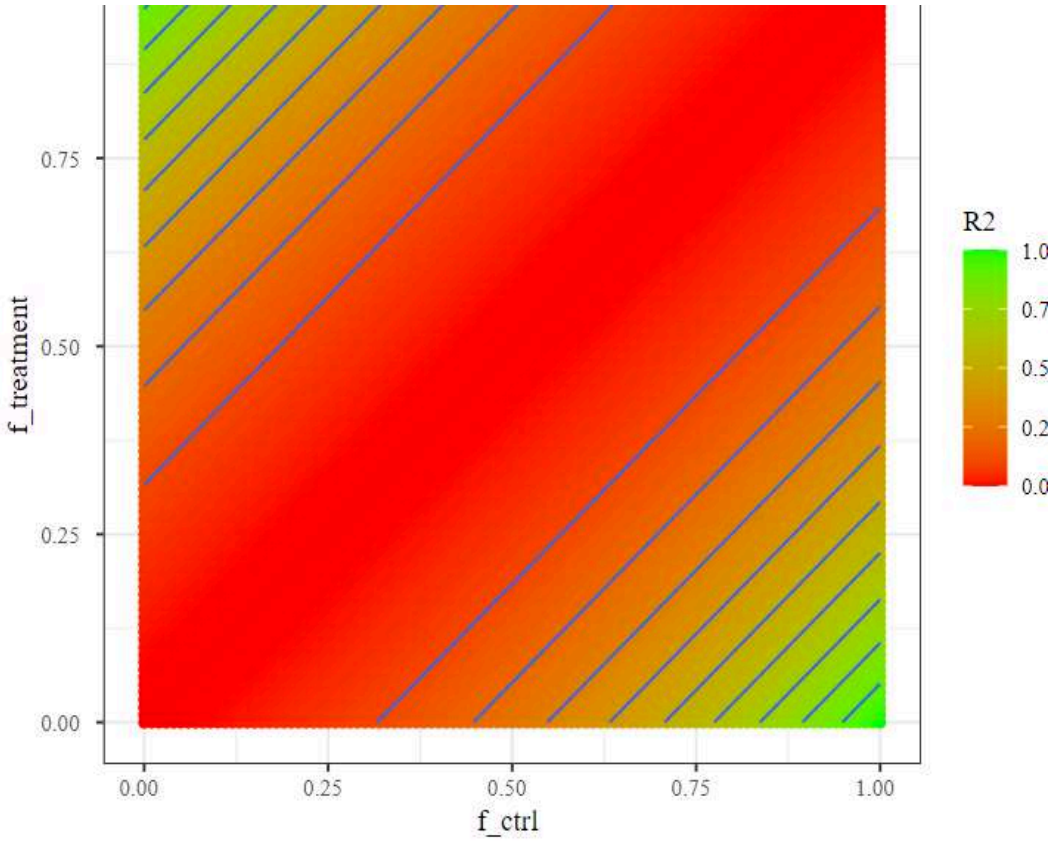

Figure 33 : Simulation des valeurs de R2 possibles pour des fréquences comprises entre 0 et 1 dans 2 conditions (contrôle : f_ctrl et traitement : f_treatment).

Bien que le R2 soit une métrique répondant à nos attentes en étant simple et plutôt conservative, nous avons identifié un point d'amélioration à cette métrique. En effet, une réaction ayant une fréquence d'activation égale ou presque à zéro dans une condition et une fréquence d'activation bien supérieure à 0 (*e.g.* égale à 0,25) dans l'autre condition pourrait représenter une perturbation significative car passant d'une inactivité totale à une activité partielle. En effet, le blocage/déblocage total d'une réaction est un effet pouvant impacter de manière importante le métabolisme cellulaire comme cela peut être le cas avec les maladies métaboliques génétiques [200], d'autant plus si la réaction impactée n'est pas une réaction redondante. De fait, même si la variation de fréquence d'activation est plus faible, il nous semble intéressant de prendre en compte ce type d'effet dans l'étude du mMoA.

La métrique du R2 étant directement liée à la valeur de la différence entre les fréquences d'activation, il faudrait drastiquement baisser le seuil pour capturer ces réactions avec un blocage/déblocage total dans une condition et donc également considérer comme DAR des réactions avec une faible différence de fréquence d'activation et sans cas de blocage/déblocage total.

2.1.2.2. Comparaison de fréquences d'activation par les propriétés de l'équation du cercle

Afin de pouvoir capturer ces perturbations de plus faible ampleur (*i.e.* une différence de fréquence d'activation relativement peu importante) mais passant d'un statut toujours inactif à faiblement actif, nous avons recherché avec le Dr Alban Ott du département de calcul scientifique de L'Oréal, une métrique légèrement plus permissive.

La solution trouvée se base sur les propriétés de l'équation d'un cercle que nous allons développer ci-dessous.

On définit l'équation générale d'un cercle par :

$$x^2 + y^2 + 2ax + 2by + c = 0 \quad (1)$$

Et on peut définir l'équation d'un cercle de centre (h,k) et de rayon r par :

$$r^2 = h^2 + k^2 - c \quad (2)$$

Connaissant les coordonnées de trois points placés sur un cercle, il est possible de calculer le rayon et le centre du cercle sur lequel sont placés ces trois points. En effet, si ces points sont placés sur le cercle alors ils doivent satisfaire l'équation du cercle et il est alors possible de calculer a, b et c (se référer à [201] pour le développement complet). En considérant trois points de coordonnées (x1, y1), (x2, y2), (x3, y3) il est possible de définir les équations suivantes :

$$a = \frac{(x1^2 - x3^2)(x1^2 - x2^2) + (y1^2 - y3^2)(x1 - x2) + (x2^2 - x1^2)(x1 - x3) + (y2^2 - y1^2)(x1 - x3)}{\left(2*((y3 - y1)(x1 - x2) - (y2 - y1)(x1 - x3))\right)} \quad (3)$$

$$b = \frac{(x1^2 - x3^2)(y1 - x2) + (y1^2 - y3^2)(y1 - y2) + (x2^2 - x1^2)(y1 - y3) + (y2^2 - y1^2)(y1 - y3)}{\left(2*((x3 - x1)(y1 - y2) - (x2 - x1)(y1 - y3))\right)} \quad (4)$$

$$c = -(x1^2) - (y1^2) - 2ax1 - 2by1 \quad (5)$$

On peut donc calculer les coordonnées du centre de ce cercle telles que :

$$h = -b$$
$$k = -a$$

On peut également calculer le rayon du cercle sur lequel sont placés les 3 points par :

$$r = \sqrt{h^2 + k^2 - c} \quad (6)$$

En appliquant ces principes, il est possible d'utiliser les propriétés d'un cercle et notamment ce lien entre les coordonnées de trois points et le rayon d'un cercle afin de comparer les

fréquences d'activation d'une réaction dans deux conditions différentes. L'objectif est de fixer deux de ces trois points de sorte à ce que les propriétés du cercle (position du centre et diamètre) soient définies par le troisième point dont les coordonnées seront définies par la fréquence d'activation d'une réaction d'intérêt dans deux conditions différentes.

On fixe les points A et B à des coordonnées arbitraires sur le cercle telles que :

$Point\ A = (-1, -1)$
$Point\ B = (1,1)$
Les coordonnées du troisième point sont définies par les fréquences d'activation de la réaction d'intérêt dans les deux conditions étudiées :

$Point\ C = (f_{ctrl}, f_{treat})$
Par exemple, prenons le cas d'une réaction ayant des fréquences d'activation similaires dans les deux conditions, les points A et B resterons inchangés et le point C aura pour coordonnées les fréquences d'activation en condition contrôle et en condition traitée :

$Point\ C = (0{,}45, 0{,}55)$
Ce qui grâce aux propriétés décrites ci-dessus et implémentées dans une fonction donnera un cercle de centre (7,47, -7,47) avec un rayon de 10,57 (Fig 34).

Prenons maintenant le cas d'une réaction ayant des fréquences d'activation très différentes dans les deux conditions, les points A et B resterons inchangés et le point C aura pour coordonnées les fréquences d'activation en condition contrôle et en condition traitée :

$Point\ C = (0{,}2, 0{,}9)$
Le cercle correspondant à ces coordonnées sera un cercle de centre (0,82, -0,82) avec un rayon de 3,66 (Fig 34).

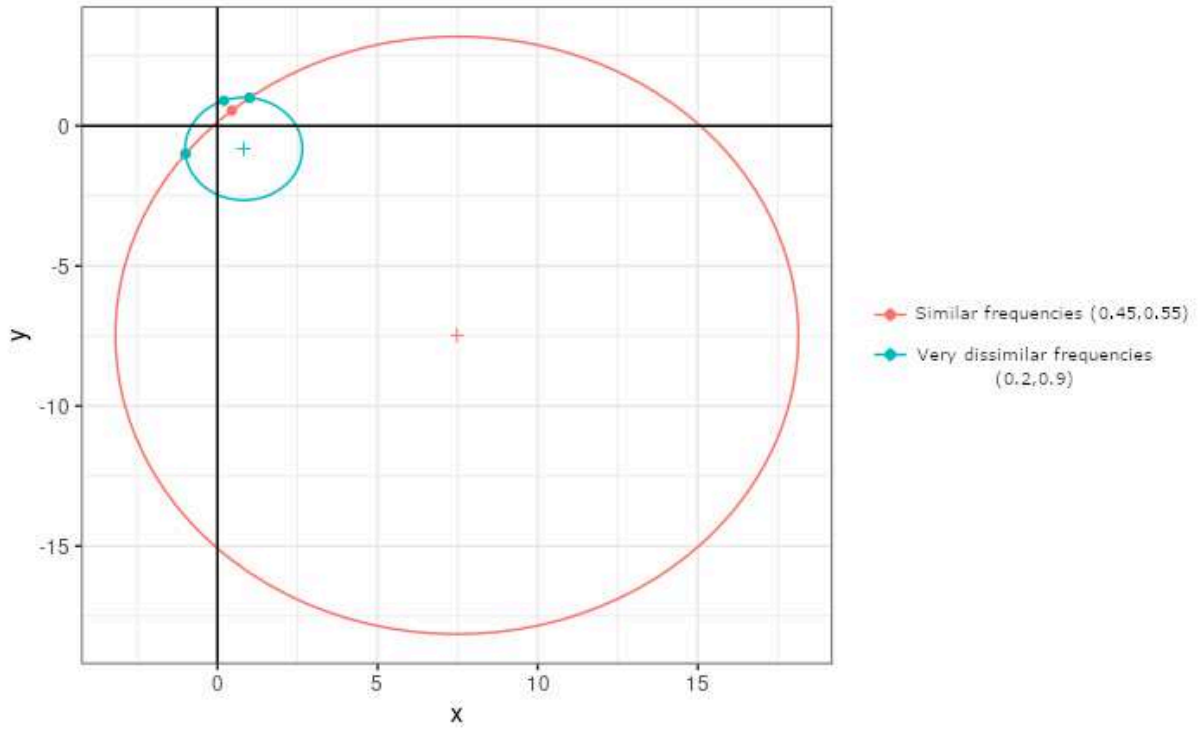

Figure 34 : Représentation d'un cercle correspondant à une faible différence entre deux fréquences d'activation (en rouge) et d'un autre cercle correspondant à une grande différence entre deux fréquences d'activation (en bleu). *En ayant fixé deux des trois points nécessaires pour déterminer le rayon et le centre d'un cercle à des coordonnées données, seul le troisième point influera sur le rayon et le centre du cercle ce qui permet d'utiliser ces deux valeurs pour comparer des différences de fréquences d'activation.*

Connaissant ces propriétés, nous avons donc développé une métrique que l'on peut nommer « center_of_circle1.2_log» (CoC). Cette métrique considère deux points A et B, de coordonnées (-1.2,-1.2) et (1.2,1.2) respectivement et un point C de coordonnées (f_{ctrl}, f_{treat}). Cette métrique se base sur l'équation du cercle pour calculer le centre et le diamètre du cercle à partir de ces paramètres et transforme le diamètre calculé par un logarithme népérien. De la même manière que pour le R2, nous avons simulé les valeurs de « centre du cercle » pour les fréquences comprises entre 0 et 1 afin de visualiser le comportement de la métrique dans l'intervalle des valeurs possibles (Fig 35). Les ellipses, en bleu sur la figure 35 représentent l'évolution de la métrique en fonction de la fréquence d'activation dans la condition contrôle et de la fréquence d'activation dans la condition traitée. Par exemple, une réaction ayant une fréquence d'activation égale à 0 en condition contrôle et égale à 0,25 en condition traitée, serait considérée comme DAR d'après le CoC (CoC = 1.74, donc inférieur au seuil de 1,75 en dessous duquel une réaction est considérée comme DAR, défini à partir des figures 33 et 35) alors qu'elle ne serait pas considérée comme DAR d'après le R2 (R2 = 0,0625, donc inférieur au seuil de 0.2 à partir duquel une réaction est considérée comme DAR, défini à partir des figures 33 et 35). Cette nouvelle métrique répond donc de manière satisfaisante en étant capable d'identifier comme différentiellement activées des réactions inactives dans une condition et partiellement active dans l'autre.

A noter que tous les points situés sur le tracé d'une ellipse (Fig 35) ont la même valeur de CoC ce qui permet de visualiser l'évolution de la métrique selon les combinaisons de fréquence d'activation dans la condition traitée et la condition contrôle.

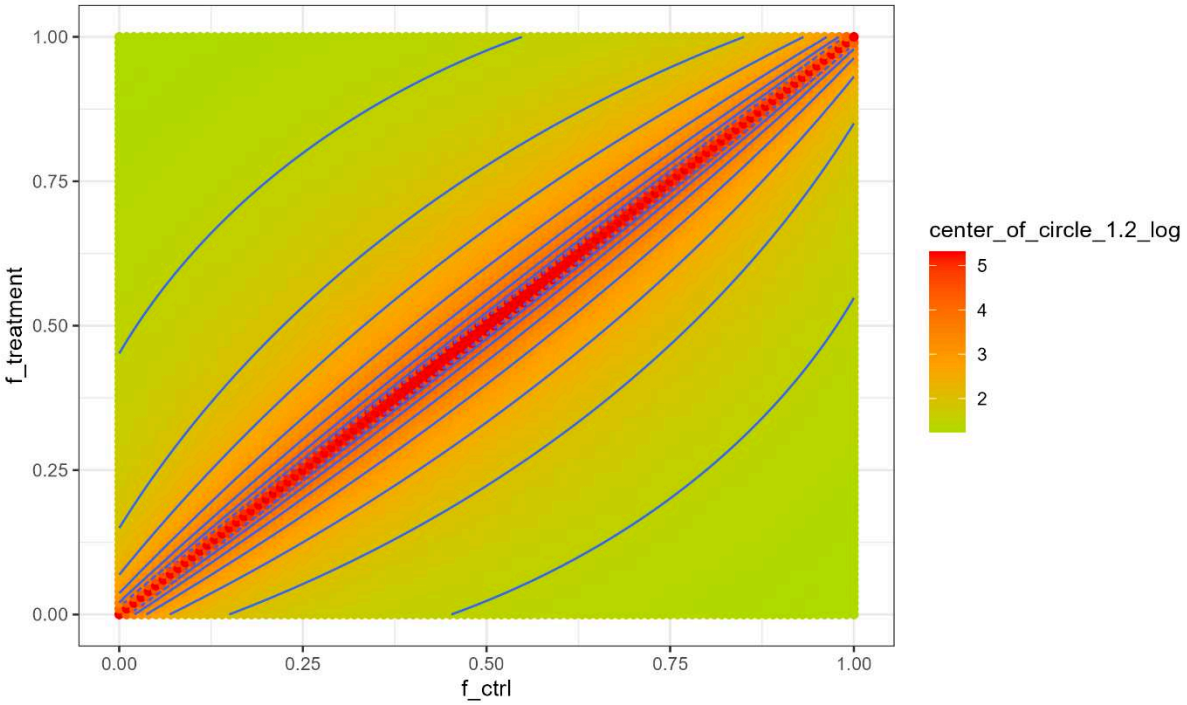

Figure 35: Simulation des valeurs de logarithme du centre du cercle possibles pour des fréquences comprises entre 0 et 1.

2.1.3. Choix de la métrique pour l'identification de réactions différentiellement activées

Afin de choisir la métrique la plus adaptée, nous avons besoin de pouvoir comparer ces deux métriques. Cependant, choisir un seuil commun pour ces deux métriques s'avère compliqué puisque ces deux métriques évoluent de manière opposée. D'une part, le R2 atteint une limite en 1, signifiant une différence totale entre les deux fréquences d'activation et d'autre part le CoC tend vers l'infini signifiant une similarité totale entre les deux fréquences d'activation. En d'autres termes, plus la valeur de R2 est élevée plus la différence entre les deux fréquences d'activation est importante et plus la valeur de CoC est élevée, moins la différence entre les deux fréquences d'activation est importante.

Il est possible de visualiser cet effet sur la figure 36, où nous avons sélectionné un seuil pour chaque métrique, sur la base de l'analyse des Figure 33 et 35. Pour le R2, le seuil au-dessus duquel une réaction est considérée comme différentiellement activée est égal à 0,2, ce qui correspond à une différence de fréquence d'activation de 45%. Pour le CoC, le seuil au-dessous duquel une réaction est considérée comme différentiellement activée est égal à 1,75, ce qui correspond à une différence de fréquence d'activation pouvant aller de 25% à 40% selon la valeur des fréquences (la métrique étant plus permissive pour les cas extrêmes d'activation/inactivation complète). Cependant la définition de ces seuils est assez subjective et rend difficile la comparaison de la performance de l'une ou l'autre des métriques sur la base du contenu des listes de DARs.

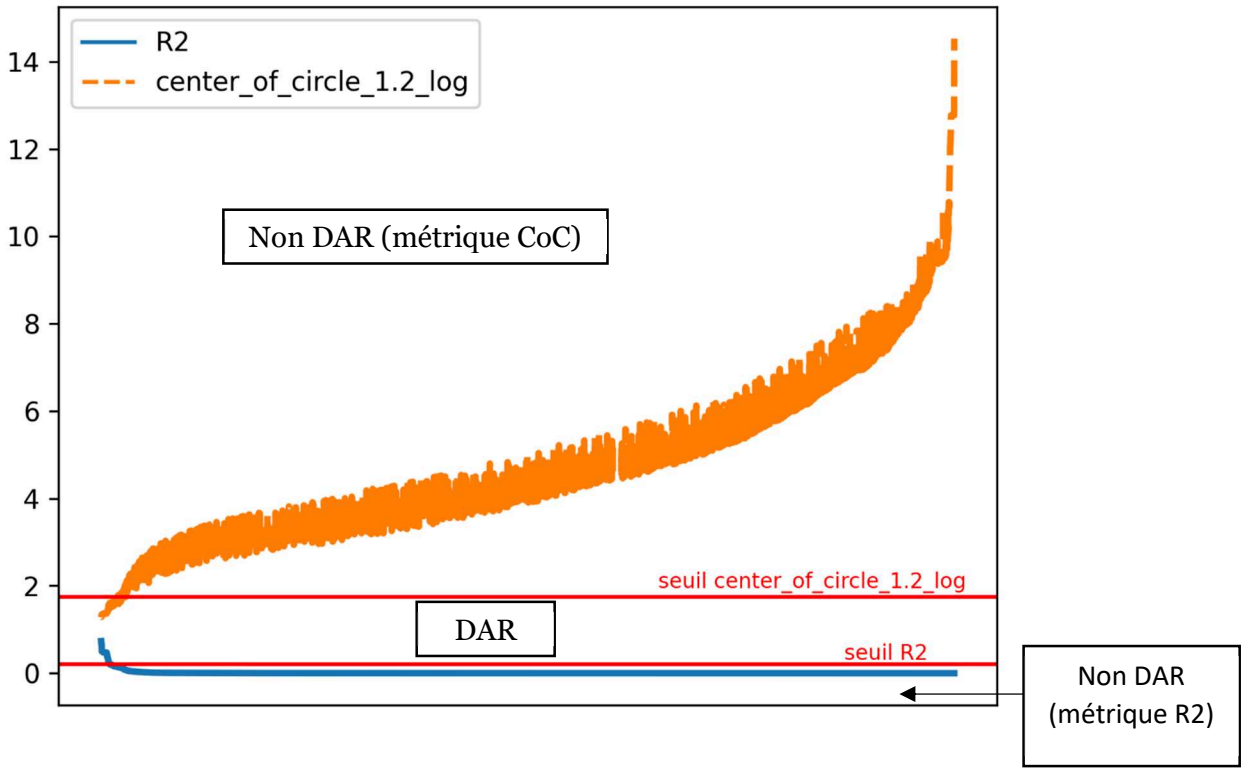

Figure 36 : Distribution des valeurs calculées pour le R2 et le center_of_circle_1.2_log. *L'axe des abscisses correspond aux réactions du modèle et l'axe des ordonnées correspond à la valeur de R2 ou de CoC calculée pour chacune des réactions entre la condition contrôle et la condition traitée par de l'amiodarone à 7µM pendant 24h. La courbe en orange correspond aux valeurs de la métrique « center_of_circle_1.2_log » transformée par un logarithme népérien. Plus cette valeur est basse, plus la différence entre les fréquences d'activation de la réaction correspondante dans les deux conditions testées est importante. La courbe en bleu correspond aux valeurs de la métrique « R2 ». Plus la valeur de R2 est élevée, plus la différence entre les fréquences d'activation de la réaction correspondante dans les deux conditions testées est importante. La zone « Non DAR (métrique R2) » correspond aux réactions dont la valeur de R2 est sous le seuil R2 (R2<0,2), la zone « DAR » correspond aux réactions dont la valeur de R2 est au-dessus du seuil R2 et/ou en dessous du seuil CoC (CoC<1,75). La zone « Non DAR (métrique CoC) » correspond aux réactions dont la valeur est supérieure au seuil de CoC.*

L'effet du choix du seuil sur les listes de DARs est visible sur la Figure 37. Toutes les DARs identifiées par le R2 au seuil de 0,3 sont identifiées par les autres combinaisons de métrique/seuil. Toutes les DARs identifiées par le R2 au seuil de 0,2 sont identifiées par le CoC au seuil de 1,75. Avec ce seuil, la métrique du CoC est également capable d'identifier 76 réactions supplémentaires en tant que DAR. Cependant en abaissant le seuil du CoC à 1,5, cette métrique devient alors plus conservative que le R2 au seuil de 0,2 en identifiant 10 DARs de moins. A partir de l'analyse des figures 36 et 37, il apparait qu'à la fois le choix du seuil mais également le choix de la métrique impacte le contenu ainsi que la taille des listes de DARs.

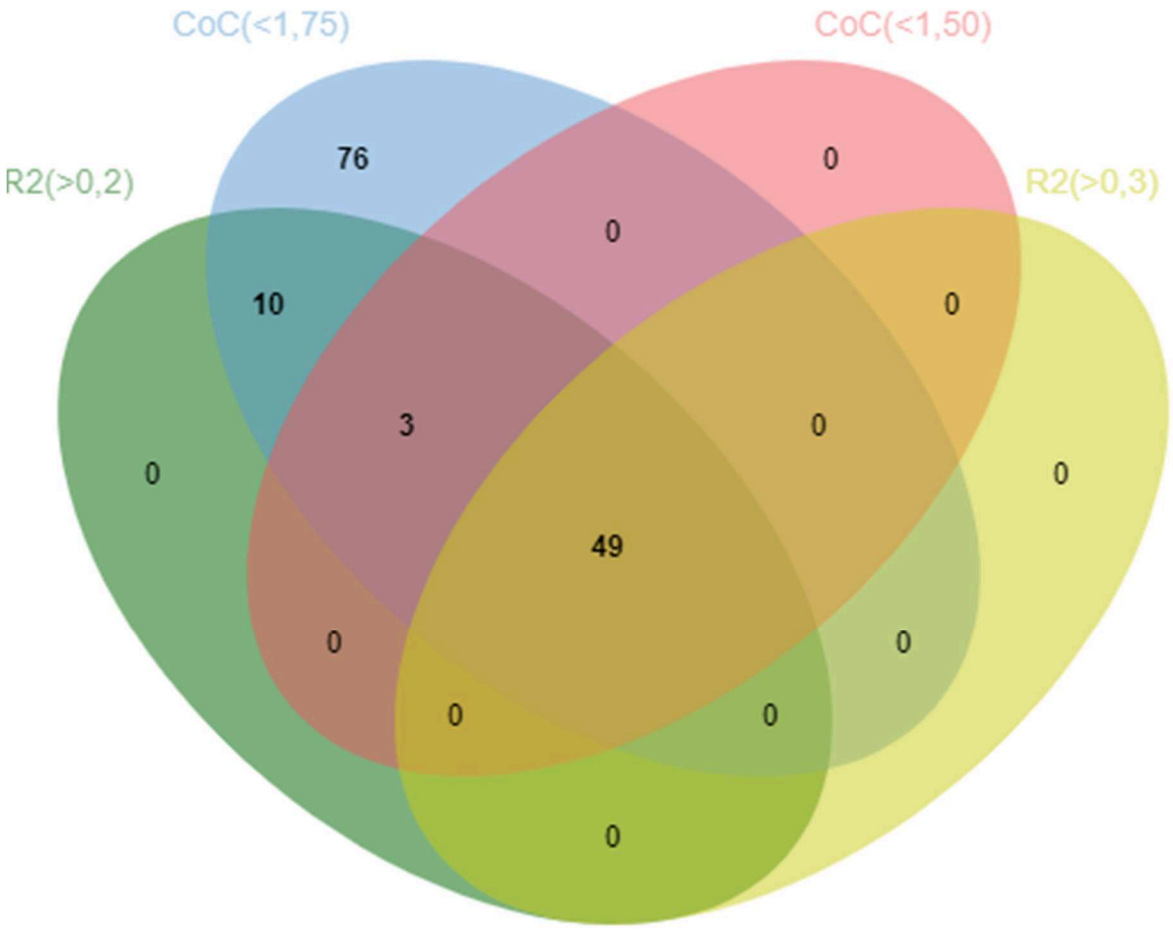

Figure 37 : Diagramme de Venn des DARs identifiées avec la métrique du R2 ou du Center of Circle (CoC). *L'ellipse verte correspond aux DARs identifiées avec la métrique du R2 au seuil de 0,2. L'ellipse bleue correspond aux DARs identifiées avec la métrique du Center of Circle 1.2 Log au seuil de 1,75. L'ellipse rouge correspond aux DARs identifiées avec la métrique du Center of Circle 1.2 Log au seuil de 1,50. L'ellipse jaune correspond aux DARs identifiées avec la métrique du R2 au seuil de 0,3. L'intersection des quatre cercles représente les DARs identifiées par le R2 (>0,2), CoC (<1,75), CoC (<1,50) et le R2 (>0,3).*

Notre choix s'est finalement porté sur la métrique du R2 car il s'agit d'une part de la métrique la plus conservative mais également de la métrique la plus « simple » d'un point de vue méthodologique et donc plus facile à justifier et faire adopter par les futurs utilisateurs.
En effet, cette simplicité facilite la compréhension de la métrique par d'autres utilisateurs tout en facilitant notre compréhension de son comportement dans un maximum de cas d'usage possible. Le CoC est une approche intéressante car elle permet de considérer comme DAR des réactions totalement inactives dans une condition et assez faiblement actives dans une autre. Cependant, au vu de sa plus grande complexité et de sa plus grande originalité, il serait nécessaire d'étudier en détail sa robustesse dans différentes conditions. En choisissant le R2 au détriment du CoC nous avons donc choisi une métrique plus conservative afin d'augmenter la confiance concernant l'impact métabolique prédit à l'issue de la modélisation et de l'identification des DARs, mais avec le risque de ne pas mettre en évidence certaines modulations.

2.1.4. Identification du bruit basal

Certaines réactions du réseau métabolique à l'échelle du génome ne peuvent être contraintes par les données transcriptomiques condition-spécifiques. En effet, une réaction pour laquelle aucune GPR n'a été déterminée (*e.g.* une réaction de transport passif) ou une réaction ayant une GPR mais pour laquelle aucune information transcriptomique n'a été intégrée, ne sera pas contrainte par les données omiques. Cette absence de contrainte peut donc engendrer de l'incertitude (ou "bruit') dans les prédictions active/inactive pour chacune des réactions et donc par extension de l'incertitude sur la fréquence d'activation calculée à partir des solutions alternatives pour ces réactions. Comme cela a été décrit dans le premier chapitre, le solveur cherche une solution optimale maximisant d'une part le nombre de réactions prédites actives et appartenant à la liste des réactions actives selon les données transcriptomiques et d'autre part le nombre de réactions prédites inactives et appartenant à la liste des réactions inactives selon les données transcriptomiques. Cependant, une réaction n'étant pas contrainte par les données transcriptomiques n'appartiendra à aucune des deux listes et donc sa prédiction active/inactive n'impactera pas l'optimalité de la solution trouvée. A noter que ces réactions non contraintes par les données transcriptomiques peuvent tout de même être contraintes de manière indirecte en étant liées dans le réseau à des réactions contraintes par les données transcriptomiques. Néanmoins, l'incertitude liée à ce manque de contrainte directe pourrait engendrer l'activation/inactivation aléatoire de certaines réactions. Afin de prendre en compte cette incertitude, nous avons développé une approche permettant d'estimer le bruit médian associé à chaque réaction dans le réseau. Le bruit médian d'une réaction fait référence à la médiane des valeurs de R2 calculées entre toutes les paires de conditions contrôle pour un solvant donné. Cela représente la valeur correspondant au point milieu des valeurs de R2 calculées entre toutes les paires de contrôle pour un solvant donné (50% des valeurs de R2 sont inférieure à la médiane et 50% sont supérieures) et permet d'identifier la variation de R2 au sein des contrôles en limitant l'impact des valeurs extrêmes (liées à des « outliers » par exemple). Plusieurs solvants (DMSO et milieu de culture) ont été utilisés par les auteurs de la base de données Open TG-GATEs. Le choix du solvant a été réalisé en fonction de la solubilité du composé testé dans l'un ou l'autre des solvants. Etant donné que pour chaque condition traitée (molécule, temps d'exposition et dose), une condition contrôle a été réalisée avec le même solvant, le choix du solvant impacte également les données « contrôle ».

Cette différence de solvant et notamment l'action du DMSO [202], risque d'induire des perturbations métaboliques qui ne seraient pas du bruit si les conditions « contrôle DMSO » étaient comparées aux conditions « contrôle medium ». Le temps d'exposition doit également être pris en compte pour le calcul du bruit basal car nous avons identifié des différences entre

les échantillons contrôles de différents temps d'exposition lors de notre exploration de la base de données Open TG-GATEs.

Le bruit médian pour chaque réaction a donc été calculé entre les conditions contrôles correspondant à un temps et un solvant donnés, et en combinant les solutions alternatives simulées pour chacune des conditions. Sur le tableau 12, il est intéressant de noter que le bruit moyen est plutôt faible avec des valeurs de 0,008 pour le groupe « DMSO » et 0,018 pour le groupe « Medium » bien que certaines réactions atteignent un bruit maximal très élevé de 0,77 pour le groupe « DMSO » et 0,83 pour le groupe « Medium ». Ce bruit maximal peut-être dû à la présence d'un effet que nous n'avons pas identifié entre les contrôles. Cependant, au vu du bruit moyen cet effet non déterminé, s'il existe, serait très marginal.

Groupe de contrôles	Nombre de paires testées	Bruit minimal (R2)	Bruit maximal (R2)	Bruit moyen (R2)
Contrôle_DMSO_24hr	9	0	0,77	0,008
Contrôle_Medium_24hr	3	0	0,83	0,018

Tableau 12. Tableau résumant le bruit minimal, maximal et moyen sur les valeurs de bruit calculées pour l'ensemble des réactions, pour un solvant et un temps d'exposition donnés. *Le nombre de paires testées correspond aux paires de contrôles dont les fréquences d'activation ont été comparées en utilisant le R2 à un solvant donné et à un temps d'exposition donné. Les valeurs de bruit minimal, maximal et moyen correspondent respectivement à la valeur de R2 médiane la plus faible, la plus forte et la moyenne des valeurs, calculées pour un groupe de contrôles donné.*

Ces valeurs de bruit basal calculées pour toutes les réactions sont ensuite utilisées pour filtrer les réactions prédites comme différentiellement activées. Les réactions différentiellement activées dont la valeur de R2 n'est pas strictement supérieure à deux fois la valeur de bruit calculé pour cette réaction sont exclues de la liste des réactions.

L'une des principales limites de l'approche de filtration basée sur le calcul du bruit basal propre à chaque réaction du GSMN est le nombre d'échantillons disponibles. En effet, pour la condition 24h medium, le bruit basal de chaque réaction a été calculé entre 3 paires de conditions contrôle Pour la condition 24h DSMO, le bruit basal de chaque réaction a été calculé entre 9 paires de conditions contrôle. Ce faible nombre de paires de conditions rend le calcul du bruit basal sensible aux valeurs extrêmes et la robustesse de l'approche pourrait être améliorée en augmentant le nombre de conditions contrôle. Bien que la valeur du bruit basal soit à mettre en perspective avec la valeur de R2 calculée entre les conditions traitées et les conditions contrôles, il aurait été intéressant d'étudier en détail les réactions présentant un bruit basal élevé (*e.g.* R2 > 0,1) afin de comprendre pourquoi l'activité de ces réactions fluctue entre les contrôles. En réalisant une analyse rapide des 186 réactions ayant un bruit basal supérieur à 0,1 pour la condition contrôle, 24h DMSO, nous avons remarqué que les voies de synthèse et de dégradation des acides gras sont plus concernées par ce phénomène de bruit basal avec respectivement 29 et 22 réactions ayant un bruit basal important (>0,1) appartenant

à ces voies. Cela pourrait être dû à un manque de contraintes transcriptomiques pour les gènes de ces voies ou à des reliquats de correction d'effet lot d'HPH (*i.e.* un effet lot encore visible sur l'intensité d'expression de certains gènes). Il serait donc intéressant d'une part d'augmenter le nombre de conditions contrôle et d'autre part de calculer le bruit basal sur d'autres données transcriptomiques présentant moins d'effets lots.

2.2. Utilisation de la stratégie développée pour l'identification de réactions métaboliques impactées pour 8 molécules hépatotoxiques

2.2.1. Identification des réactions métaboliques impactées pour les 8 molécules sélectionnées

L'intégration des données transcriptomiques puis la modélisation sous-contraintes avec énumération par notre version adaptée de DEXOM a été réalisée pour les conditions correspondant aux 8 molécules hépatotoxiques sélectionnées, à la plus forte dose disponible dans la base de données Open TG-GATEs ainsi qu'au temps d'exposition le plus long pour les données générées sur HPH (24 heures). Cela correspond à modéliser environ 10 000 réseaux condition-spécifiques par échantillon, en sachant que pour chaque molécule testée 2 échantillons ont été traités et 2 contrôles ont été réalisés. Les solutions calculées pour des échantillons correspondant à des réplicas biologiques (2 par conditions) ont été combinées, ce qui revient à environ 20 000 solutions alternatives par condition modélisée, qui sont représentatives de l'état du métabolisme cellulaire de HPH dans la condition d'exposition.

Comme précisé précédemment, les distributions de flux prédites par DEXOM sont binarisées (cf. section 1.3.2) et sont stockées sous la forme de vecteurs binaires (Tableau 13) indiquant pour chaque réaction si elle est active ou inactive dans la solution.

Solution DEXOM	**Réaction 1**	**Réaction 2**	**Réaction 3**	**…**	**Réaction N-1**	**Réaction N**
Solution 1	0	0	0	…	1	1
….	…	…	…	…	…	…
Solution N	1	1	0	…	0	1

Tableau 13. Exemple d'un tableau de solutions énumérées par DEXOM pour une condition d'exposition. *Ces solutions alternatives ont été énumérées avec une version adaptée de DEXOM et stockées sous forme de vecteurs binaires. Une valeur de 1 indique une réaction active alors qu'une valeur de 0 indique une réaction inactive. N≈20 000 solutions*

Il est intéressant de remarquer que les nombres minimal, maximal et moyen de réactions prédites comme actives par notre adaptation de DEXOM sont du même ordre de grandeur

(Tableau 14) pour toutes les conditions testées (*i.e.* les 8 molécules testées ainsi que leur contrôles). Le nombre minimal de réactions prédites actives varie entre 3423 (DMSO) et 3639 (tétracycline, 25µM, 24h), le nombre maximal de réactions prédites actives varie entre 4396 (acide valproïque, 5000µM, 24h) et 4645 (indométacine 200µM, 24h) et enfin le nombre moyen de réactions varie entre 4070 (acide valproïque 5000µM, 24h) et 4570 (amiodarone 7µM, 24h). Cette première observation suggère que la taille des réseaux condition-spécifique n'est pas forcément indicative de potentiels effets hépatotoxiques étant donnée l'absence de différence d'ordre de grandeur entre les réseaux condition-spécifiques des conditions contrôle et les réseaux condition-spécifiques des conditions traitées.

	ethanol (10000µM, 24h)	valproic acid (5000µM, 24h)	indomethacin (200µM, 24h)	amiodarone (7µM, 24h)	allopurinol (140µM, 24h)	rifampicin (70µM, 24h)	sulindac (3000µM, 24h)	tetracycline (25µM, 24h)	medium	DMSO
Nombre minimal de reactions actives	3475	3522	3498	3623	3534	3535	3499	3639	3578	3423
Nombre maximal de reactions actives	4615	4396	4645	4570	4563	4620	4479	4580	4626	4610
Nombre moyen de reactions actives	4260	4070	4266	4570	4176	4284	4122	4209	4216	4192

Tableau 14. Nombre de réactions prédites comme actives pour chaque molécule testée et les solvants utilisés*. Ensemble de réseaux condition-spécifiques énumérés par une version adaptée de DEXOM. Les solutions correspondant aux réplicas d'une même condition ont été combinées avant de calculer le nombre minimal, maximal et moyen de réactions actives par condition.*

De fait, si la différence entre les ensembles de réseaux condition-spécifiques modélisés pour chacune des conditions testées n'est pas représentée par la taille de ces réseaux, nous pouvons supposer que ces différences résident dans leur composition.

Afin d'identifier l'impact métabolique de chacune des 8 molécules sélectionnées, nous avons calculé les DARs pour chaque molécule en comparant les fréquences d'activation de chaque réaction dans la condition « exposition pendant 24 heures, à la plus forte dose » aux fréquences d'activation dans la condition contrôle correspondante (même lot de cellule, même temps d'exposition, même solvant, …). Le nombre de DARs identifiées était au minimum de 57 DARs pour les HPH exposés à l'indométacine (200µM, 24h) et au maximum de 477 DARs pour les HPH exposés à l'acide valproïque (5000µM, 24h) (Tableau 15). Les DARs identifiées pour l'acide valproïque, l'indométacine, l'amiodarone et l'allopurinol n'ont été que très légèrement affectées par la filtration des réactions identifiées comme bruitées par l'approche de calcul du

bruit basal (cf. section 2.1.5) avec respectivement 12.6%, 5.3%, 6.7%, et 11.4% de DARs filtrées (*i.e.* avec un R2 inférieur à deux fois le bruit basal). A l'inverse, les DARs identifiées pour la tétracycline et l'éthanol ont été plus fortement impactées avec respectivement 43,4% et 62,8% des DARs éliminées par l'étape de filtration (Tableau 15).

	ethanol (10000µM, 24h)	valproic acid (5000µM, 24h)	indomethacin (200µM, 24h)	amiodarone (7µM, 24h)	allopurinol (140µM, 24h)	rifampicin (70µM, 24h)	sulindac (3000µM, 24h)	tetracycline (25µM, 24h)
Nombre de DARs	94	477	57	60	88	121	242	99
Nombre de DARs après filtration	35	417	54	56	78	98	181	56
% de reactions retirees lors de la filtration	62.8	12.6	5.3	6.7	11.4	19	25.2	43.4

Tableau 15. Nombre de DARs identifiées pour chaque condition. *Les DARs identifiées pour des réseaux condition spécifique de HPH après 24hr d'exposition à la dose la plus forte n'induisant pas plus de 20% de cytotoxicité pour les 8 molécules sélectionnées.*

Afin d'identifier les DARs spécifiques à une condition (i.e. qui ne sont pas prédites comme DAR pour aucune des autres molécules étudiées) nous avons calculé le ratio de spécificité des DAR pour chacune des conditions.

$$\boldsymbol{DARspec_x = \frac{nSpecificDARs_x}{nTotalDARs_x}} \quad (1)$$

Ce ratio (1) correspond au nombre de DARs identifiées uniquement dans la condition étudiée (x) divisé par le nombre total de DARs identifiées pour cette même condition. L'acide valproïque est le composé qui partage le moins de DARs avec les autres molécules car elle a un ratio de spécificité de 91,1% alors que l'indométacine est le composé qui partage le plus de DARs avec les autres molécules car elle a un ratio de spécificité de 22,2% (Tableau 16). Il est important de noter que cette métrique est à interpréter avec précaution car il suffit que deux molécules parmi celles sélectionnées aient des mécanismes d'actions métaboliques proches et donc un nombre de DARs en commun important pour qu'elles aient toutes les deux des scores de spécificité faibles, suggérant que ces molécules partagent de nombreuses DARs avec tous les autres composés, ce qui n'est pas forcément vrai. Afin d'apporter plus de précision lors de la lecture de cette métrique, nous avons calculé le pourcentage moyen du nombre de DARs partagées par une molécule avec toutes les autres molécules. Cette métrique permet notamment de faire la différence entre une molécule qui partage un mécanisme d'action avec un grand nombre de molécule et une molécule qui partage un mécanisme d'action avec une seule autre molécule. Dans le premier cas, le pourcentage moyen du nombre de DARs

partagées avec toutes les autres molécules sera élevé alors que le ratio de spécificité sera faible. Ceci est, dans une certaine mesure, le cas de l'indométacine (Tableau 16) qui a le plus haut pourcentage moyen du nombre de DARs partagées avec toutes les molécules testées mais le plus faible ratio de spécificité. Dans le second cas, le pourcentage moyen du nombre de DARs partagées avec toutes les autres molécules sera faible et le ratio de spécificité faible également. Ce cas de figure n'est pas retrouvé parmi les 8 molécules sélectionnées mais pourrait être retrouvé lorsque l'on compare deux analogues biologiques par exemple.

	Nombre de DARs spécifiques d'une molécule	Ratio de spécificité (%)	Nombre total de DARs	Pourcentage moyen de l'intersection entres les ensembles de DARs
ethanol (10000µM, 24h)	19	54.3	35	9.0
valproic acid (5000µM, 24h)	380	91.1	417	1.6
indomethacin (200µM, 24h)	12	22.2	54	12.4
amiodarone (7µM, 24h)	14	25.0	56	11.2
allopurinol (140µM, 24h)	23	29.5	78	10.6
rifampicin (70µM, 24h)	36	36.7	98	10.9
sulindac (3000µM, 24h)	111	61.3	181	6.8
tetracycline (25µM, 24h)	43	76.8	56	4.1

Tableau 16. Ratios de spécificité et pourcentages moyens du nombre de DARs partagées avec les autres composés étudiés. *Ce tableau contient plusieurs métriques dont l'objectif est de comprendre comment les DARs identifiées sont partagées entre les 8 composés étudiés. Les métriques présentées dans ce tableau sont : le nombre de DARs spécifiques à chaque molécule, le ratio de spécificité, le nombre total de DARs et le nombre moyen de DARs partagées avec les autres composés.*

Le nombre de DARs identifiées ainsi que les métriques de spécificité et de taille d'intersection moyenne sont des métriques descriptives intéressantes car elles permettent de visualiser l'importance et la diversité des perturbations métaboliques entre nos 8 molécules d'intérêt. Elles peuvent notamment orienter les analyses suivantes selon les objectifs de l'étude. Par exemple si l'on recherche des molécules analogues alors on s'intéressera en premier lieu aux molécules ayant un ratio de spécificité faible. D'après les métriques présentées dans le tableau 16, aucune des 8 molécules testées n'est très proche d'une autre molécule testée (faible ratio de spécificité et faible pourcentage moyen de DARs partagées avec l'ensemble des molécules testées). Cependant, il semble que l'amiodarone, l'allopurinol, l'indométacine, et dans une moindre mesure la rifampicine, partagent une partie de leurs mécanismes d'actions car leurs ratios de spécificité sont faibles et le pourcentage moyen de DARs partagées avec l'ensemble des autres molécules sont parmi les plus importants. Cette observation montre cependant la dépendance de la métrique « pourcentage moyen de DARs partagées avec l'ensemble des

autres molécules testées » au nombre ainsi qu'à la diversité des molécules testées. Les métriques présentées dans le tableau 16 sont informatives lors d'une première phase de d'exploration et de comparaison des listes de DARs mais doivent être nécessairement complétées par une analyse mécanistique plus poussée. D'une manière générale, il est important d'être vigilant quant à l'interprétation brute du nombre de DARs identifiées. En effet, comme nous allons le voir dans les prochaines sections, certaines réactions identifiées comme DARs peuvent être des réactions peu informatives (*e.g.* réactions de transport extracellulaire et réactions de modélisation).

2.2.2. Interprétation des DARs par une analyse de sur-représentation

Pour étudier les différences et l'éventuelle complémentarité de la prédiction des DARs avec l'identification de gènes différentiellement exprimés (DEGs), nous avons réalisé une analyse de sur-représentation sur les signatures transcriptomiques (*i.e.* les listes de DEGs) ainsi qu'une analyse de sur-représentation sur les listes de DARs pour les mêmes conditions.

2.2.2.1. Analyse de sur-représentation sur les ensembles de gènes

La taille des signatures transcriptomiques est très variable selon les molécules testées. Comme indiqué précédemment, ne sont considérés que les échantillons exposés à la plus forte dose pendant 24 heures et les contrôles correspondants. La molécule ayant la plus petite signature transcriptomique est l'amiodarone avec seulement 2 DEGs dont aucun gène métabolique (*i.e.* associé à une réaction métabolique) alors que la molécule ayant la plus grande signature transcriptomique est le sulindac avec 6434 DEGs dont 632 sont des gènes métaboliques. Les signatures transcriptomiques calculées pour les 8 molécules sélectionnées représentent un défi pour l'analyse de sur-représentation. En effet, bien que le test exact de Fisher (utilisé pour l'étude de sur-représentation) soit adapté aux échantillons de petite taille ($n<5$) [203], il est recommandé de ne pas réaliser d'analyse de sur-représentation sur des signatures transcriptomiques constituées de seulement quelques gènes. A l'inverse des signatures de très grande taille comme celles de l'acide valproïque et du sulindac ont tendance à être significativement enrichies pour un très grand nombre de voies rendant l'interprétation fonctionnelle difficile.

Composé	Dose (µM)	Temps (heures)	Nombre de DEGs	Nombre de DEGs métaboliques
Ethanol	10000	24	**1483**	**163**
Valproic Acid	5000	24	**5710**	**611**
Indomethacin	200	24	**890**	**180**
Amiodarone	7	24	**2**	**0**
Allopurinol	140	24	**1271**	**129**
Rifampicin	70	24	**810**	**164**
Sulindac	3000	24	**6434**	**632**
Tetracycline	25	24	**503**	**58**

Tableau 17. Tailles des signatures transcriptomiques (listes de DEGs) pour chaque molécule exposée à la plus forte dose pendant 24hr. *Ce tableau contient la taille des signatures transcriptomiques obtenues pour chacune des 8 molécules testées à la plus forte dose disponible dans la base de données Open TG-GATEs pendant 24hr sur des hépatocytes primaires humains. Ne sont considérés comme différentiellement exprimés que les gènes ayant un log2(abs(FC)) > 0,26 et une p-valeur corrigée (FDR) inférieure à 0,05.*

Nous avons réalisé une analyse de sur-représentation pour les signatures transcriptomiques des 8 molécules sélectionnées (Tableau 17) sur la base Reactome 2022 avec le package R « ReactomePA ». Les figures montrant les 50 meilleures (classées par p-valeur corrigée) voies enrichies pour les 8 molécules d'intérêt sont disponibles en Annexe (Figure 55 à 61). Il est intéressant de remarquer que les gènes différentiellement exprimés après exposition à l'éthanol n'ont été enrichies dans aucune voie de la base Reactome. Etant donné la taille de la signature transcriptomique de l'éthanol, cela semble assez surprenant. Cela pourrait être dû à une répartition des 1483 DEGs de l'éthanol sur l'ensemble des voies de Reactome et donc sur-représentés dans aucune voie.

A noter qu'utiliser la p-valeur pour classer les voies significativement sur-représentées est une approche couramment utilisée mais statistiquement discutable. La p-valeur n'étant pas une mesure de la force de la significativité, utiliser sa valeur pour classer des résultats pourrait mener à une interprétation erronée de l'importance de la voie identifiée dans le mécanisme d'action de la molécule [204].

Néanmoins, les gènes métaboliques ne représentent qu'une sous-partie (entre 10 et 20% environ) des DEGs identifiés pour chacune des molécules et un grand nombre de « voies Reactome » significativement enrichies ne sont pas des voies métaboliques, ce qui suggère que les mécanismes identifiés par cette analyse basée uniquement sur les données transcriptomiques pourrait être complémentaire des analyses que nous allons détailler dans les prochaines parties et dont l'objectif est d'élucider le mécanisme d'action métabolique des composés modélisés.

2.2.2.2. Analyse de sur-représentation sur les voies métaboliques

A l'instar des analyses de sur-représentation pour les données transcriptomiques, il est possible de réaliser une analyse de sur-représentation des DARs dans les voies métaboliques définies dans Recon2.2. L'analyse de sur-représentation dans les voies métaboliques souffre des mêmes limites que celles évoquées pour l'analyse de sur-représentation sur les ensembles de gènes (GO, Reactome, …), c'est-à-dire qu'elle est sensible à la taille des signatures et à la taille des voies métaboliques [46,47]. Les GSMNs tels que Recon2.2 contiennent un ensemble de réactions peu informatives d'un point de vue mécanistique, mais qui peuvent néanmoins être identifiées comme DARs : il s'agit par exemple de réactions d'échange avec le milieu extérieur. Nous avons donc choisi de les retirer lors de l'analyse de sur-représentation des DARs afin de focaliser l'analyse sur les réactions propres au métabolisme cellulaire. Les DARs identifiées pour ces 8 molécules sont sur-représentées dans un total de 30 voies métaboliques. La répartition du nombre de voies enrichies par molécule est assez inégale avec une seule voie sur-représentée pour l'indométacine et 12 voies sur-représentées pour l'acide valproïque. La voie de synthèse des acides gras ainsi que la voie de synthèse des pyrimidines sont les voies qui sont les plus souvent sur-représentées (4 molécules sur 8), viennent ensuite la voie du métabolisme des stéroïdes ainsi que la voie du transport nucléaire (3 molécules sur 8). 4 voies (métabolisme des sphingolipides, métabolisme de l'alanine et de l'aspartate, métabolisme du NAD et métabolisme de la Thiamine) sont sur-représentées pour deux molécules et 21 voies sont uniquement sur-représentées par l'une des 8 molécules testées. Il est intéressant de noter que les voies citées ci-dessus sont assez peu présentes parmi les voies Reactome identifiées lors de l'analyse de sur-représentation sur les listes de DEG. La voie du métabolisme des stéroïdes est la voie la plus fréquemment identifiée (4 molécules sur 8) dans les analyses d'enrichissement sur les DEGs. La voie de synthèse des acides gras (2 molécules sur 8) ainsi que la voie de synthèse des pyrimidines (3 molécules sur 8) sont également identifiées lors de ces analyses. Cependant, la majorité des voies métaboliques significativement enrichies ne sont pas identifiées par les analyses de sur-représentation des DEGs sur les voies Reactome. L'inverse est également vrai, c'est-à-dire que la majorité des voies Reactome significativement enrichies ne sont pas identifiées (car pas liées au métabolisme) lors des analyses de sur-représentation des DARs dans les voies métaboliques. Ces observations suggèrent d'une part une certaine complémentarité entre l'interprétation directe des données transcriptomiques à partir d'analyse de sur-représentation dans des ensembles de gènes tels que Reactome et l'utilisation d'une stratégie de modélisation sous-contraintes et d'identification de DARs comme présentée au cours de ce chapitre. D'autre part cela montre la difficulté que représente la comparaison d'analyses de sur-représentation réalisées sur des ensembles de gènes/réactions différents. En effet, nous avons comparé ces deux analyses en utilisant les

noms de voies métaboliques qui sont généralement indicatif de la fonction associée à l'ensemble de gène/réactions mais cette comparaison est assez subjective et sensible à la granularité des voies et ensembles de gènes qui peut varier entre les différentes bases disponibles [205].

En analysant les résultats de l'analyse de sur-représentation des DARs sur les voies métaboliques (Fig 38), nous pouvons constater que les perturbations du métabolisme cellulaire sont majoritairement spécifiques du composé étudié. En effet, plus des 2/3 des voies significativement perturbées ne le sont que par un seul des 8 composés testés. Il est intéressant de remarquer que la perturbation généralisée du métabolisme cellulaire identifiée par notre stratégie de modélisation pour l'acide valproïque pourrait être due à l'effet inhibiteur de l'histone de-acetylase [206] décrit dans la littérature pour l'acide valproïque. L'histone de-acetylase est une enzyme jouant un rôle majeur dans la régulation de l'expression génique et dont la perturbation peut donc avoir des effets généralisés sur le fonctionnement cellulaire [207] et donc perturber un grand nombre de voies métaboliques comme l'a révélé l'analyse d'enrichissement (Fig 38). Le métabolisme des acides gras est significativement perturbé par 4 des 8 molécules testées, ce qui a été décrit dans la littérature pour l'amiodarone [208,209] mais pas explicitement décrit pour les 3 autres molécules. La voie d'oxydation des acides gras est significativement enrichie pour une seule molécule, l'indométacine. Etant donné qu'il s'agit d'une voie fortement impliquée dans les mécanismes hépatotoxiques [210], nous aurions pu nous attendre à ce que cette voie soit significativement plus enrichie pour plus de conditions. Cela peut être dû à plusieurs cas de figures : utilisation de doses trop faibles pour déclencher des effets sur la dégradation des acides gras, DARs associées à la dégradation des acides gras qui ont tendance à être impactées par la filtration sur le bruit basal et également la taille importante de la voie de la beta-oxydation dans Recon2.2 qui peut pénaliser cette voie dans l'analyse de sur-représentation [46].

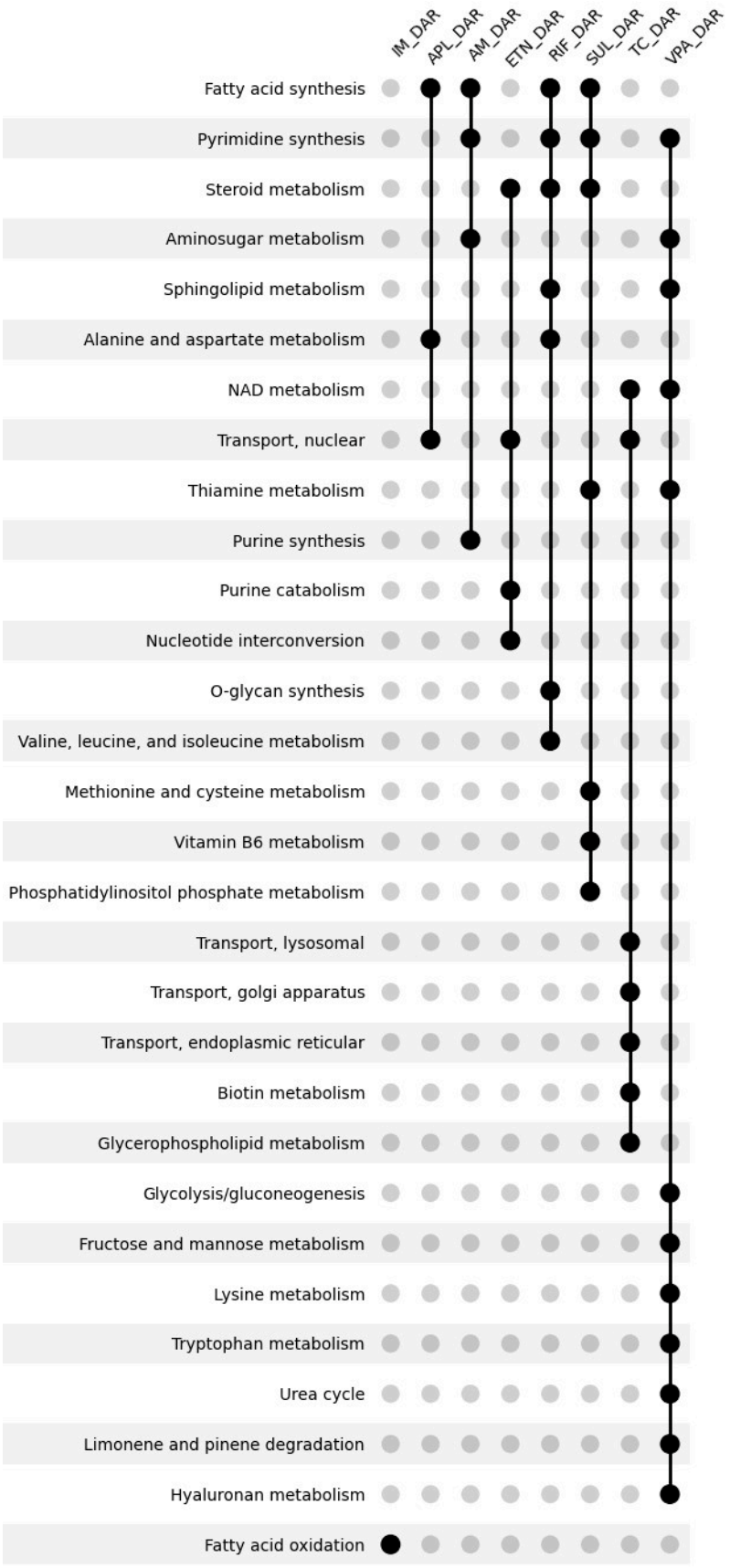

Figure 38 : Enrichissement fonctionnel sur les voies métaboliques de Recon2.2 pour les listes de DARs filtrées (bruit basal et réactions d'échange) des 8 molécules sélectionnées avec MetExplore. *L'étude de sur-représentation a été réalisée avec un test exact de Fisher, les p-valeurs ont été corrigées par une correction de Benjamini-Hochberg.*

2.2.2.3. Limites des analyses de sur-représentation

Comme nous avons pu l'évoquer au cours des précédentes sections, l'analyse de sur-représentation est sensible à la taille des signatures mais également à la taille des voies métaboliques (ou des ensembles de gènes). Cependant, les processus métaboliques ne sont pas strictement isolés les uns des autres comme le laisse penser ce découpage en voies métaboliques. Au contraire, les processus métaboliques interagissent les uns avec les autres et il peut être difficile de trouver le « découpage » adéquat pour séparer ces processus en voies métaboliques bien délimitées [46,47].

La voie de synthèse des acides gras est une des voies les plus sur-représentées parmi les 8 molécules testées. Cette voie est significativement sur-représentée pour l'allopurinol, l'amiodarone, la rifampicine et le sulindac, nous avons donc cherché à savoir si ces 4 molécules impactaient la voie de synthèse des acides gras de manière identique ou s'il pouvait exister différents mécanismes au sein de cette voie. Pour cela, nous avons représenté la voie de synthèse des acides gras décrite dans Recon2.2 grâce à MetExploreViz [70] afin de comparer visuellement les réactions perturbées par chacune des 4 molécules ainsi que les interactions entre ces réactions perturbées. L'amiodarone et l'allopurinol perturbent les mêmes réactions (Fig 39A et Fig 39C) de la voie de synthèse des acides gras, ce qui suggère qu'elles partagent un mécanisme d'action métabolique similaire concernant cette voie. En revanche, la rifampicine et le sulindac n'impactent pas les mêmes réactions métaboliques que l'allopurinol et l'amiodarone mais perturbent plutôt une autre partie de la voie de synthèse des acides gras (Fig 39B et Fig 39D). Ce constat montre que malgré l'intérêt que présente l'analyse de sur-représentation des voies métaboliques pour la compréhension générale de l'impact métabolique d'un xénobiotique, ce type d'analyse n'est pas suffisant pour étudier plus précisément les mécanismes d'action de ces composés, qui peuvent différer, même si tous ces composés sont hépatotoxiques [211]. En effet, en se basant uniquement sur les voies métaboliques, on pourrait penser que des composés sont de bons analogues biologiques (*i.e.* molécules similaires d'après l'étude de leurs impacts biologiques) alors qu'en réalité ils impactent ces voies de manières différentes. Nous aborderons ce sujet en détail dans le prochain chapitre afin de tenter d'apporter une réponse à ce besoin de précision et de compréhension des liens entre processus métaboliques.

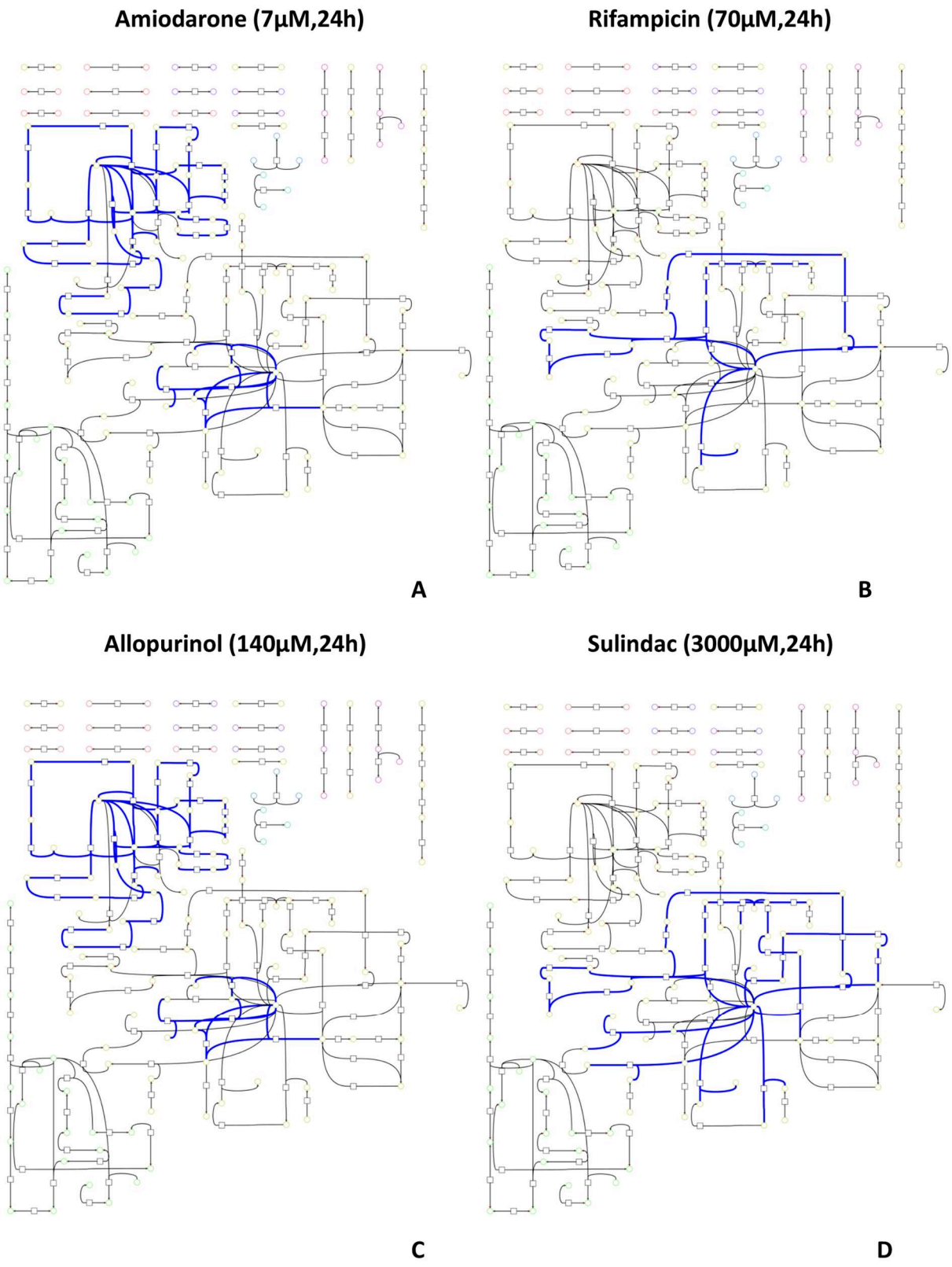

Figure 39 : Comparaison de l'impact métabolique de 4 molécules pour lesquelles la voie de synthèse des acide gras est sur-représentée. *Les réactions différentiellement activées identifiées pour chaque molécule et faisant partie de la voie de synthèse des acides gras sont colorées en bleu. Les nœuds carrés représentent les réactions et les nœuds ronds représentent les métabolites. Les liens entre les DARs coloriés en bleu représentent les parties de la voie de synthèse des acides gras perturbées par la molécule*

2.3. Conclusion et perspectives

Au cours de ce chapitre, nous avons proposé une adaptation de DEXOM permettant d'arriver à un compromis entre coût computationnel et diversité des solutions alternatives. Nous avons également mis l'accent sur l'interprétabilité et la robustesse des résultats issus de l'énumération partielle. En effet, pour traiter ces dizaines de milliers de solutions alternatives énumérées par condition, nous avons recherché une méthode permettant d'identifier des réactions différentiellement activées. Les approches statistiques, envisagées pour identifier des réactions différentiellement activées n'ont finalement pas pu être retenues : en effet, le très grand nombre d'échantillons par condition n'est pas compatible avec le calcul d'une p-valeur qui devient alors systématiquement statistiquement significative (inférieure à 0,05), même lorsque les différences entre les deux conditions sont minimes. Face à cette difficulté, nous avons opté pour le calcul de fréquences d'activation par réaction pour chacune des conditions testées, combiné à une métrique, le R2, permettant de comparer ces fréquences entre 2 conditions. Afin de prendre en compte le manque de contraintes ainsi que la redondance de certaines réactions, nous avons mis au point une méthode de calcul du bruit basal pour chaque réaction. Cette approche nous a notamment permis d'améliorer la robustesse des DARs identifiées en retirant les DARs dont le R2 calculé entre la condition traitée et la condition contrôle n'est pas au moins deux fois supérieur au bruit basal et dont la prédiction pourrait être due à un effet aléatoire. Enfin, nous avons appliqué l'ensemble de ces méthodes à 8 molécules sélectionnées pour leur hépatotoxicité. Nous avons réalisé la modélisation condition-spécifique avec notre version adaptée de DEXOM, calculé les fréquences d'activation, identifié et filtré les DARs afin de réaliser une analyse de sur-représentation sur les voies métaboliques de Recon2.2. Cette application sur 8 molécules nous a notamment permis de montrer l'apport du calcul des DARs pour l'analyse des résultats de modélisation condition-spécifiques avec énumération. L'identification de DARs nous a également permis de réaliser une analyse de sur-représentation sur les voies métaboliques de Recon2.2 et donc d'obtenir une première analyse fonctionnelle de la perturbation métabolique via des approches classiquement utilisées et comparable à celles utilisées lors de l'analyse fonctionnelle de listes de DEGs. Nous avons également pu identifier certaines limites. Ces limites sont d'une part dues à des choix méthodologiques tels que la définition de seuils, de quantité de conditions modélisées ou propres à l'analyse de sur-représentation comme vous avons pu l'évoquer.

Comme nous avons également pu le mentionner lors de l'analyse des résultats de sur-représentation des DARs dans les voies métaboliques pour les 8 molécules testées, l'impact métabolique est majoritairement spécifique de chaque composé. Il serait intéressant d'appliquer la stratégie de modélisation présentée dans ces travaux sur un plus grand nombre de composés. En effet, identifier l'impact métabolique d'un grand nombre de molécules à des

concentrations et des temps d'exposition différents pourrait permettre de calculer la spécificité de chacune des réactions pour différentes conditions : nous pourrions ainsi identifier si une réaction est perturbée dans un très grand nombre de conditions, concentrations, ou seulement après un certain temps d'exposition. Il serait également possible d'identifier des groupes de réactions dont les fréquences d'activation sont corrélées et donc identifier des mécanismes de régulation du métabolisme cellulaire. Calculer la spécificité des réactions métaboliques pourrait donc permettre une analyse plus globale des perturbations métaboliques en catégorisant cet impact métabolique par rapport à l'impact métaboliques d'autres composés. Cela permettrait également d'aller vers l'identification d'analogues biologiques en identifiant des signatures de DARs propres à certaines classes de molécules.

Le prochain chapitre a pour objectif d'apporter une réponse aux limites analytiques évoquées précédemment en nous appuyant sur des approches basées sur la théorie des graphes et la visualisation afin d'approfondie l'analyse des mécanismes d'action métabolique pour 2 des 8 molécules étudiées dans ce chapitre.

Chapitre 4 : Utilisation d’approches de graphe pour mettre en évidence et visualiser le mécanisme d’action métabolique

Comme nous avons pu l'aborder au cours du chapitre précédent, les analyses de sur-représentation sont une première étape intéressante pour l'interprétation des mécanismes d'action d'une molécule mais ne permettent pas à elles seules de comprendre le mécanisme d'action métabolique. Nous allons donc tenter d'améliorer l'interprétation des mécanismes d'actions métaboliques prédits par notre stratégie en s'appuyant sur des approches basées sur des graphes. Dans un premier temps, nous utiliserons la topologie du graphe des réactions pour calculer des distances métaboliques entre toutes les DARs, distances que l'on considère représentatives du lien fonctionnel existant entre ces réactions. Ensuite, nous utiliserons ces distances pour identifier des groupes fonctionnels de DARs pour lesquels nous extrairont des sous-graphes de taille facilement analysable visuellement (moins d'une centaine de nœuds) permettant ainsi de tendre vers une meilleure identification et une meilleure compréhension du mMoA des xénobiotiques testés.

Le formalise des graphes s'applique particulièrement bien à la représentation du métabolisme. En effet, comme nous avons pu le mentionner dans le chapitre 1, ce formalisme permet de représenter des systèmes dits complexes, c'est-à-dire constitués de plusieurs milliers d'éléments (*e.g.* métabolites et réactions) interagissant les uns avec les autres.

Comme cela a été mentionné précédemment, les approches d'analyse de sur-représentation des DARs dans les voies métaboliques considèrent les DARs comme indépendantes les unes des autres et ne tirent donc pas parti des interactions existantes entre les DARs au sein du réseau métabolique. A l'inverse, les approches de graphe sont capables de prendre en compte les interactions entre les DARs et donc de proposer un mécanisme d'action métabolique intégrant des réactions métaboliques n'étant pas des DARs mais permettant d'interconnecter ces DARs.

Avant d'aborder les aspects méthodologiques de ce chapitre, nous allons rapidement décrire les particularités topologiques des réseaux métaboliques à l'échelle du génome qu'il nous semble important de connaitre afin d'obtenir un graphe métabolique ayant la topologie qui soit la plus pertinente possible pour l'utilisation d'algorithmes de théorie des graphes.

1. Particularités topologiques des réseaux métaboliques à l'échelle du génome

1.1. Compartimentation des réseaux métaboliques

Le but d'un GSMN est d'une part de représenter fonctionnellement et structurellement (par la topologie du réseau) le métabolisme d'un organisme et également de servir de plateforme pour simuler le métabolisme dans différentes conditions [212]. Les cellules eucaryotes sont constituées de sous-unités (ou compartiments) réalisant des fonctions biologiques spécifiques. On dénombre 13 sous-unités appelées organites dans une cellule eucaryote (Fig 40). Les organites principaux sont le cytoplasme, le noyau, le peroxysome, le lysosome, le réticulum

(lisse et rugueux), l'appareil de golgi et la mitochondrie.

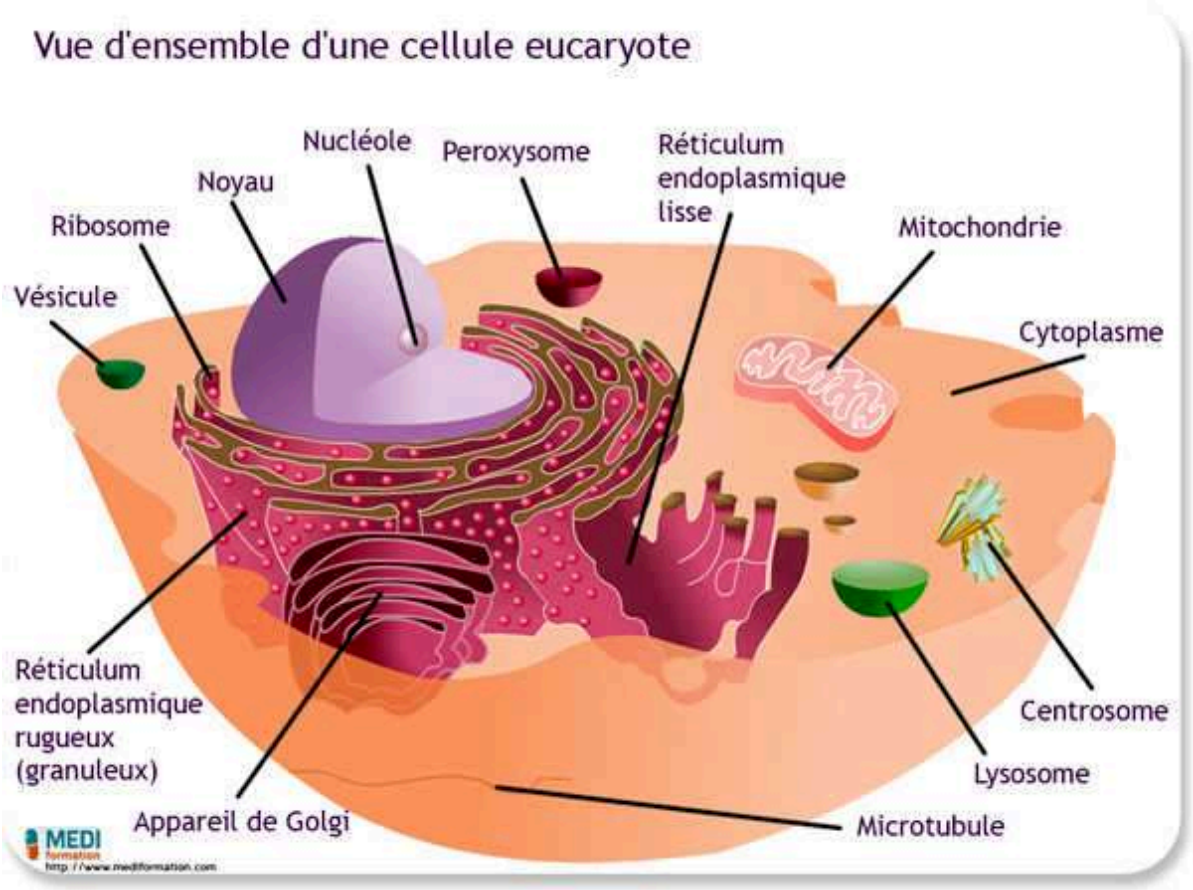

Figure 40: Schéma représentant une cellule eucaryote et les différents organites dont elle est composée. *Schéma provenant de* [213]

La compartimentation cellulaire joue un rôle important dans le fonctionnement de la cellule et notamment pour le métabolisme cellulaire en créant des environnements avec des propriétés chimiques particulières, en protégeant les autres organites d'espèces réactives et en permettant la régulation de certains processus métaboliques[214]. De fait, afin de représenter le plus fidèlement possible le fonctionnement du métabolisme cellulaire, ces compartiments cellulaires ont été pris en compte dès 2007 avec la publication de Recon1 [215] et continueront à l'être au fil des publications de nouveaux GSMN humains [72,95,216,217].

La compartimentation des réseaux métaboliques a des répercussions sur la topologie de ces réseaux et donc sur le fonctionnement des algorithmes s'appuyant sur les propriétés topologiques telles que la distance (*i.e.* shortest-paths) ou la modularité (*i.e.* détection de communauté). Comme nous l'avons mentionné précédemment, nous avons choisi d'utiliser Recon2.2, qui est constitué de 8 compartiments représentants 8 organites différents, 1 compartiment représentant le milieu extracellulaire ainsi qu'un compartiment nécessaire à la modélisation (impliqué dans la régulation des flux d'imports [218]). Le tableau 18 résume la répartition des réactions et métabolites parmi ces 10 compartiments. La taille des

compartiments dans Recon2.2 est variable, allant de 6 à 4556 réactions. La continuité biochimique est assurée par des réactions de transport entre ces différents compartiments.

Compartiment	*Nombre de réactions*	*Nombre de métabolites*
Limites (compartiment théorique pour la modélisation)	745	722
Cytoplasme	4556	1919
Réticulum endoplasmique	876	675
Milieu extracellulaire	2648	770
Appareil de Golgi	373	312
Lysosome	336	291
Espace mitochondrial intermembranaire	6	1
Mitochondrie	1133	756
Noyau	200	161
Péroxysome	527	440

Tableau 18. Répartition des réactions et métabolites par compartiment cellulaire dans Recon2.2.

Il est important de noter que l'attribution des réactions à des compartiments peut être subjective, notamment pour les réactions de transport à l'interface entre deux compartiments. La compartimentation des réseaux métaboliques est également à l'origine d'une augmentation du nombre de métabolites dans le réseau puisque chaque métabolite sera dupliqué un nombre de fois équivalent au nombre de compartiments dans lequel il participe à des réactions biochimiques. Ce qui peut également avoir un impact topologique et notamment pour les approches d'identification automatique des cofacteurs dont nous allons discuter dans les prochaines sections.

1.2. Rôle central des cofacteurs dans la topologie des graphes métaboliques

Les réseaux métaboliques ont un temps été considérés comme des réseaux dits « scale-free » (réseau à invariance d'échelle) [219–221] . Un réseau « scale-free », est un réseau dont la plupart des nœuds ont un faible degré (*i.e.* le nombre d'arêtes connectées à un nœud) et sont donc très peu connectés aux autres nœuds alors que quelques nœuds sont à l'inverse très connectés aux autres nœuds du réseau et présentent donc un fort degré. Dans ces réseaux, la

distribution des degrés suit une loi de puissance indiquant que pour deux quantités, l'une varie à la puissance de l'autre. Pour un graphe $G = (V, E)$ suivant une loi de puissance, le nombre y_i de nœuds de degré i, est proportionnel à $i^{-\beta}$: $y_i \propto i^{-\beta}$, avec $\beta > 1$ [222], une constante correspondant à l'exposant de la loi de puissance du graphe G. Pratiquement cela signifie que la proportion d'un ensemble de nœuds de degré i décroit en suivant un certain ratio de puissance β jusqu'à un nombre de degré d. Dans un réseau « scale-free », les nœuds de haut degré sont peu nombreux et « dominent » le réseau. La plupart des nœuds du réseau sont organisés autour de ces nœuds (i.e. hubs). Enfin la longueur moyenne des chemins entre toutes les paires de nœuds du réseau correspond au minimum théorique d'un graphe aléatoire. Enfin, dans un réseau biologique « scale-free », on considère généralement que les hubs (*i.e.* les nœuds de haut degré) correspondent à des fonctions biologiques importantes [221,223].

Arita *et al.* ont montré qu'en prenant en compte la conservation atomique (suivi des atomes de entre les substrats et les produits d'une réaction) des réactions métaboliques, la longueur moyenne des chemins entre les paires de nœuds était plus importante qu'initialement publié par [219–221] et donc que les réseaux métaboliques, dans ces conditions, n'était pas des réseaux « petit monde ». En prenant en compte les transitions carbonées pour chacune des réactions, Arita *et al.* ont pris en compte l'une des particularités topologiques des réseaux métaboliques, à savoir la présence de nœuds très connectés mais non « mécanistiquement informatifs ». Ces hubs correspondent aux cofacteurs des réactions biochimiques, qui bien que jouant un rôle essentiel dans le métabolisme ne sont, dans la très grande majorité des cas, pas spécifiques des réactions auxquelles ils sont associés et agissent comme des « raccourcis » sur la topologie du réseau métabolique.

En effet, bien que ces composés soient biochimiquement importants et utilisés par de très nombreuses réactions, ils n'ont généralement pour rôle que d'apporter l'énergie ou l'équilibre redox (réduction/oxydation) nécessaire à la réaction. De par leur impact sur la topologie du réseau et les « raccourcis » qu'ils peuvent engendrer, ces composés posent donc un problème pour l'utilisation d'approches de théorie des graphes sur les réseaux métaboliques.

GSMN	*Degré maximal*	*Degré moyen*	*Moyenne des plus courts chemins*	*Nombre de composantes connexes*	*Diamètre*
Recon2.2	3979	19,56	3,83	16	11
Recon2.2 (retrait des cofacteurs)	1510	8,14	7,12	239	27

Tableau 19. Propriétés topologiques de Recon2.2 avec et sans retrait des cofacteurs. *Propriétés calculées grâce à l'application NetworkSummary de la librairie java Met4J (https://forgemia.inra.fr/metexplore/met4j). Les composés identifiés comme cofacteurs sont issus de la liste provenant du serveur web Metexplore* [1]

Sur la Figure 41, nous avons représenté sous la forme d'un graphe biparti 3 réactions de la voie de la glycolyse. Sur la Figure 41A, les cofacteurs n'ont pas été retirés (à l'inverse de la Figure 41B) et on s'aperçoit que le chemin le plus court (dans le cas non dirigé) entre l'hexokinase et la phosphofructokinase ne passe pas par la G6P isomérase mais plutôt pas l'un des 3 cofacteurs (ATP, ADP ou proton). Or ces chemins n'ont pas de réel sens biologique. En retirant ces cofacteurs, la topologie du réseau respecte les interactions attendues dans le cadre de la voie de la glycolyse pour ces 3 réactions (Fig 41B).

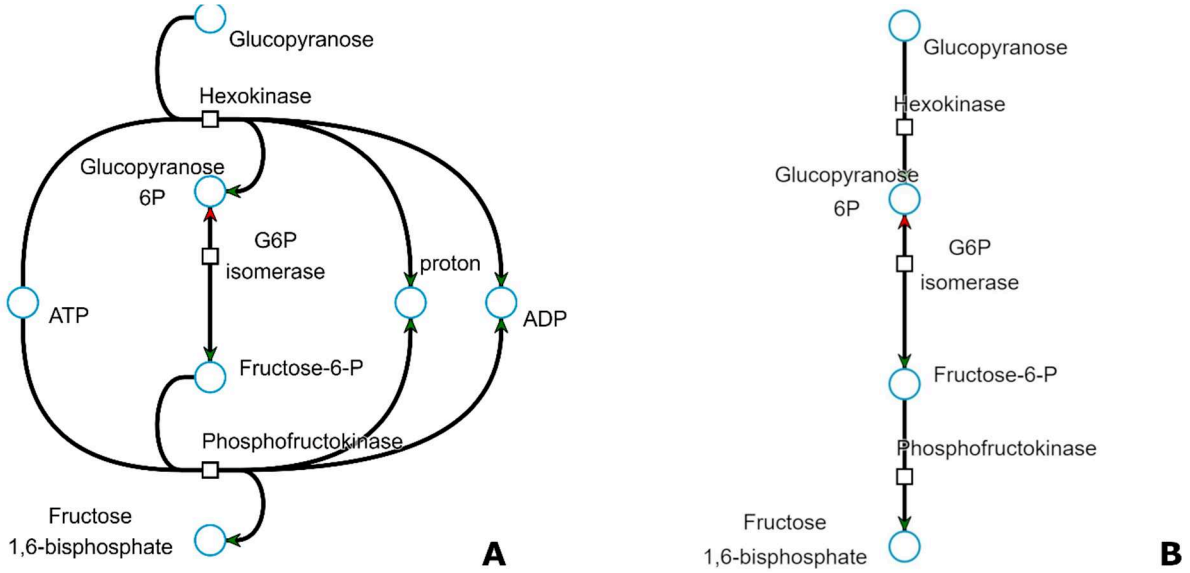

Figure 41 : Illustration de l'impact des cofacteurs sur la topologie d'un réseau de 3 réactions appartenant à la voie de la glycolyse. *Les nœuds représentés par des carrés représentent les réactions et les nœuds représentés par des cercles représentent les métabolites. La figure 41A correspond au sous-réseau avec l'ensemble des métabolites associées à ces 3 réactions. La figure 41B correspond au sous-réseau avec les métabolites « cofacteurs » et inorganiques retirés.*

Il convient donc d'identifier et retirer ces cofacteurs afin de garantir une topologie plus pertinente d'un point de vue mécanistique, notamment lors de l'utilisation d'algorithmes de recherche des plus courts chemins, de partitionnement de graphes ou d'extraction de sous-graphes comme ce sera le cas dans les prochaines sections.

2. Simplifier l'identification des cofacteurs via le développement d'une approche d'annotation automatique des arêtes du graphe

L'identification des cofacteurs dans le réseau métabolique n'est cependant pas triviale. Elle peut être réalisée manuellement, à partir d'une liste de co-facteurs définis *a priori* ou automatiquement via des approches basées sur le degré [224,225] ou la similarité chimique [226,227]. Il est important de noter qu'un composé peut être cofacteur d'une réaction et métabolite « principal » dans une autre, ce qui rend l'identification de composés comme cofacteurs de manière générale pour l'ensemble du réseau source d'erreur dans certain cas. Par exemple, l'ATP est un cofacteur dans un très grand nombre de réactions cependant dans certains cas particuliers comme la synthèse des nucléotides, l'ATP n'est pas un cofacteur mais le composé source principal [228]. Afin de limiter ce biais, nous avons choisi de développer une approche capable de différencier automatiquement et par rapport au contexte local les arêtes connectant les réactions partageant (une réaction produisant, l'autre consommant le composé) un composé « principal » des arêtes connectant les réactions partageant un cofacteur.

2.1. Filtration des arêtes correspondant à des transitions carbonées dans le graphe des composés

On appelle « transition carbonée » la conservation de tout ou partie de la structure carbonée (le nombre et la position des atomes de carbones) lors de la transformation d'un métabolite source vers un métabolite produit. Par exemple, pour la réaction de décarboxylation du malate en pyruvate (Fig 42), aucun atome de carbone n'est partagé entre le malate et le NADPH. En revanche, le malate partage 3 atomes de carbone avec le pyruvate et 1 atome de carbone avec le CO2. De la même manière, le NADP+ ne partage aucun atome de carbone avec le Pyruvate ou le CO2 mais tous ses atomes de carbone avec le NADPH.

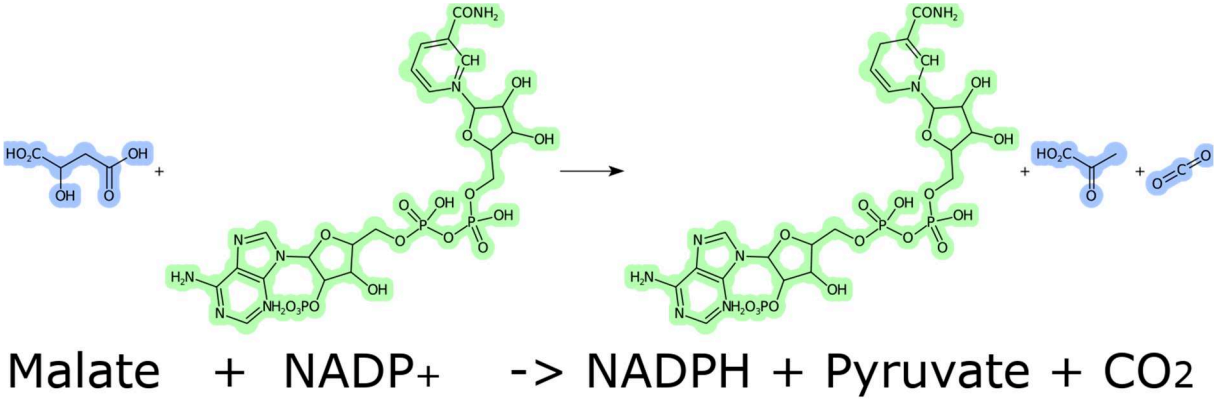

Figure 42 : Exemple de suivi des atomes entre les métabolites source et les métabolites produits de la réaction de décarboxylation du Malate en Pyruvate. *Figure réalisée avec CDKDEPICT (https://www.simolecule.com/cdkdepict/depict.html).*

Cela peut se résumer avec les transitions carbonées suivantes :

- Malate -> Pyruvate (3 atomes de carbone)
- Malate -> CO2 (1 atome de carbone)
- NADP+ -> NADPH (21 atomes de carbone)

Ainsi, prendre en compte les transitions carbonées lors de la construction du graphe des composés reviendrait à connecter le nœud malate au nœud pyruvate, le nœud malate au nœud CO2 et le nœud NADP+ au nœud NADPH (Fig 43A). En revanche sans tenir compte des transitions carbonées, dans un graphe des composés « classique », le nœud malate serait également connecté au nœud NADPH, le nœud NADP+ serait connecté au nœud CO2 et enfin le nœud NADP+ serait connecté au nœud Pyruvate (Fig 43B).

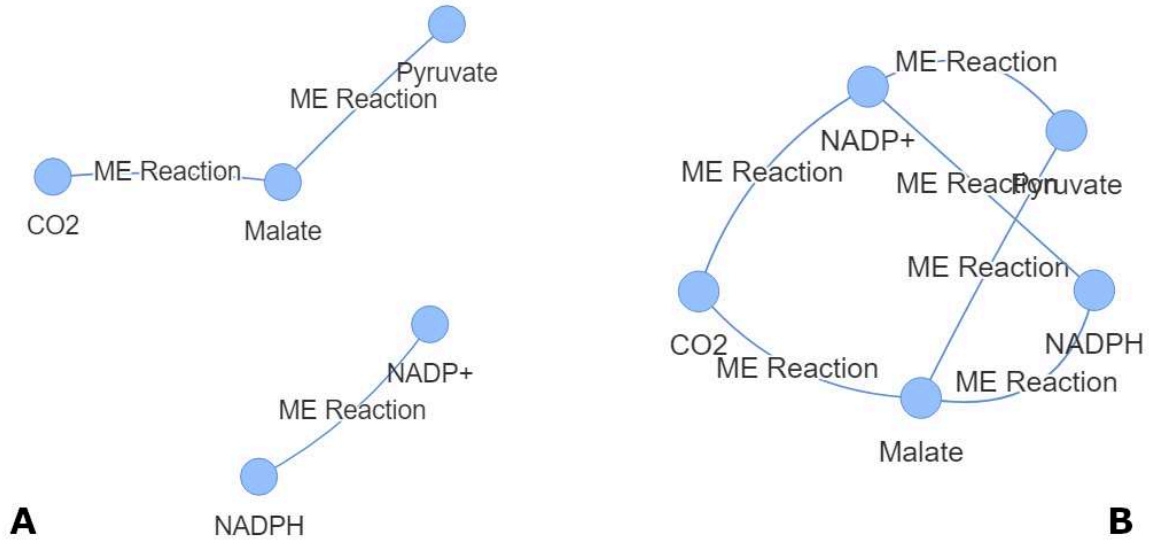

Figure 43 : Graphe des composés pour la réaction de décarboxylation du malate en pyruvate. *La figure 43A correspond au graphe des composés lorsque les transitions carbonées sont prises en compte. La figure 43B correspond au graphe des composés lorsque les transitions carbonées ne sont pas prises en compte.*

La prise en compte des transitions carbonées permet donc d'éviter les arêtes secondaires connectant un composé principal à un composé cofacteur, arêtes qui sont à l'origine des « raccourcis » dans les chemins métaboliques que l'on a mentionnés précédemment. En revanche on peut s'apercevoir sur la figure 43A que les composés à un seul carbone tel que le CO2 (qui est généralement considéré comme cofacteur) peuvent avoir une transition carbonée avec le métabolite principal.

A partir de ces observations et de l'état de l'art [223,229–232], de premiers travaux ont été réalisés au sein de l'équipe par le Dr Clément Frainay afin de calculer le graphe des composés basé sur les squelettes carbonés à partir d'un GSMN. En effet, les molécules organiques possèdent un squelette de carbones lié par des liaisons covalentes qui correspond à la « base » de la molécule. L'hypothèse posée est que si l'on ne considère que les transformations

biochimiques impliquant cette base moléculaire, modifiée ou non, la topologie du graphe des composés correspond sera plus pertinente.
Dans un premier temps, les transitions carbonées sont calculées avec GSAM (https://forgemia.inra.fr/metexplore/gsam) pour toutes les réactions du modèle pour lesquelles la structure 2D des métabolites (sous la forme de SMILES) est connue. Ensuite le graphe des composés pour l'ensemble du GSMN est construit grâce à Met4J (*https://forgemia.inra.fr/metexplore/met4j*) et les arêtes connectant des composés source/produit pour une réaction sans partager de carbone (sans transition carbonée) sont retirées. Les nœuds correspondant à des composés dont la formule brute contient moins de 2 carbones, qui par définition n'ont pas de « squelette » carboné, sont également retirés.
On obtient alors le graphe des composés du GSMN initial avec une topologie plus pertinente d'un point de vue mécanistique et qui permet de s'affranchir des limites associées à la définition d'une liste de cofacteurs telles que le manque de généricité du statut cofacteur (un métabolite cofacteur en général mais pas pour certaines réactions métaboliques) ou encore le manque de reproductibilité (la liste de cofacteurs retirés n'est pas systématiquement donné lorsque cette approche est utilisée dans la littérature). Ce graphe correspond au graphe des squelettes carbonés du réseau métabolique, noté CSN.

Dans un graphe des réactions, les métabolites sont représentés par les arêtes qui connectent deux nœuds réactions si ce métabolite est substrat de l'une et produit de l'autre. Cependant, cette approche n'est pas directement adaptable au graphe des réactions. En effet, contrairement au graphe des composés, le sous-graphe induit des réactants (un sous graphe contenant uniquement les réactants et les arêtes correspondantes d'après le graphe initial) d'une réaction n'est pas nécessairement connexe, ce qui impliquera de connecter des nœuds ne devant pas l'être lors du passage du CSN vers son équivalent sur le graphe des réactions.
Il n'est donc pas possible d'utiliser directement les transitions carbonées pour retirer des arêtes entre les nœuds réactions. L'identification automatique d'arêtes principales (*e.g.* une arête connectant le Malate au Pyruvate) et d'arêtes secondaires (*e.g.* une arête connectant le NADP+ au NADPH) nécessite donc le développement d'une nouvelle approche qui s'appuie sur le CSN et sa topologie afin d'identifier automatiquement et pour chacune des réactions les arêtes principales des arêtes secondaires lors de la construction du graphe des réactions.

2.2. Extension au graphe des réactions pour l'annotation automatique d'arêtes

Kotera *et al.* ont développé en 2004 les RPAIRS [233]. Les RPAIRS sont un jeu de données constitués de paires de substrat/produit pour un grand nombre de réactions métaboliques de la base de données KEGG. Ces paires substrat/produit ont été séparées en 5 catégories : « main », « cofac », « trans », « ligase », and « leave ». Cette assignation a été réalisée

manuellement ce qui a permis d'obtenir une annotation de qualité mais a représenté une grande quantité de travail, ce qui est certainement l'une des raisons justifiant l'arrêt du projet en 2016. Plusieurs approches [234–236] s'appuyaient sur ces annotations jusqu'à cette date.

Afin de combler le vide laissé par l'arrêt du projet d'annotation RPAIRS et faciliter le nettoyage des graphes des réactions métaboliques, nous avons donc développé une méthode permettant d'identifier automatiquement et pour chaque réaction les paires de substrat/produit « main » et « side ».

La méthode que nous avons mis au point et implémentée dans la librairie Met4J est la suivante :

La première étape consiste à construire le graphe des composés du squelette carboné d'un GSMN d'intérêt. Cette étape est réalisée grâce à l'approche décrite lors de la section précédente et développée par le Dr Clément Frainay.

La seconde étape consiste à itérer sur les arêtes du graphe des composés du squelette carboné. Pour rappel, dans un graphe des composés, deux nœuds sont connectés par une arête représentant une réaction si ces nœuds sont substrats/produits de cette réaction. Pour chacune de ces arêtes, le graphe des composés propre à cette réaction est reconstruit (Fig 44). On identifie ensuite les composantes connexes de ce graphe des composés et pour chaque composante un score de redondance est calculé (1). Ce score mesure la redondance des arêtes de la composante étudiée dans le graphe des composés représentant le squelette carboné du GSMN.

$$comp_{score} = N_{edge\_comp} / \sum_{e \epsilon comp_{edges}} e_{parallel} \epsilon CSN \quad (1)$$

Avec N_{edge_comp}, le nombre d'arêtes de la composante connexe étudiée divisé par la somme des arêtes parallèles retrouvées dans le CSN pour chacune des arêtes de la composante. Un score de composante élevé signifie que les arêtes de la composante sont faiblement redondantes dans le CSN et que les arêtes de cette composante sont probablement des arêtes « principales ». En effet, les arêtes « side » sont des arêtes existant en grand nombre dans le graphe des composés (*e.g.* l'arête ATP-ADP, l'arête NADP+-NADP). Pour rappel, des arêtes sont dites parallèles lorsqu'elles connectent les mêmes nœuds source/cible. Les scores de toutes les composantes sont ensuite classés par ordre décroissant. Les arêtes de la composante ayant le score le plus élevé parmi toutes les autres composantes du graphe de la réaction étudiée seront annotées « main » alors que les autres arêtes du graphe de la réaction étudiée seront annotées « side ».

Prenons par exemple le cas de la réaction de décarboxylation du malate en pyruvate étudiée dans la section précédente. Le sous-graphe des composés correspondant à cette réaction

d'après la topologie du squelette carboné est constitué de deux composantes connexes (Fig 44). La première composante est constituée de l'arête reliant le nœud malate au nœud pyruvate et la seconde est constituée de l'arête reliant le nœud NADP+ au nœud NADPH. L'arête malate -> pyruvate n'a pas d'arête parallèle dans le CSN (*i.e.* elle n'existe qu'en un seul exemplaire dans le graphe), ce qui donne à cette première composante un score de 1. L'arête NADP+ -> NADPH (qui correspond à une arête entre deux cofacteurs) existe en 235 exemplaires dans le CSN, soit un score très faible de 0,004. Pour cette réaction, l'arête principale sera donc l'arête malate -> pyruvate.

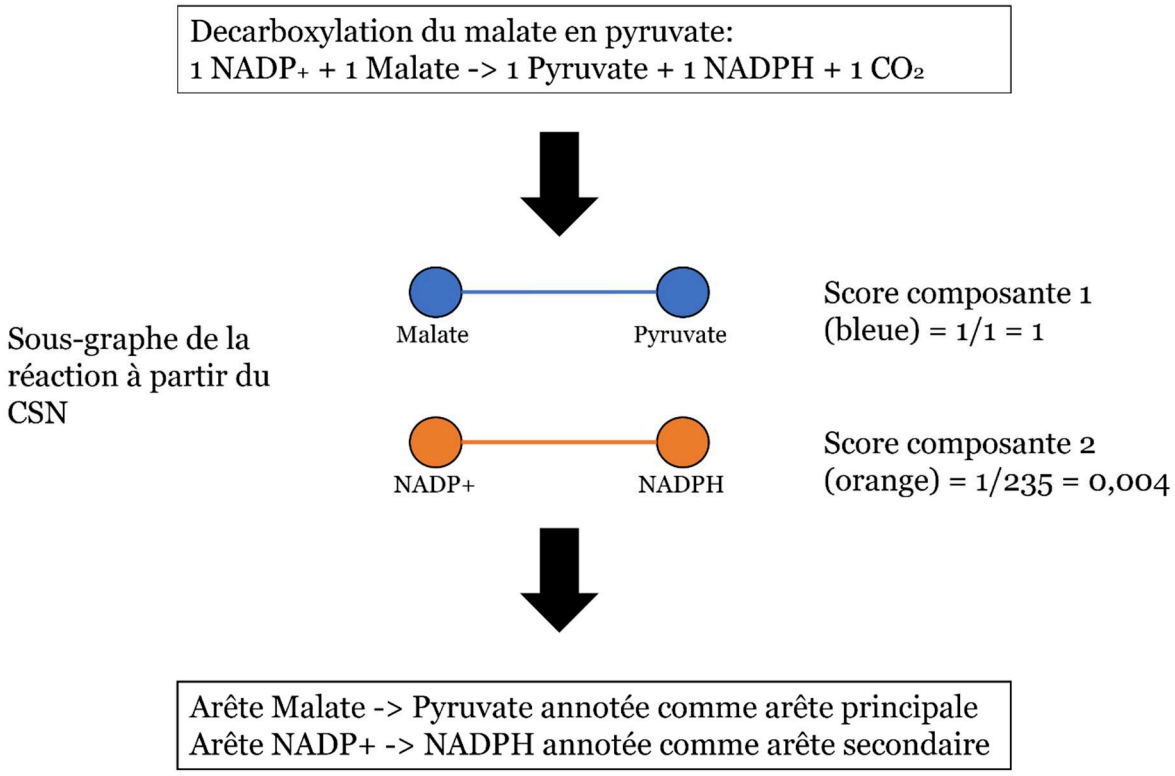

Figure 44 : Schéma du fonctionnement de l'approche d'annotation automatique des arêtes.
Exemple sur la réaction de décarboxylation du malate en pyruvate.

En calculant ce score pour les composantes connexes de l'ensemble des sous-graphes de réactions du CSN de Recon2.2, il est ainsi possible de déterminer automatiquement pour chaque réaction quelles arêtes sont principales et quelles arêtes sont secondaires. Il est intéressant de noter que si le graphe du squelette carboné prend en compte la compartimentation (ce qui est le cas pour le CSN de Recon2.2), alors cette approche sera en mesure d'identifier correctement des arêtes qui sont secondaires dans un cas mais principales dans d'autres puisque la méthode d'annotation automatique prend plus en considération le contexte local (compartiment et graphe de la réaction) qu'une approche basée sur l'identification générale d'une liste de cofacteurs. En effet, dans un graphe prenant en compte la compartimentation, chaque métabolite existe en plusieurs exemplaires (donc plusieurs nœuds). Chaque exemplaire correspond au même métabolite mais localisé dans un

compartiment cellulaire différent donc susceptible d'interagir avec des réactions et des métabolites différents. Cela aura donc pour effet d'augmenter la précision du calcul des arêtes parallèles qui dans ce type de graphes prend en compte la compartimentation cellulaire. Par exemple, considérons le métabolite 1 et le métabolite 2, connectés dans un graphe des composés non compartimenté par 200 arêtes parallèles. Ces mêmes métabolites connectés dans un graphe des composés prenant en compte 3 compartiments pourraient avoir ce type de connexions :

Metabolite_1_compartiment_1 – Metabolite_2_compartiment_1 : 25 arêtes parallèles

Metabolite_1_compartiment_2 – Metabolite_2_compartiment_1 : 125 arêtes parallèles

Metabolite_1_compartiment_3 – Metabolite_2_compartiment_2 : 50 arêtes parallèles

Cependant, nous avons identifié un cas pour lequel notre approche d'annotation automatique des arêtes ne fonctionne pas correctement. Lorsque le graphe de la réaction a plusieurs composantes que l'on considère principales (*e.g.* une réaction de co-transport ou une réaction de transfert de fonction chimiques), le classement des composantes connexe du sous graphe de la réaction impliquera d'annoter arbitrairement certaines arêtes comme secondaire alors que ces arêtes représentent des transitions principales. Par exemple, la réaction de tyrosine transaminase n'a pas d'arête secondaire mais deux arêtes principales (Fig 45).

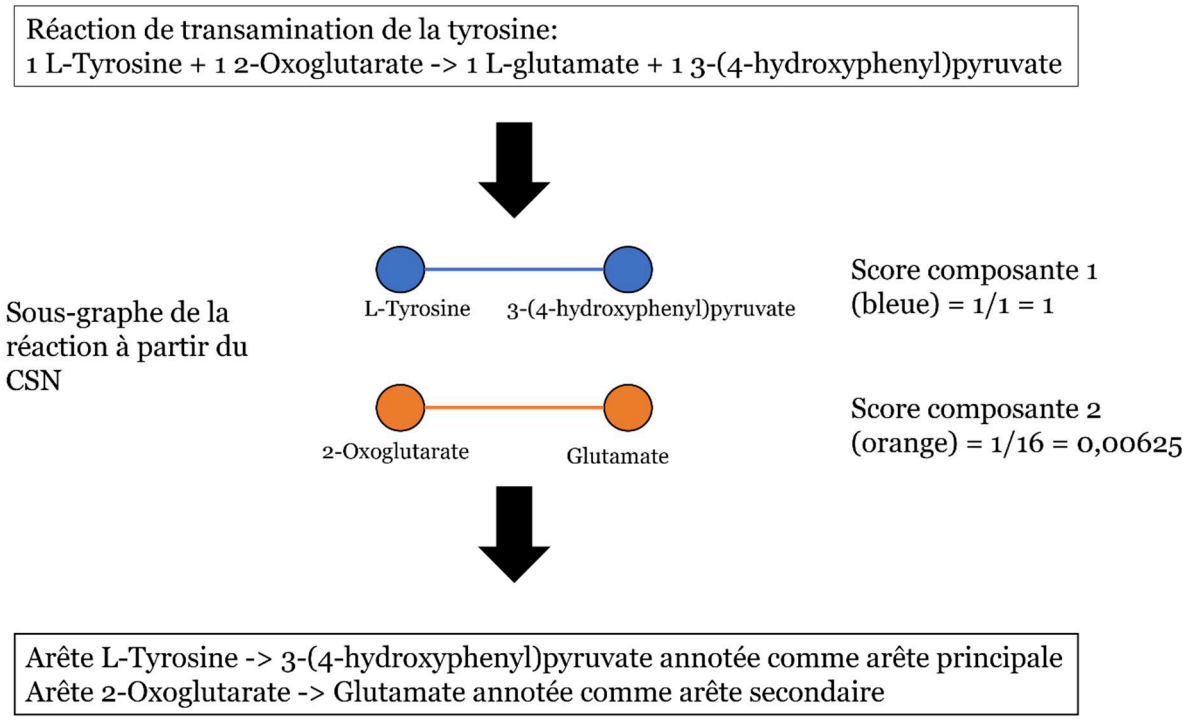

Figure 45 : Schéma du fonctionnement de l'approche d'annotation automatique des arêtes dans un cas limite de l'approche. *Exemple sur la réaction de transamination de la tyrosine.*

Etant donné qu'il n'y a pas de partage de carbones entre la L-Tyrosine et le Glutamate ni entre le 2-Oxoglutarate et le 3-(4-hydroxyphenyl)pyruvate, ces composés ne sont pas connectés dans le CSN. Cela explique pourquoi il y a deux composantes connexes dans le sous-graphe des composés de la réaction de tyrosine transaminase obtenu à partir du CSN. Les arêtes de cette réaction seront donc mal annotées par notre approche qui considérera l'arête 2-Oxoglutarate -> glutamate comme arête secondaire alors qu'il s'agit probablement plus d'une arête principale également. Notons tout de même que dans ce type de cas de figure, l'annotation des arêtes reste tout de même fortement subjective (certains experts considéraient peut-être que l'arête doit être annotée « secondaire » alors que d'autres considérerais cette arête « principale »)

Il est acceptable de faire l'hypothèse que chaque réaction comporte au moins une transition principale, et que la plus exclusive en fasse partie. La méthode permet donc de définir une composante principale pour chaque réaction. En revanche, comme montré dans l'exemple précédent, il reste néanmoins difficile d'établir le statut « secondaire » pour les autres composantes. L'approche peut donc fournir une base de travail pour assister la sélection d'une liste de composés « side » par réaction, qui demeure néanmoins toujours dépendante d'une sélection manuelle. La poursuite de ces travaux est donc nécessaire afin d'obtenir une méthode purement automatique pour le prétraitement des graphes métaboliques.

Dans l'attente de l'aboutissement de ces travaux en cours, nous avons choisi d'avoir recours à une liste manuelle de cofacteurs et de molécules inorganiques (Tableau 21, en annexe) afin de corriger la topologie du graphe des réactions de Recon2.2 utilisé pour la suite des analyses.

3. Utiliser la topologie du graphe pour calculer des distances métaboliques

La distance entre deux réactions dans un graphe métabolique simple correspond au nombre d'intermédiaire minimal nécessaire pour connecter ces deux réactions. Ainsi, dans le graphe des réactions, la distance métabolique entre deux réactions est le nombre (minimal) de réaction permettant de connecter ces deux réactions. Etant donné que deux réactions sont connectées si l'un produit ce que l'autre consomme, il est possible de faire l'hypothèse que des réactions proches dans le réseau métaboliques réalisent une fonction similaire ou complémentaire. Nous allons donc calculer la distance entre toutes les paires de DARs prédites précédemment afin d'identifier, pour chaque condition, des groupes de DARs partageant des fonctions similaires/complémentaires.

Le calcul des distances métaboliques étant sensible à la topologie du graphe métabolique, il est important de retirer les cofacteurs et les molécules inorganiques lors de la construction du graphe afin d'avoir une topologie pertinente. Pour calculer ces distances métaboliques à partir de graphes métaboliques, il existe plusieurs algorithmes avec différentes capacités et performances selon l'objectif et le type de graphe. Nous avons envisagé l'utilisation de 3 algorithmes différents pour calculer les distances métaboliques entre toutes les paires de DARs pour une condition. Les principales propriétés nous ayant permis de choisir l'algorithme le plus adapté sont résumées dans le tableau 20.

Algorithme	Complexité(temps)	Objectif	Parallélisation	Liste de seeds/targets
Djikstra	$\theta(\|E\| + \log\|V\|)$	Plus court chemin entre un nœud et tous les autres	Non	1 nœud seed
Floyd-Warshall	$\theta(\|V\|^3)$	Plus courts chemins entre les paires de nœuds d'un graphe	Non	Non
Many to Many SP	Inconnue	Plus courts chemins entre les seeds/targets	Oui	Oui

Tableau 20. Caractéristiques des trois algorithmes de calcul des plus court-chemins envisagés.

L'algorithme de Djikstra est plus performant que l'algorithme de Floyd-Warshall pour trouver tous les chemins partant d'un nœud source mais moins performant que Floyd-Warshall

lorsqu'il s'agit de trouver les plus courts chemins entre toutes les paires de nœuds d'un graphe ($\theta(|V|^3\ Log\ |V|)$. Dans notre cas, nous cherchons à calculer le plus court chemin entre toutes les paires de DARs. Il serait possible de le faire avec l'algorithme de Dijkstra en lançant une recherche de chemins pour chaque DAR (nœud seed) mais le temps de calcul augmentera fortement avec le nombre de DAR et la recherche de chemins ne serait alors pas optimisée car elle impliquerait de recalculer des chemins déjà calculés lors des recherches de chemin précédentes (*i.e.* pour les DARs précédentes) [237]. Une autre option serait de calculer l'ensemble des plus courts chemins du graphe métabolique avec l'algorithme de Floyd-Warshall puis de filtrer la matrice de distance obtenue afin de ne conserver que les distances entre les paires de DARs. Cette option est pertinente si l'on désire développer une approche centrée sur un GSMN particulier car il suffit alors de calculer ces distances une seule fois puis de filtrer la matrice selon les DARs prédites pour chaque condition. Cependant, notre objectif est de développer une stratégie qui soit réutilisable avec différents GSMNs. Nous avons donc opté pour l'algorithme de « Many to Many Shortest-path » qui permet de rechercher l'ensemble des chemins entre un ensemble de nœuds sources et un ensemble de nœuds cibles. Cet algorithme est basé sur le concept des « highway hierarchies » qui est une méthode de pré-traitement du graphe permettant de simplifier la topologie de ce graphe (sur plusieurs niveaux). Ces approches de pré-traitement puis de recherche du plus court chemin sont couramment utilisées pour la recherche du plus court chemin dans un réseau routier [237].

Afin de calculer des distances entre réactions métaboliques, nous avons choisi le formalisme du graphe des réactions. Les arêtes correspondant à la connexion de deux réactions par un cofacteur ou une molécule inorganique ont donc été retirées lors de la création du graphe des réactions. Comme mentionné précédemment, nous avons utilisé une liste de cofacteurs (Tableau 21 en annexe) définie manuellement. Nous avons également retiré les réactions bloquées, inactives dans toutes les solutions énumérées pour toutes les conditions (inactives dans les HPH) ainsi que les réactions d'échange avec le milieu extracellulaire. Cette liste de réactions retirées est identique à la liste de réactions retirées lors de l'analyse de sur-représentation. Retirer ces réactions et ces arêtes permet d'avoir un graphe métabolique dont les chemins entre les réactions sont plus pertinents d'un point de vue biologique. Par exemple, cela permet d'éviter de calculer un chemin passant par une réaction qui est bloquée ou toujours inactive dans le modèle cellulaire modélisé. Un tel chemin bien que pertinent topologiquement ne serait pas forcément pertinent d'un point de vue mécanistique.

Après avoir construit et prétraité le graphe des réactions de Recon2.2, nous avons pu calculer les matrices de distances correspondant à toutes les listes de DARs prédites lors de la première phase de la stratégie (modélisation, énumération et calcul des DARs). Dans la prochaine section, nous allons voir comment ces matrices de distances peuvent permettre d'identifier des

groupes de DARs proches dans le réseau métaboliques et donc faciliter l'interprétation de ces listes de DARs.

4. Identification de groupes de DARs proches métaboliquement pour mettre en évidence des fonctions métaboliques modulées

Regrouper les DARs (les réactions prédites comme perturbées) en fonction de leur proximité dans le réseau métabolique permet d'une part de faciliter la compréhension des liens entre les DARs et d'autre part de pouvoir faciliter la représentation des résultats. Comme nous l'avons mentionné dans la section précédente, nous faisons l'hypothèse que cette distance métabolique peut être utilisée comme une mesure de la relation fonctionnelle entre deux DARs. Il est donc possible d'identifier via des approches de partitionnement s'appuyant sur ces distances métaboliques, des groupes de DARs en interaction et susceptibles de réaliser une fonction métabolique commune ou complémentaire.

Il existe un grand nombre de méthodes de partitionnement basées sur des distances. L'approche classiquement utilisée est l'approche de partitionnement hiérarchique agglomératif (« hierarchical agglomerative clustering ») qui considère dans un premier temps chaque point comme une partition. Ces partitions sont ensuite regroupées de proche en proche de manière itérative. La méthode des k-moyennes (« k-means ») est également une méthode fréquemment utilisée et se base sur la minimisation de la somme des carrés des distances entre les points d'une partition (pour chaque point on considère sa distance à la moyenne des points de son cluster).

Nous nous sommes également intéressés à d'autres méthodes de partitionnement telles que DBSCAN [238] (un algorithme de partitionnement par densité) et des méthodes de détection de communautés telles que l'algorithme de Leiden [144]. Le partitionnement basé sur le calcul de la densité (DBSCAN) est capable d'une part d'identifier des partitions dans un jeu de données mais également de considérer des éléments comme « aberrants » si ces éléments sont trop éloignés des autres éléments et ne peuvent être ajoutés à aucune partition. Bien que cela soit une caractéristique intéressante qui aurait pu nous permettre d'obtenir des partitions denses et sans valeurs (réactions) aberrantes, nous n'avons pas obtenu de résultats satisfaisants (forte sensibilité au choix des paramètres) avec cet algorithme.

Ici le partitionnement a pour principal objectif de regrouper les DARs dans des partitions comprenant quelques dizaines de DARs afin de pouvoir ensuite extraire des sous-réseaux qui soient lisibles et contenant le moins possibles de réactions non différentiellement perturbées ajoutées par l'algorithme d'extraction de sous-graphe. Nous avons donc sélectionné la méthode de partitionnement hiérarchique agglomératif d'une part pour la robustesse des partitions

calculées et d'autre part pour la possibilité de définir le nombre de partitions recherchées. Bien que la méthode des k-moyennes permette également de définir un nombre de partitions, nous avons préféré ne pas utiliser cette méthode à cause de sa sensibilité aux valeurs d'initialisations définies au hasard [239].

5. Faciliter l'interprétation fonctionnelle des groupes de DARs via l'extraction de l'arbre couvrant de poids minimum

Comme cela a été mentionné précédemment, Recon2.2 est un GSMN constitué de 7785 réactions et 5323 métabolites. Un tel réseau, bien que de taille modeste si l'on compare à la taille des réseaux utilisés dans certains domaines tels que les télécommunications, les réseaux routiers ou les réseaux sociaux, est trop dense pour être facilement analysable et interprétable par l'œil humain (Fig 46). L'objectif de cette approche est donc d'améliorer l'interprétation des groupes de DARs (ou partitions) identifiés précédemment en calculant un sous-graphe pour chaque partition qui soit de la plus petite taille possible tout en contenant les DARs de la partition considérée et un nombre minimal de réactions ajoutées pour la connectivité de sous-graphe. Visualiser les connexions entre les DARs d'une partition permettra de mettre en évidence les liens métaboliques entre ces DARs sans pour autant être confronté à la grande complexité d'analyse que représenterait une analyse des connexions entre ces DARs dans le contexte du réseaux Recon2.2 complet (Fig 46).

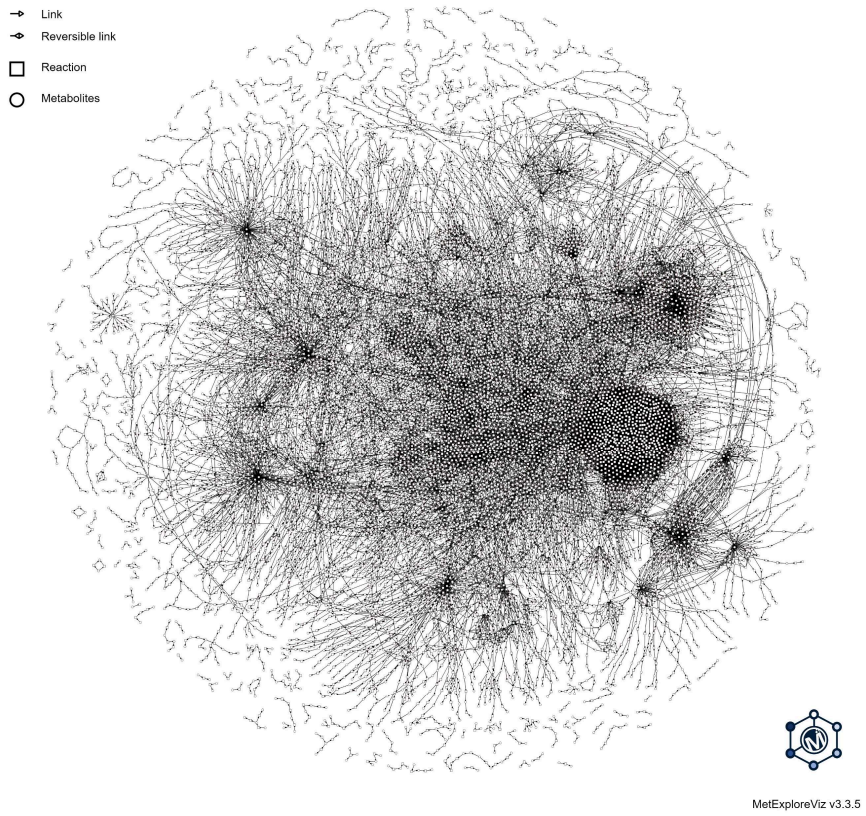

Figure 46: Visualisation de Recon2.2 sous la forme d'un graphe bipartite avec MetExploreViz.

C'est pourquoi, afin d'étudier en détail le mMoA d'une molécule nous avons choisi d'extraire un sous-réseau contenant les DARs identifiées pour chacune des partitions afin de focaliser notre attention sur l'analyse des sous-parties du réseau métabolique qui soient les plus susceptibles d'être impactées. Pour arriver à cet objectif, nous nous sommes appuyés sur l'extraction d'arbres couvrant de poids minimal grâce à la résolution d'une approximation du problème de l'arbre de Steiner.

En effet, la résolution du problème de l'arbre minimal de Steiner est computationnellement difficile (prouvé comme NP-Complet [240]). Il est donc nécessaire d'utiliser une approche heuristique permettant de trouver une solution approchée.

Nous avons utilisé l'implémentation de l'extraction de sous-réseau par approximation de l'arbre de Steiner disponible dans la librairie java Met4J (https://forgemia.inra.fr/metexplore/met4j).

Cette implémentation s'appuie notamment sur le calcul du « Metric closure graph » [241] qui consiste à calculer le graphe des plus courtes distances. Le « metric closure graph » d'un graphe G est le graphe complet pour lequel chaque arête est pondérée par la distance du plus court chemin entre le nœud source et le nœud cible de cette arête dans le graphe G (Fig 47). Un graphe complet est un graphe dont les nœuds sont adjacents (reliés par une arête) deux à deux.

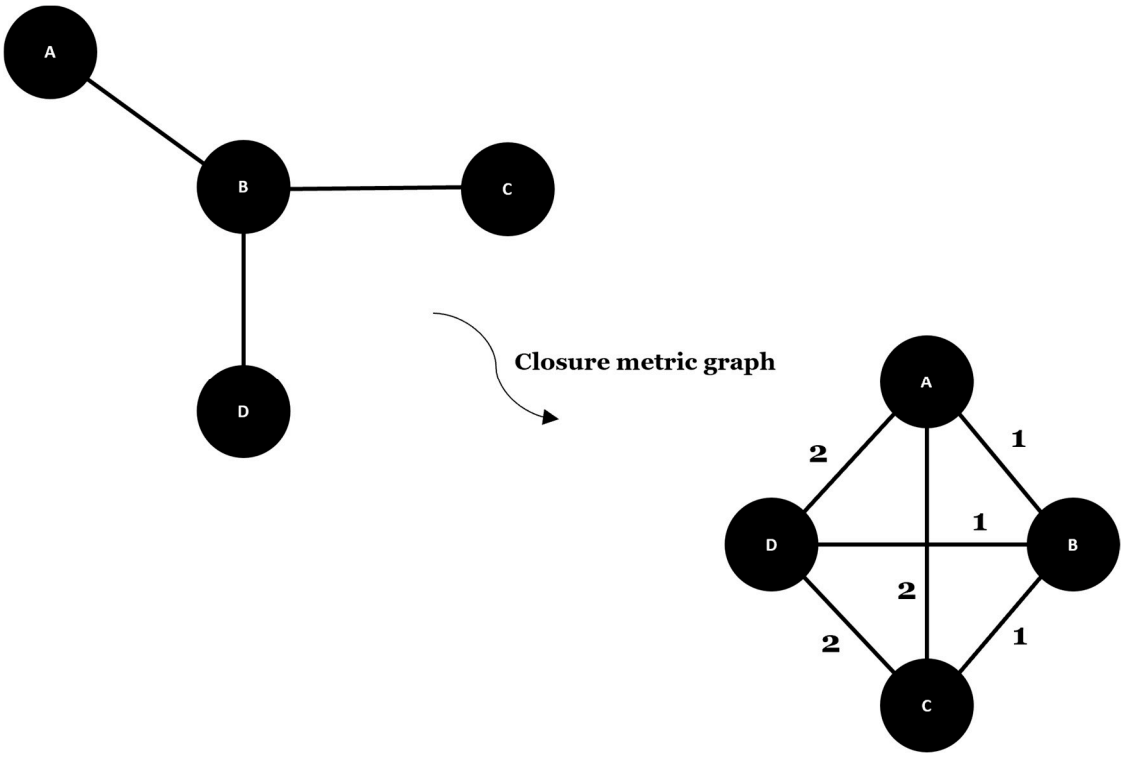

Figure 47 : Example jouet du "metric closure graph" d'un graphe simple. *Le « metric closure graph » du graphe G(V,E) avec V = {A,B,C,D} et E = {A-B,B-C,B-D} correspond au graphe complet pondéré P(V,E) avec V = {A,B,C,D} et E = {A-B,B-C,C-D,D-A,A-C,B-D}. Le poids des arêtes du graphe P correspondant à la distance entre les nœuds de P dans G.*

A partir du « metric closure graph », l'arbre couvrant de poids minimal (Minimum Spanning Tree) peut être calculé permettant ainsi de connecter les nœuds sources et les nœuds cibles (définit par l'utilisateur et correspondants aux DARs dans notre cas) en minimisant le poids total. Le poids total représente la somme du poids des arêtes (qui dans un « metric closure graph » représente la distance dans le graphe G entre les deux nœuds qu'elle connecte). En minimisant ce poids total, on minimise la distance entre les nœuds source/cible ainsi que l'ajout de nouveaux nœuds au graphe final.

Le rationnel derrière le choix du graphe minimal est que les DARs représentent l'impact métabolique et que les autres réactions qui pourraient être ajoutées par la méthode d'extraction de sous-graphe bien qu'intéressantes pour comprendre les interactions entre les DARs ne sont pas aussi représentatives du mMoA de la molécule étudiée que les DARs. Minimiser le nombre de réactions ajoutées lors de l'extraction du sous-graphe permet donc de se focaliser sur les DARs et leurs fonctions.

L'extraction de sous-réseaux à partir d'une approche basée sur la recherche d'un arbre couvrant de poids minimal permet effectivement d'obtenir des réseaux de la plus petite taille

possible, contenant toutes les réactions perturbées d'une partition en ajoutant le moins possible de réactions non significativement perturbées mais nécessaires pour la connectivité de ce sous-réseau. Cependant, comme nous avons pu le décrire précédemment, l'identification de cet arbre couvrant de poids minimal se fait grâce à une méthode heuristique. Plusieurs arbres de poids couvrant minimal sont donc possibles et l'algorithme doit ainsi choisir arbitrairement une solution parmi les différentes solutions possibles, induisant un possible biais lors de l'analyse du sous-graphe de la partition. De plus, rechercher le graphe minimal, c'est le risque de ne pas prendre en compte des chemins alternatifs pertinents entre les DARs. Il pourrait donc être intéressant d'utiliser l'algorithme des « k-shortest paths » afin de trouver les différentes alternatives (et s'affranchir des limites associées aux arbres). Cependant comme nous avons pu le préciser précédemment, l'objectif était de minimiser la taille des sous-graphes afin de maximiser leur lisibilité. Etant donné le nombre potentiellement important de chemin alternatif, l'utilisation de l'algorithme des « k-shortest paths » aurait pu fortement augmenter la taille des sous-graphes propres à chaque partition et donc fortement complexifier l'analyse. Ce type d'approche pourrait cependant être utilisé sur des partitions de très petite taille (une dizaine de DARs maximum) contenant des DARs très proches les unes des autres (ce qui tend à limiter le nombre de chemins alternatifs).

6. Exploration des mMoA de 2 molécules hépatotoxiques (amiodarone et acide valproïque) par notre stratégie d'analyse des DARs

Nous avons choisi d'appliquer notre stratégie d'analyse des DARs via des approches de graphes sur deux molécules parmi les 8 molécules sélectionnées initialement : l'amiodarone (7µM, 24h) et l'acide valproïque (5000 µM, 24h). Nous avons choisi ces deux molécules d'une part pour la littérature scientifique les concernant et décrivant les mécanismes d'action induisant des dommages hépatiques tels que des stéatoses [208,242] et d'autre part, pour l'importante différence de taille entre leurs listes de DEGs. En effet, 5709 DEGs ont été identifiés en comparant les HPH exposés à l'acide valproïque (5000 µM, 24h) à leur condition contrôle alors que seulement 2 DEGs ont été identifiés en comparant les HPH exposés à l'amiodarone (7µM, 24h). Une telle différence dans la taille des signatures transcriptomiques permettra d'évaluer l'apport de notre approche dans deux situations opposées : (1) lorsque la liste de DEGs est trop petite pour réaliser une analyse de sur-représentation pertinente à partir des gènes ; (2) lorsque la liste de DEGs est de grande taille (plusieurs milliers de gènes), rendant l'interprétation en termes de mécanisme d'action difficile. Comme nous avons pu le mentionner dans le chapitre précédent, 57 DARs ont été identifiées pour les HPH exposés à l'amiodarone à la dose de 7µM pendant 24 heures et 413 DARs pour les HPH exposés à l'acide valproïque à la dose de 5000µM pendant 24 heures. Dans les prochaines sections nous allons

donc décrire les méthodes qui composent notre stratégie d'analyse du mécanisme d'action métabolique (mMoA) et évaluer l'apport de ces approches pour la compréhension du mMoA.

6.1. Analyse globale des DARs prédites pour les PHH exposés à l'amiodarone et l'acide valproïque

Dans un premier temps et pour nous rendre compte de la complexité que représente la compréhension ainsi que la visualisation de l'impact métabolique d'une molécule, visualisons les DARs (ainsi que le sens de la perturbation de leur activité par rapport à la condition contrôle) sur le GSMN complet Recon2.2 à l'aide de MetExploreViz. MetExploreViz est une librairie javascript dédiée à la visualisation et la manipulation des GSMNs intégrée à MetExplore. Sur la figure 48, sont représentées les DARs identifiées pour l'amiodarone (Fig 48A) et l'acide valproïque (Fig 48B) sur l'ensemble de Recon2.2. Les DARs significativement plus actives dans la condition traitée par rapport à la condition contrôle sont coloriées en rouge alors que les DARs significativement moins actives dans la condition traitée par rapport à la condition contrôle sont coloriées en vert. Pour les DARs identifiées après exposition des HPH à l'amiodarone (Fig 48A), la majorité des DARs sont suractivées (*i.e.* significativement plus actives dans la condition traitée que la condition contrôle) alors que quelques groupes de DARs de plus petite taille et plus dispersés sont déconnectées de ce groupe principal et n'interagissent pas directement avec les DARs du groupe principal. Concernant les DARs identifiées pour l'acide valproïque (Fig 48B), de nombreux groupes de petite taille sont dispersés et répartis sur l'ensemble du réseau. Il ne semble pas y avoir une région du réseau métabolique qui soit plus impactée que les autres. Une telle analyse reste très macroscopique et nous apprend uniquement que le mMoA de ces deux molécules est différent à la fois en termes de taille mais également en termes de localisation métabolique. Ce type de représentation globale a donc un intérêt pour comparer rapidement deux molécules mais ne permet pas d'analyses mécanistiques, pour lesquelles d'autres méthodes doivent être mises en place.

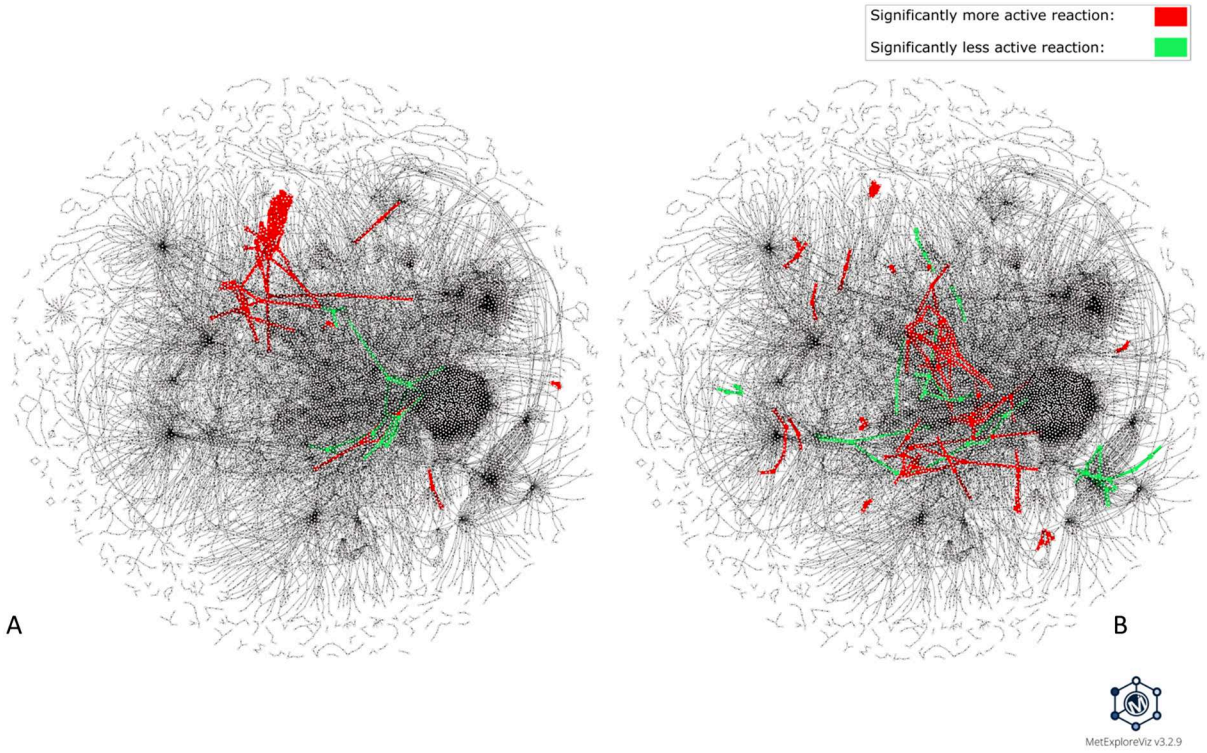

Figure 48: Visualisation par MetExploreViz de DARs prédites après exposition de HPH à l'amiodarone et l'acide valproïque sur le réseau Recon2.2, complet. *Cette visualisation a été réalisée à l'aide de MetExploreViz, en retirant les cofacteurs et molécules inorganiques (Tableau 21). Les DARs identifiées pour l'amiodarone (7µM, 24h) sont coloriées selon le sens de leur perturbation (plus active = rouge, moins active = vert) sur la Fig 48A et les DARs identifiées pour l'acide valproïque (5000µM, 24h) sont coloriées également selon leur sens de perturbation sur la Fig 48B. Les nœuds représentent les réactions et les métabolites et sont connectés si un métabolite est le substrat/produit des réactions. La structure des réseaux (positions des nœuds) ainsi que leur contenu sont identiques pour la Figure 48A et 48B, permettant la comparaison visuelle entre les deux figures.*

Dans la prochaine section nous allons donc appliquer notre approche de partitionnement de graphes et d'extraction de sous-graphes à partir des distances métaboliques avec l'objectif d'identifier et visualiser plus précisément les sous-parties perturbées du réseau métabolique et donc permettre une analyse plus fine de l'impact métabolique.

6.2. Identification de sous-réseaux de DARs à partir du partitionnement de graphes et de l'extraction de sous-graphes

Comme nous avons pu le discuter dans le cadre de l'analyse de sur-représentation, la granularité des voies métaboliques joue un rôle important sur l'analyse fonctionnelle des DEGs ou des DARs. Afin de s'affranchir de cette limite et ainsi représenter de la manière la plus précise possible le mMoA d'un composé chimique, nous allons nous appuyer sur les distances métaboliques entre les DARs. Pour rappel, notre hypothèse est que la distance entre les réactions dans le graphe métabolique peut être utilisée comme une mesure de la proximité métabolique entre les réactions. En effet, les réactions sont liées par des composés qui sont produits et consommés par d'autres réactions. Plus la chaîne entre deux réactions est courte, plus la relation métabolique est supposée être importante. En se basant sur ces distances

métaboliques, il est donc possible regrouper les réactions susceptibles d'interagir et potentiellement impliquées dans la même fonction métabolique. En appliquant l'approche de partitionnement hierarchique sur la matrice des distances métaboliques entre les DARs identifiées, après modélisation du métabolisme cellulaire de HPH exposés à chacune des 2 molécules, amiodarone et acide valproïque, nous avons identifié respectivement deux partitions (C1 et C2, pour l'amiodarone) et trois partitions (C1, C2 et C3, pour l'acide valproïque) (Fig 49).

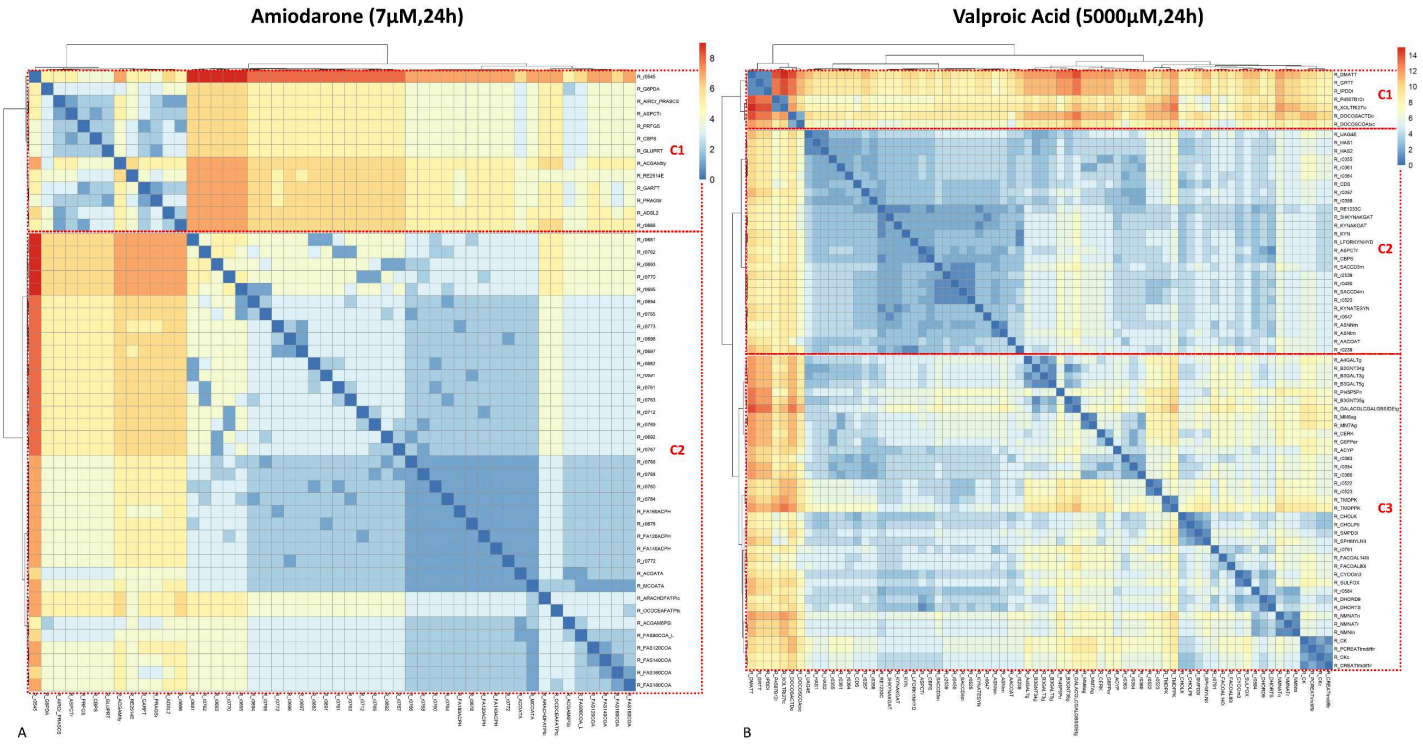

Figure 49: Heatmap avec partitionnement sur la matrice de distance métabolique entre les paires de DARs identifiées pour l'amiodarone (Fig 49A) et pour l'acide valproïque (Fig 49B). *Le partitionnement sur la matrice de distance a été réalisé via une approche de partitionnement hierarchique avec l'algorithme de Ward. Ce partionnement est visualisé sous la forme d'une heatmap construite par le package R « Pheatmap ». L'échelle de couleur à droite de chaque figure représente la distance entre deux réactions. Cette distance va de 0 (cellules colorées en bleu) à 8 (cellules colorées en rouge) pour la matrice de distance des DARs identifiées pour l'amidoarone (Fig 49A) et 14 (cellules colorées en rouge) pour la matrice de distance des DARs identifiées pour l'acide valproïque (Fig 49B). Deux partitions (C1 et C2) ont été identifiées pour l'amiodarone (Fig 49A) et trois partitions (C1, C2 et C3) pour l'acide valproïque (Fig 49B)*

Pour tirer parti de ce partitionnement et aller plus loin dans l'interprétation mécanistique des perturbations métaboliques engendrées par l'exposition à l'amiodarone et à l'acide valproïque, nous avons réalisé l'extraction d'un sous-réseau correspondant à l'arbre couvrant de poids minimal (se référer à la section, 1.5 de ce chapitre) pour chaque partition identifiée précédemment. L'extraction de ces sous-réseaux permet notamment de visualiser les DARs de chaque partition, la manière dont elles sont interconnectées et les réactions ayant été ajoutées pour la connectivité du sous-réseau par l'algorithme d'extraction basé sur les arbres de Steiner. En raison du temps nécessaire pour analyser en détail le mMoA de chacune des partitions, nous avons choisi de prioriser l'analyse de certaines partitions par rapport à d'autres. Cette

priorisation a été réalisée en calculant la proprotion de DARs dans un sous-graphe minimal et en considérant que plus cette proportion est importante, plus la partition est pertinente d'un point de vue mécanistique car indicatif de réactions proches avec une forte relation métabolique. Cette métrique peut être définie de la manière suivante :

$$DARs_{subnet\_coverage} = \frac{nDARs_{subnet}}{nTotalReactions_{subne}} * 100$$

En nous permettant d'estimer la proportion de DARs identifiées dans un sous-réseau, cette métrique nous indique les partitions dans lesquelles les DARs identifiées sont proches et en interactions.

Concernant la condition HPH exposée à 5000µM d'acide valproïque pendant 24 heures, nous avons extrait un sous-réseau contenant 77% de DARs à partir de la partition C2, ce qui signifie que la majorité des réactions de ce sous-réseau sont des réactions perturbées par l'exposition à l'acide valproïque, proches dans le réseau métabolique et interagissant les unes avec les autres. Parmi les DARs de ce sous réseau, 21 étaient suractivées et 4 étaient sous-activées (Fig 50A). 5 DARs prédites comme suractivées sont associées avec le métabolisme des -oses tels que le métabolisme du fructose, du mannose ou encore la voie de glycolyse et de gluconeogénèse. Ces DARs sont notamment impliquées dans la phosphorylation des hexoses (Fig 50B). Un autre groupe de DARs identifiées comme suractivées est associé avec le métabolisme de la lysine dans la mitochondrie et particulièrement la dégradation de la lysine via la production de L-saccahropinate à partir de 2-oxoglutarate. Une DAR associée avec le transport de la L-asparagine dans la mitochondrie a également été prédite. La L-asparagine est utilisée par la L-asparaginase (également prédite comme DAR) pour produire du L-aspartate, transporté par un transporteur aspartate-glutamate qui n'a pas été prédit comme DAR mais qui a été ajoutée par l'algorithme d'extraction de sous-réseau car nécessaire pour la connectivité de ce dernier. La phosphatidate citidylyltransferase, faisant partie du métabolisme des glycérophospholipides était également prédite comme suractivée. Enfin, 8 DARs prédites comme suractivées sont associées au métabolisme du tryptophane et impliquent des réactions produisant ou consommant du kynurenate, du L-glutamate et du 2-oxoglutarate. Les réactions sous-activées identifiées dans ce sous-réseau sont associées au métabolisme de l'hyaluronane et du propanoate pour lequel seulement la conjugaison de l'acétoacétate et du CoEnzymeA en acétoacétate-CoA est identifiée comme perturbée par l'acide valproïque. Une partie de la voie de la synthèse des pyrimidines est également perturbée avec notamment une sous-activation de réactions associées à la production/consommation du glutamate et de l'aspartate. Il est important de noter que l'interprétation fonctionnelle des notions de suractivation et de sous-activation des réactions métaboliques est à considérer avec précaution, notamment pour les réactions réversibles pour lesquelles l'interprétation d'une

sur/sous-activation est ambigüe. 9 réactions non prédites comme DARs ont été ajoutées par l'algorithme d'extraction de sous-réseaux afin d'obtenir le réseau minimal connectant toutes les DARs de la partition étudiée. Ces réactions sont associées à des voies telles que le transport mitochondrial, la glycolyse et la gluconéogenèse, la synthèse des pyrimidines ainsi que le métabolisme de la lysine (Fig 50B).

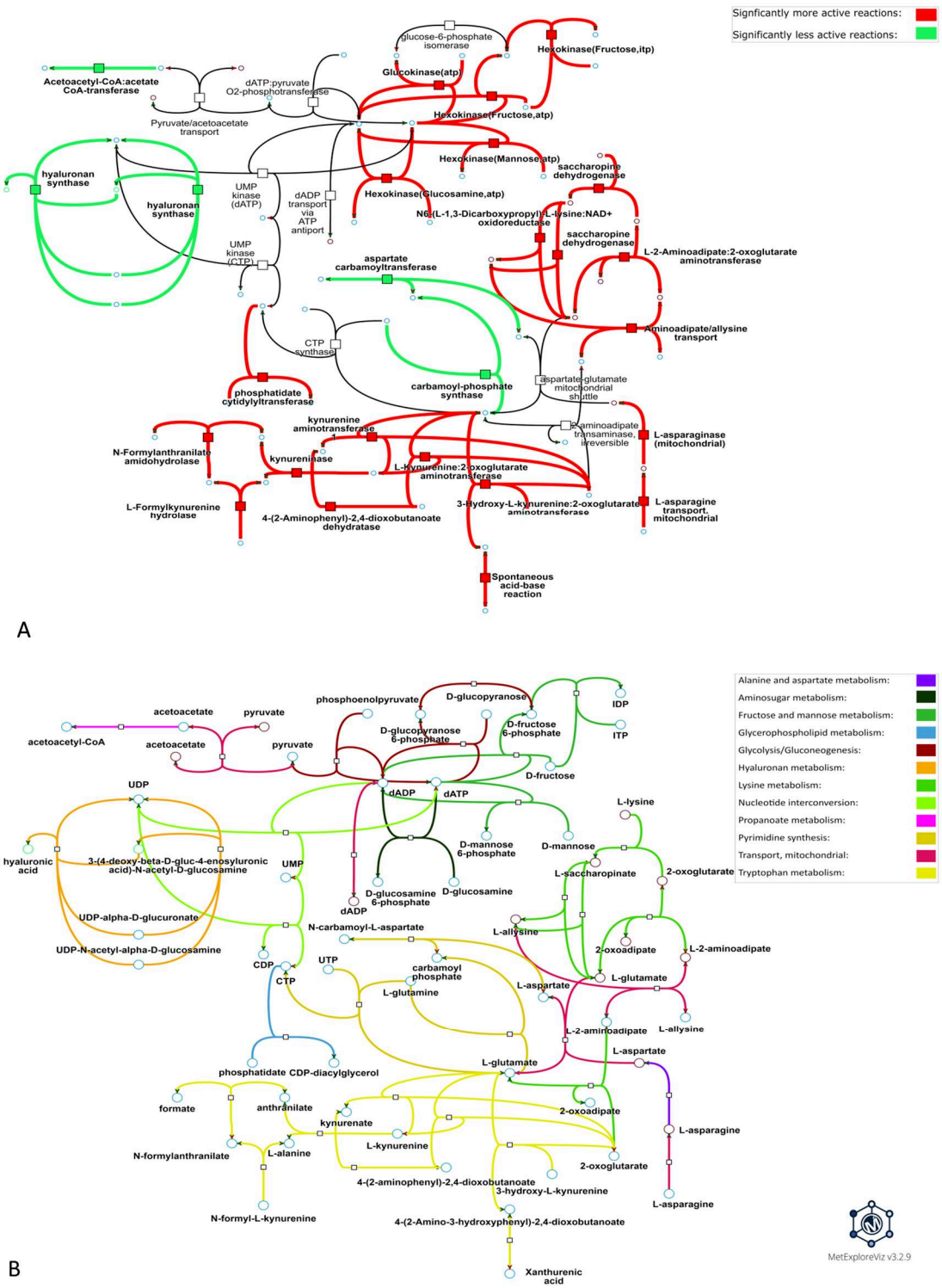

Figure 50 : Visualisation du sous-réseau métabolique minimal extrait à partir des DARs de la partition 2 (C2) prédites pour l'acide valproïque. Les DARs ont été prédites à partir des résultats de modélisation condition-spécifique avec énumération avec une version adaptée de DEXOM afin de simuler le métabolisme cellulaire de HPH exposés à 5000µM pendant 24h. Le sous-réseau visualisé sur les figures 50A et 50B correspond à la partition 2, qui est la partition à l'origine du sous-réseau ayant la plus grande proportion de DARs (77%). Les nœuds représentés par des carrés représentent des réactions et les nœuds représentés par des cercles représentent les métabolites. Sur la figure 50A, les liens représentent le sens de la perturbation : si la réaction est plus fréquemment active dans la condition traitée par rapport à la condition contrôle (suractivée), alors elle est coloriée en rouge ; à l'inverse, si elle est sous-activée, elle est coloriée en vert. Sur la figure 50B, les couleurs des liens correspondent aux voies métaboliques. Les visualisations interactives des figures 50A et 50B sont disponibles via ces liens :

https://metexplore.toulouse.inrae.fr/userFiles/metExploreViz/index.html?dir=/72ff7fdc7031b880ef4f3532134aa326/networkSaved_292937465
https://metexplore.toulouse.inrae.fr/userFiles/metExploreViz/index.html?dir=/72ff7fdc7031b880ef4f3532134aa326/networkSaved_1994092833

Concernant la condition HPH exposés à 7µM d'amiodarone pendant 24 heures, nous avons extrait un sous-réseau contenant 95% de DARs. Cela signifie que le sous-réseau contient presque exclusivement des réactions perturbées suite à l'exposition à l'amiodarone et qui interagissent entre elles car une seule réaction n'ayant pas été prédite comme DAR a été ajoutée par l'algorithme d'extraction de sous-réseau. Sur la figure 51A, nous pouvons observer que 36 DARs sont suractivées dans la condition traitée par rapport à la concentration contrôle et seulement une DAR est sous-activée dans la condition traitée par rapport à la condition contrôle. La majorité des réactions suractivées (31) du sous-réseau présenté en figure 51 sont associées avec la voie de synthèse des acides gras, 5 réactions suractivées sont associées à la voie de β-oxydation des acides gras et enfin une réaction suractivée est associée au métabolisme des sucres aminés. Lorsque l'on s'intéresse aux réactions perturbées de la voie de synthèse des acides gras, on peut constater que de nombreuses réactions sont impliquées dans la conjugaison/déconjugaison des acyl carrier proteins (ACP). Certaines réactions associées à la synthèse des Fatty-Acyl-CoA sont également perturbées avec notamment des réactions impliquées dans l'ajout de groupes malonyl-CoA à des acides gras qui sont prédites suractivées. Des perturbations de réactions associées à la voie de dégradation des acides gras (β-oxydation) sont également perturbées comme l'acetyl-CoA-ACP transacyclase, la malonyl-CoA-ACP transacyclase ainsi que deux fatty-acyl-ACP hydrolases. Enfin la seule réaction sous-activée de ce sous-réseau est une réaction associée au métabolisme des sucres aminés impliquée dans la production de N-acetylglucosamine-6-phosphate à partir d'acetyl-CoA et de D-glucosamine-6-phosphate. Seulement deux réactions ont été ajoutées par l'algorithme d'extraction de sous-graphe pour la connectivité du sous-réseau (Fig 51). Ces réactions sont responsables de l'élongation des acides gras et font partie de la voie de synthèse des acides gras.

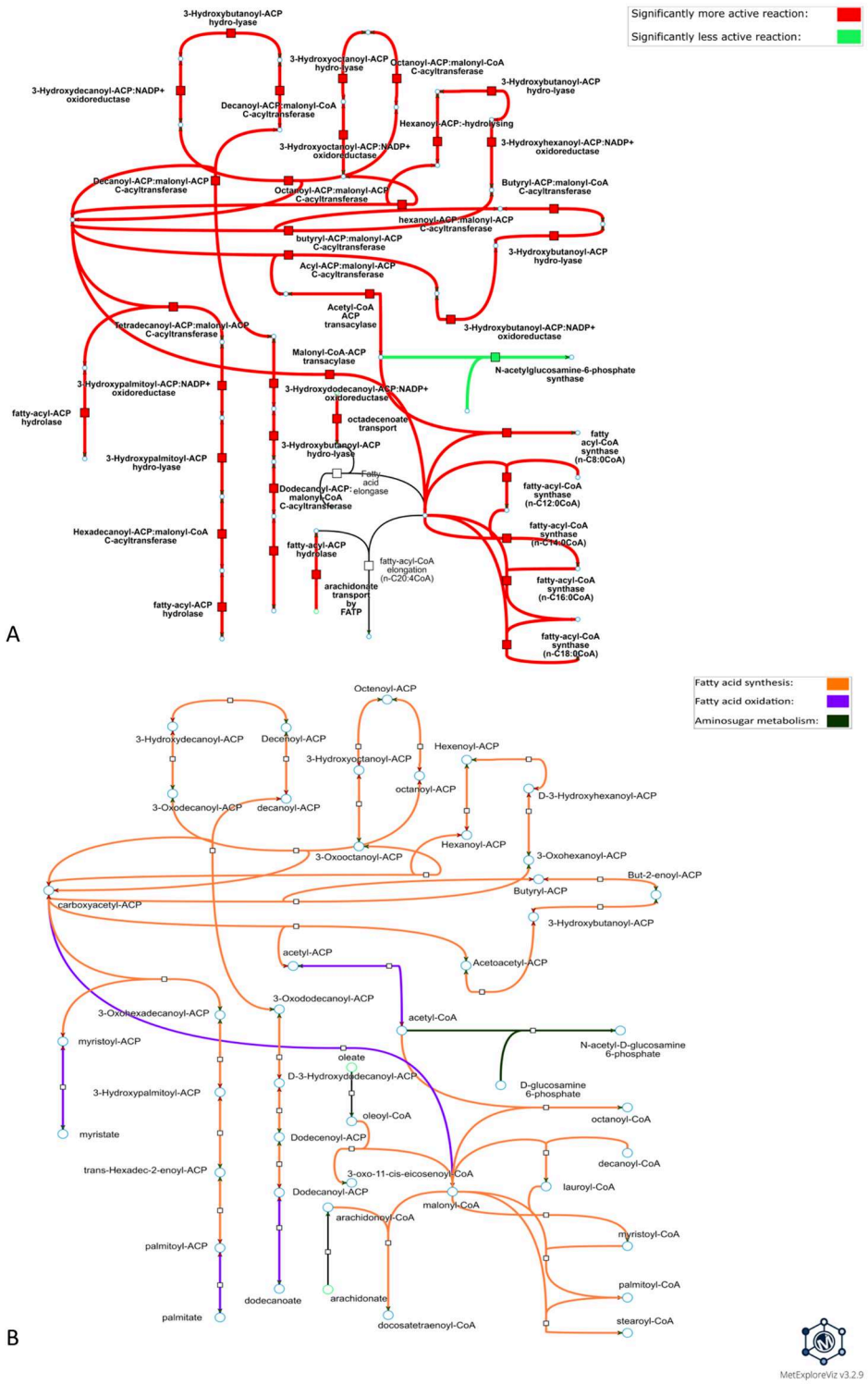

Figure 51 : Visualisation du sous-réseau métabolique minimal couvrant les DARs de la partition 2 (C2) prédites pour l'amiodarone. *Les DARs ont été prédites à partir des résultats de modélisation condition-spécifique avec énumération modélisés avec une version adaptée de DEXOM afin de simuler le métabolisme cellulaire de HPH exposés à 7µM pendant 24h. Le sous-réseau visualisé sur les figures 51A et 51B correspond à la partition 2, qui est la partition à l'origine du sous-réseau ayant la plus grande proportion de DARs (95%). Les nœuds représentés par des carrés représentent des réactions et les nœuds représentés par des cercles représentent les métabolites. Sur la figure 51, les liens représentent la direction de la perturbation.* Si la réaction est plus fréquemment active dans la condition traitée par rapport à la condition contrôle, alors elle est suractivée et est coloriée en rouge. A l'inverse, si elle est sous-activée elle est coloriée en vert. Les visualisations interactives des figures 51A et 51B sont visibles via ces liens : *https://metexplore.toulouse.inrae.fr/userFiles/metExploreViz/index.html?dir=/72ff7fdc7031b880ef4f35321344a326/networkSaved_373423088*

https://metexplore.toulouse.inrae.fr/userFiles/metExploreViz/index.html?dir=/72ff7fdc7031b880ef4f35321344a326/networkSaved_725935955

6.3. Analyse des sous-réseaux pour la compréhension du mMoA de HPH exposés à l'amiodarone ou à l'acide valproïque

L'amiodarone et l'acide valproïque sont deux molécules bien étudiées dans la littérature et connues pour induire des phénomènes hépatotoxiques par l'intermédiaire de stéatoses [208,242], ce qui est cohérent avec les perturbations métaboliques prédites par notre stratégie de modélisation et d'analyse des perturbations du métabolisme cellulaire pour l'acide valproïque (Fig 50) et l'amiodarone (Fig 51). L'acide valproïque est également connu pour son impact sur les fonctions mitochondriales [242] et son rôle d'inhibiteur d'histone de-acétylase [206], une classe de molécules connues pour impacter une grande diversité de fonctions cellulaires, ce qui est également cohérent avec l'importance et la diversité des perturbations métaboliques prédites et visualisées pour l'acide valproïque (Fig 50).

Pour rappel, nous avons sélectionné la partition correspondant au sous-réseau ayant la plus grande proportion de DARs par rapport au nombre total de réactions constituant le sous-réseau. Ce choix est principalement déterminé par le fait que certaines partitions présentent moins d'intérêt fonctionnel que d'autres (*e.g.* une partition composée de quelques réactions ou de réactions distantes les unes des autres n'ayant pu être intégrées à aucune autre partition).

L'étude du sous-réseau correspondant à la partition 2 des DARs identifiées à la suite de la modélisation de l'exposition des HPH à l'acide valproïque (5000 µM, 24h) nous a permis d'observer que cette partition était associée à 12 voies. Ces 12 voies correspondent aux 12 voies métaboliques identifiées comme significativement enrichies pour la condition acide valproïque lors de l'analyse de sur-représentation menée précédemment (cf. chapitre 3, section 2.2). Nous avons également pu identifier un autre groupe de quatre réactions suractivées qui est associé à la dégradation de la lysine dans la mitochondrie, un phénomène associé à la perturbation de l'homéostasie mitochondriale chez la souris [243]. Ce phénomène pourrait également être dû à une diminution du pool de L-carnitine associée à l'exposition à l'acide valproïque [244,245] qui pourrait déclencher un mécanisme de compensation pour restaurer la L-carnitine qui nécessite la lysine comme substrat. Enfin, un groupe de huit réactions ayant été prédit comme suractivées est associé au métabolisme du tryptophane. Il est intéressant de noter qu'une augmentation du métabolisme du tryptophane et de la kynurénine a déjà été signalée comme un effet potentiel de l'acide valproïque chez les rats [246], et que l'augmentation de la conversion du tryptophane en nicotinamide induite par l'acide valproïque a été observée chez les rats [247].

L'étude du sous-réseau correspondant à la partition 2 regroupant la majorité des DARs prédites pour l'amiodarone a révélé que la majorité des DARs étaient associées à la voie de synthèse des acides gras. Une augmentation de la lipogenèse *de novo* a été observée sur des cultures cellulaires de cellules HepaRG [211]. Cette augmentation pourrait être due à l'activation de SREBP1, un facteur de transcription responsable de la régulation de la lipogenèse *de novo*. L'un de ces gènes, FASN, code pour la fatty-acid synthase, une enzyme catalysant plusieurs réactions prédites comme DARs après comparaison des solutions de la condition traitée aux solutions de la condition contrôle (Fig 51A). Il est intéressant de noter que les gènes SREBP1 et FASN ne font pas partie des DEGs de la signature transcriptomique pour l'amiodarone, suggérant que notre stratégie de modélisation est capable de prédire une partie du mécanisme d'action publié pour l'amiodarone même lorsque les données transcriptomiques disponibles ne permettent pas d'identifier de mécanisme d'action comme c'est le cas ici pour l'amiodarone. Une augmentation de la lipogenèse *de novo* a également été décrite sur une lignée adipocytaire 3T3L1 [209] et est associée à une augmentation de la production de palmitate que l'on identifie indirectement par le biais de la prédiction de la fatty-acyl-ACP hydrolase responsable de l'hydrolyse du palmitoyl-ACP en palmitate comme suractivée (Fig 51). Il est cependant important de noter que cette réaction étant réversible, l'inverse (conjugaison du palmitate et de l'ACP en palmitoyl-ACP) est également possible. Dans le sous-réseau extrait pour la partition 2, seulement deux réactions ont été ajoutées par l'algorithme d'extraction de sous-réseau, suggérant que les DARs identifiées pour ce sous-réseau sont très proches dans le réseau métabolique et donc fonctionnellement liées. Ce sous-réseau correspond également à 74% des DARs prédites pour l'amiodarone. Etant donné que la majorité des DARs du sous-réseau (Fig 51) sont associées avec la voie de synthèse des acides gras, les prédictions provenant de notre stratégie suggèrent que le mMoA de l'amiodarone sur des HPH est principalement associé à une perturbation de cette voie métabolique et plutôt localisé dans le réseau métabolique humain. Il est important de noter que la caractérisation de l'impact métabolique de l'amiodarone sur les HPH n'aurait pas été possible en se basant uniquement sur l'analyse des données transcriptomiques puisque seulement deux DEGs ont été identifiés pour les HPH exposés à 7µM pendant 24h.

Ces résultats montrent d'une part que l'intégration de données omiques aux GSMNs via la modélisation sous-contraintes avec énumération de solutions alternatives permet d'enrichir l'information contenue dans les données transcriptomiques et d'autre part que ces approches permettent de réaliser un focus sur les mMoA ce qui peut permettre de faciliter l'interprétation du mécanisme d'action de molécules impactant un grand nombre de fonctions cellulaires tel que l'acide valproïque.

7. Conclusion

Au cours de ce chapitre, nous avons pu aborder l'apport que représentent les approches de théorie des graphes pour analyser de manière plus détaillée l'impact métabolique d'une exposition à un xénobiotique et progresser vers une meilleure compréhension de leur mMoA, à partir des modèles de l'état métabolique obtenus par modélisation sous-contraintes, comme décrite dans les chapitres précédents. Les graphes métaboliques présentent des particularités topologiques qui ont nécessité le développement de méthodes spécifiques. Tout d'abord, comme nous l'avons détaillé, dans un graphe métabolique des composés, les nœuds de plus haut degré ne sont pas forcément les nœuds les plus importants d'un point de vue biologique dans la représentation des interactions entre réactions et métabolites, mais correspondent souvent plutôt à des cofacteurs. Bien qu'essentiels pour le fonctionnement du métabolisme cellulaire, ces cofacteurs ne sont pas spécifiques et interviennent dans de nombreuses réactions métaboliques : d'un point de vue des graphes, ils connectent donc de nombreuses réactions qui peuvent être métaboliquement très éloignées, agissant ainsi comme des raccourcis dans la topologie du graphe métabolique, ce qui rend l'utilisation d'approches de théorie des graphes peu pertinentes en l'état. Nous avons donc décidé de retirer ces cofacteurs, à l'aide d'une liste définie manuellement. Nous avons également proposé une méthode permettant d'annoter, de manière automatique, dans le graphe des réactions, les arêtes connectant les réactions comme « principales » ou « secondaires », selon si elles contiennent ou non des cofacteurs. Cette méthode, qui permettrait de s'affranchir de la définition manuelle et arbitraire des cofacteurs, nécessite encore quelques développements et adaptations avant de pouvoir être appliquée de manière automatique.

Afin d'analyser l'impact métabolique de xénobiotiques et proposer des mécanismes d'actions associés nous avons établi une approche combinant calcul de distances métaboliques, partitionnement de graphes et extraction de sous-graphes. Nous avons pu montrer que la répartition des réactions identifiées comme modulées en sous-groupes, en se basant sur les distances métaboliques entre ces DARs, permet d'identifier des ensembles de réactions perturbées que l'on suppose fonctionnellement liées d'un point de vue métabolique, et donc de s'affranchir, en partie, des problèmes de définition des voies métaboliques (et des ensembles de gènes tels que Reactome). Ces groupes de réactions peuvent rassembler des réactions appartenant à différentes voies métaboliques, permettant ainsi l'étude du mMoA de manière continue plutôt que de manière isolée en voies métaboliques. Pour tirer parti de ces groupes de DARs et enrichir l'interprétation mécanistique de l'impact métabolique, nous avons ensuite chercher à extraire un sous-réseau minimal pour chacun des sous-groupes identifiés permettant d'obtenir *in fine,* pour chaque sous-groupe un petit sous-réseau connecté : ce sous-réseau présente l'avantage d'être d'une part plus facilement exploitable du fait de sa taille

réduite, et d'autre part plus facilement interprétable d'un point de vue mécanistique par rapport à une liste de réactions (ou de gènes) déconnectées. Cette stratégie a été appliquée pour étudier l'effet métabolique associé à deux des 8 molécules testées : l'amiodarone et l'acide valproïque. Pour l'amiodarone, l'analyse d'un sous-réseau minimal a permis de suggérer une possible suractivation de la lipogenèse de novo après exposition à 7µM pendant 24 heures, ce qui est en accord avec les mécanismes d'action engendrant des phénomènes hépatotoxiques et décrits dans la littérature pour ce composé. Concernant l'acide valproïque, nous avons pu identifier un sous-réseau de DARs qui sont proches en termes de distances métaboliques mais appartenant pourtant à différentes voies Cela suggère un effet relativement généralisé et peu spécifique de l'acide valproïque qui peut être expliqué par le rôle décrit dans la littérature de l'acide valproïque comme inhibiteur de l'histone de-acetylase, dont l'altération est connue pour engendrer des effets généralisés sur le fonctionnement cellulaire.

Bien que les sous-groupes de DARs identifiés pour l'amiodarone et l'acide valproïque soient cohérents car regroupant des réactions proches les unes des autres et pour lesquelles l'identification d'un sous-réseau connecté ne nécessite que peu d'ajout de réactions non-DAR, nous avons identifié deux limites à ce type d'approche. La première limite est liée au fait que l'algorithme de partitionnement (clustering) hiérarchique utilise la distance entre les réactions pour séparer les réactions dans les différents sous-groupes. Par exemple, dans le cas où deux réactions situées chacune à l'extrémité d'une cascade de réactions (*i.e.* un enchainement linaire de réactions réalisant une fonction métabolique particulière) il est probable que ces deux réactions soient considérées comme distantes et classées dans deux sous-groupes différents, suggérant qu'elles ne sont donc pas directement fonctionnellement liées alors que selon la topologie du réseau ces réactions sont effectivement fonctionnellement liées. Dans le cas de cet exemple simplifié de chaine linéaire, ces réactions sont effectivement interdépendantes puisque sans les métabolites produits par la réaction au début de la chaine, la réaction à l'autre extrémité ne peut l'être. La seconde limite de cette approche de répartition des DARs en sous-groupes réside dans le choix du nombre de sous-groupes. Cette limite est d'une part liée au choix d'utiliser une méthode de clustering hiérarchique qui nécessite (dans l'implémentation que nous avons utilisée) de déterminer le nombre de partitions de manière subjective mais correspond également à la nécessité d'étudier la liste initiale de DARs en sous-groupes à cause d'une part de sa taille et d'autre part de la diversité de zone métaboliques impactées dans le réseau. En effet, la visualisation des sous-réseaux permet une analyse très précise des perturbations métaboliques mais cette analyse ne peut être réalisée que si le sous-réseau est lisible et donc de taille restreinte (environ une centaine de réactions au maximum). Certaines de ces limites pourraient peut-être être résolues grâce à des approches de détection de communautés telles que l'algorithme de Leiden [144] mais cela impliquerait de perdre le contrôle sur la taille maximale des partitions (avec l'algorithme de Leiden) et donc de

complexifier l'analyse et la visualisation détaillée du mMoA. De plus ces approches de détection de communauté sont sensibles aux propriétés topologiques propres aux réseaux de grande tailles (tels que les réseaux sociaux) [248] et donc peut-être, dans une moindre mesure, aux réseaux tels que les GSMNs. Cela nécessiterait donc une évaluation détaillée des performances des différents algorithmes de détection de communauté disponibles et applicables sur les GSMNs.

Enfin, comme nous avons pu l'observer avec les cas d'étude de l'amiodarone et de l'acide valproïque, les perturbations du métabolisme peuvent impacter un large panel de voies métaboliques qui sont plus interconnectées que leur séparation en voies métaboliques ne le laisse penser. L'utilisation de méthodes de clustering et d'extraction de sous-réseaux basées sur le calcul de distances métaboliques a permis d'identifier des partitions ainsi que des sous-réseaux de DARs permettant de prendre en compte cette « multi-localisation » des impacts métaboliques et d'analyser en détail le rôle et les connexions des réactions perturbées. L'utilisation d'un algorithme d'extraction basé sur la recherche de l'arbre couvrant de poids minimal nous a également permis d'apporter du contexte entre les DARs. Par exemple, dans le sous réseau présenté en figure 50, la majorité des réactions ajoutées par l'algorithme d'extraction de sous-réseau sont des réactions de transport permettant de connecter des DARs situées dans différents organites et ainsi d'analyser le mMoA d'une molécule comme un phénomène continu impactant différents compartiments cellulaires et voies métaboliques.

Notons également que cette stratégie d'interprétation des DARs basée sur des approches de théorie des graphes a été mise au point pour permettre une analyse détaillée de l'impact métabolique représentée par les DARs qui sont le résultat de la première partie de notre stratégie reposant sur l'exploitation des dizaines de milliers de solutions alternatives issues de la modélisation condition spécifique avec énumération. Cependant, elle peut être facilement adaptée pour fonctionner avec d'autres approches capables de déterminer une liste de réactions différentiellement activées entre deux conditions (*e.g.* MOOMIN[249]). Fort heureusement, les limites identifiées pour cette approche ne sont pas insolubles et sont au contraire source de perspectives et de futurs travaux comme nous allons pouvoir l'aborder au cours du dernier chapitre de ce manuscrit.

Chapitre 5 : Conclusion et Perspectives

1. Conclusion

Les travaux présentés au cours de cette thèse proposent une stratégie permettant de modéliser, visualiser et interpréter l'impact métabolique d'un xénobiotique grâce à l'intégration de données transcriptomiques au réseau métabolique humain. Cette stratégie rentre donc dans le cadre de l'objectif initial de ce projet qui était de développer de nouvelles approches de biologie systémique pour l'évaluation de la toxicité de matières premières cosmétiques. En s'appuyant à la fois sur l'état de l'art et en développant de nouvelles approches pour résoudre les nombreux défis rencontrés au fil du développement de cette stratégie, nous avons proposé un ensemble d'approches permettant de (1) Intégrer des données transcriptomiques sous la forme de contraintes à un problème de modélisation sous-contraintes du métabolisme cellulaire via une version adaptée de DEXOM ; (2) Proposer une nouvelle approche d'exploitation des résultats de modélisation sous-contraintes avec énumération partielle des solutions alternatives permettant d'identifier des réactions différentiellement activées entre deux conditions tout en limitant l'impact du bruit basal : (3) Proposer une nouvelle approche d'interprétation de ces listes de réactions qui s'appuie sur la notion de proximité dans le réseau métabolique, une méthode de partitionnement et enfin l'extraction de sous-graphes permettant la visualisation et une meilleure compréhension du potentiel mMoA prédit par l'approche de modélisation.

Afin de disposer d'une grande quantité de données pour le développement de cette stratégie, nous nous sommes appuyés sur une base de données publiques, Open TG-GATEs. L'exploration de cette base de données de transcriptomique nous a notamment permis de mettre en évidence l'importance d'une compréhension fine des conditions expérimentales dans lesquelles la base de données a été générée. Nous avons notamment identifié les métadonnées manquantes (lot d'hépatocytes et classification du degré d'hépatotoxicité) et mené les recherches bibliographiques permettant d'obtenir ces métadonnées et de déployer les méthodes de normalisation et de correction adaptées.
Bien que cette phase d'exploration n'ait pas nécessité de nouveaux développements elle a favorisé certains choix méthodologiques tels que le traitement des effets lots de cellule ou encore le choix de méthodes de comparaison entre conditions plutôt que de méthodes traitant l'ensemble de la base (apprentissage machine par exemple). Cette phase d'exploration et d'enrichissement des métadonnées a également été importante pour le choix des xénobiotiques à étudier ainsi que les interprétations réalisées lors des dernières étapes de notre stratégie.

Après cette première phase de traitement des données, nous nous sommes attachés à représenter l'état métabolique d'une cellule hépatique dans une condition donnée (exposée à un xénobiotique, à un temps donné et une dose donnée, ou non exposée) de la manière la plus exhaustive possible. Des méthodes « non-biaisées » de modélisation sous-contraintes avec

exploration de l'espace de solutions existent dans la littérature. Cependant, pour utiliser ce type d'approches dans notre stratégie, nous avons dû répondre à deux défis principaux : (1) le temps de calcul ; (2) l'exploitation des résultats. Ce premier défi a été résolu en adaptant la phase d'énumération de solutions alternatives de DEXOM, permettant ainsi d'utiliser cette méthode pour modéliser de la manière la plus exhaustive possible l'état du métabolisme cellulaire dans plusieurs dizaines de conditions différentes avec un coût computationnel par condition modéré (environ 24h par condition sur un ordinateur classique, quelques heures sur un cluster de calcul). Le second défi a pu être résolu grâce au développement d'une approche permettant d'identifier des réactions différentiellement activées, nommées DARs, entre deux conditions (*e.g.* traitement *vs* contrôle). L'utilisation de méthodes statistiques classiquement utilisées (Test exact de Fisher, Odds-Ratio) dans la littérature n'étant pas possible à cause du très grand nombre d'échantillons (ici les solutions alternatives de chaque condition), nous avons proposé une nouvelle métrique basée sur le calcul de fréquences d'activations permettant d'identifier des réactions significativement plus ou moins fréquemment actives dans une condition par rapport à une autre. Soucieux de la robustesse des prédictions réalisées par notre stratégie et conscient du manque de contraintes omiques sur certaines réactions du réseau métabolique, nous avons développé une approche de calcul du bruit basal propre à une réaction afin de renforcer notre confiance dans le fait que les réactions prédites comme perturbées sont effectivement représentatives de l'impact métabolique du xénobiotique étudié.

Enfin, nous avons proposé une approche originale pour l'interprétation des listes de DARs prédites par les premières étapes de notre stratégie. Cette approche se différencie des approches classiques de sur-représentation dans les voies métaboliques en utilisant la notion de proximité dans le réseau métabolique afin d'identifier des groupes de DARs proches dans le réseau et donc que l'on suppose liées par une relation métabolique. Pour aller plus loin dans l'interprétation du ou des mécanismes d'actions métaboliques proposés par notre stratégie, nous avons développé une approche d'extraction de sous-graphes pour chaque groupe de DAR identifié mettant ainsi en évidence les liens métaboliques entre ces réactions. Nous avons pu montrer via une application que notre stratégie était capable de prédire des mécanismes d'action publiés pour deux molécules, l'amiodarone et l'acide valproïque.

Bien que cela n'a pas été directement abordé au cours de ce manuscrit de thèse, cette stratégie a été automatisée afin de simplifier l'utilisation de l'ensemble des méthodes implémentées et discutées. Le choix des méthodes a également été réalisé dans un objectif de généralisation et de modularité de la stratégie de modélisation et d'interprétation de l'impact métabolique développée au cours de ces travaux.

En combinant une approche de modélisation sous-contraintes avec énumération partielle de solutions avec des approches de graphes notre stratégie se veut originale et permet de progresser vers l'identification du mMoA de xénobiotiques. L'ensemble de ces méthodes se veut complémentaire mais également modulaire puisque chacune des 3 grandes étapes (intégration des données transcriptomiques, modélisation et analyse basée sur les graphes) présentée dans ce manuscrit peut être combinée avec d'autres approches existantes ou futures. Cette modularité sera également clé pour mettre à jour certaines étapes de la stratégie afin d'améliorer certains points sensibles tels que l'intégration des données ou le partitionnement des DARs.

Nous avons donc pu répondre à plusieurs défis permettant ainsi d'améliorer l'étude et la compréhension du mécanisme d'action métabolique des xénobiotiques via la modélisation du métabolisme cellulaire. Nous allons à présent aborder quelques perspectives sur lesquelles il serait intéressant de travailler afin de compléter la stratégie développer au cours de ces travaux.

2. Perspectives

2.1. Extension à d'autres types de données omiques

Comme nous avons pu le mentionner précédemment, l'une des limites de la modélisation sous-contraintes est le manque de contraintes biologiques. Moins le problème est contraint par des données expérimentales et plus l'espace de solutions possibles risque d'être important. Il est possible de réduire cet espace de solutions alternatives en apportant des informations expérimentales complémentaires. Dans le cadre de ces travaux, nous avons utilisé uniquement des données transcriptomiques. Il serait intéressant d'évaluer l'apport que pourrait représenter la combinaison de contraintes provenant de données transcriptomiques avec des contraintes provenant de données d'exométabolomique. L'exometabolomique consiste à mesurer la concentration de métabolites dans le milieu extracellulaire entre deux conditions et ainsi de permettre de contraindre les flux des réactions produisant et consommant ces métabolites (qui correspondent aux réactions d'échanges danas le modèle du réseau métabolique).

Il faudrait pour cela générer des données transcriptomiques ainsi que des données d'exométabolomique à partir des mêmes échantillons biologiques exposés et non exposés à un composé potentiellement hépatotoxique. L'intégration simultanée de ces données omiques serait réalisée par l'ajout de nouvelles contraintes lors de la création du MILP permettant de prendre en compte les données d'exométabolomique. Nous pourrions pour cela nous appuyer sur des travaux ayant modifié iMAT [99,100] afin de pouvoir intégrer des données de métabolomique lors de la modélisation [101].

En suivant la même approche que [101] et en l'adaptant à DEXOM, il serait possible d'intégrer différents types de données omiques tant que ces données sont générées de manière simultanée pour les mêmes conditions biologiques.

2.2. Apprentissage machine et apprentissage profond sur les graphes métaboliques

L'un des objectifs initiaux de la thèse était de développer un modèle prédictif de l'hépatotoxicité à partir des résultats de modélisation condition-spécifique avec énumération. Plus précisément, il s'agissait de développer, par des méthodes d'apprentissage, un modèle permettant de prédire le caractère hépatotoxique ou non d'un composé chimique, à partir du réseau métabolique reconstruit caractérisant l'exposition à ce composé. Cependant, plusieurs défis ont émergé au fil de l'avancée du projet. Le premier défi a été le déséquilibre de classe (le ratio molécule hépatotoxique/molécule non hépatotoxique) existant dans la base de données Open TG-GATEs, avec beaucoup plus de données disponibles pour des molécules hépatototoxiques par rapport à des molécules non hépatotoxiques. Une autre difficulté a été que les doses utilisées dans la littérature pour établir la classification des molécules comme hépatotoxiques/non hépatotoxiques sur laquelle le modèle a été entrainé étaient souvent plus élevées que les doses (faiblement cytotoxiques) utilisées dans Open TG-GATEs et avec lesquelles avaient été générées les données de transcriptomique utilisées pour la modélisation sous-contraintes et/ou la construction du modèle. Au final, les effets expérimentaux mesurés n'étaient pas réellement représentatifs du statut d'hépatotoxicité établi sur des doses différentes.

Différentes méthodes d'apprentissage machine ont été expérimentées, certaines comme les approches de multi-kernel [250] et d'intégration de données hétérogènes [251–255] nous ont permis d'intégrer des données transcriptomiques et structurales aux résultats de modélisation condition-spécifique. Cependant, nous n'avons pas été en mesure de construire un modèle prédictif dont les performances étaient supérieures aux performances déjà publiées pour la prédiction de l'hépatotoxicité (représentée par le risque DILI) [6,256]. Comme expliqué précédemment, l'une des causes de cette faible performance des modèles construits réside probablement à la fois dans le déséquilibre des données et dans l'écart entre les données testées dans la base de données Open TG-GATEs et celles ayant permis de caractériser le niveau et le type d'hépatotoxicité dans la littérature.

Un des points d'amélioration que nous avons identifié réside dans l'utilisation des résultats de modélisation pour la prédiction de l'hépatotoxicité. Nous avons entrainé les modèles sur les vecteurs binaires d'activation/inactivation des réactions. Cependant, en réalisant l'apprentissage sur ce type de données, on ne tire pas directement partie de l'interconnexion existante entre les réactions métaboliques.

Il serait donc intéressant d'utiliser les propriétés topologiques des graphes métaboliques correspondant à chacune des solutions alternatives énumérées afin d'entrainer un modèle prédictif capable de tirer profit des propriétés topologiques des graphes métaboliques correspondant aux solutions énumérées de chacune des conditions.

L'une des approches classiquement utilisée pour prendre en compte la topologie des graphes est de calculer des caractéristiques topologiques propre à chaque graphe (*e.g.* comptage des graphlets, exploitation de la matrice d'adjacence, etc) et d'entrainer le modèle à identifier les molécules hépatotoxiques des molécules non hépatotoxiques sur la base de ces caractéristiques topologiques. Il existe un grand nombre de propriétés topologiques calculables à partir d'un graphe (Fig 52). Une des étapes importantes sera de sélectionner un ensemble de propriétés topologiques qui soient complémentaires et dont le temps de calcul n'est pas trop important. En effet, du fait de l'énumération partielle, le nombre de graphes métaboliques pour lesquels il faudra calculer les propriétés topologiques risque d'être important.

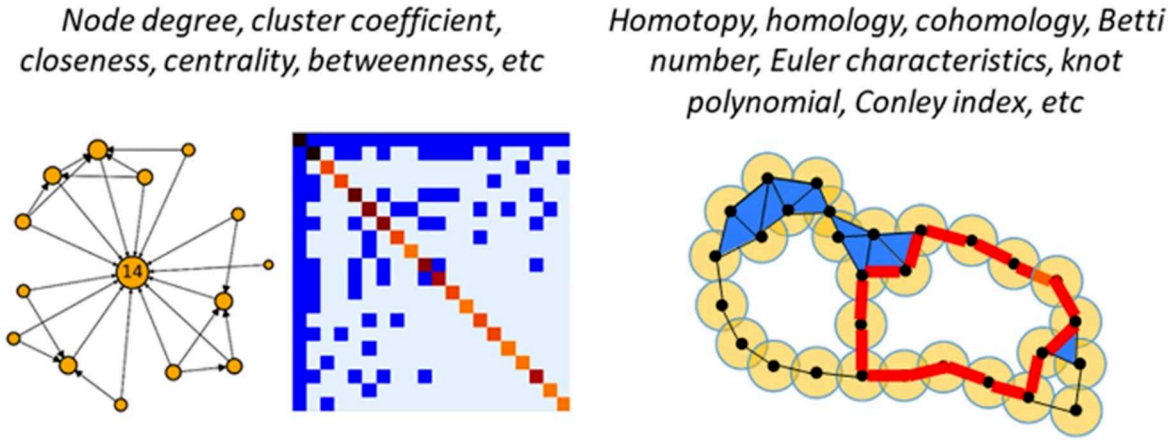

Figure 52 : Exemple de propriétés topologiques calculables à partir de graphes. *Figure adaptée de* [257]

Enfin une autre approche qui pourrait également s'avérer intéressante serait de faire appel à des méthodes d'apprentissage profond capables d'apprendre directement à partir de données structurées sous la forme de graphes. Ce type d'approche est connue sous le nom de « Graph Neural Networks » (GNN). Il existe de nombreuses variantes des GNNs, les plus connues sont les Graph Convulational Networks (GCN), les Graph Attention Network (GAT) et les Graph Recurrent Networks (GRN) [258].

En utilisant l'approche de GNN adaptée, l'objectif serait donc de classifier les graphes correspondant aux solutions alternatives en deux catégories : hépatotoxiques et non hépatotoxiques. Des travaux utilisant les GNN pour prédire des marqueurs de toxicité à partir de données structurales ont déjà été publiés [259,260]. Etant donné que la tache finale (prédire des phénomènes toxiques) est similaire à celle que nous souhaitons réaliser, il serait intéressant d'étudier la « Loss function » (la fonction permettant de faire converger le modèle, vers un modèle prédictif) qui a été utilisée et ainsi s'appuyer sur leurs travaux pour mettre au

point un modèle GNN adapté à la prédiction de l'hépatotoxicité à partir des données de modélisation condition-spécifique avec énumération.

Les GNNs pourraient également nous permettre d'estimer la similarité entre des ensembles de graphes métaboliques provenant de conditions (cellules exposées à des xénobiotiques) différentes (Fig 53) et ainsi permettre l'extension de la stratégie présentée au cours de ces travaux à des applications telles que le Read-Across biologique.

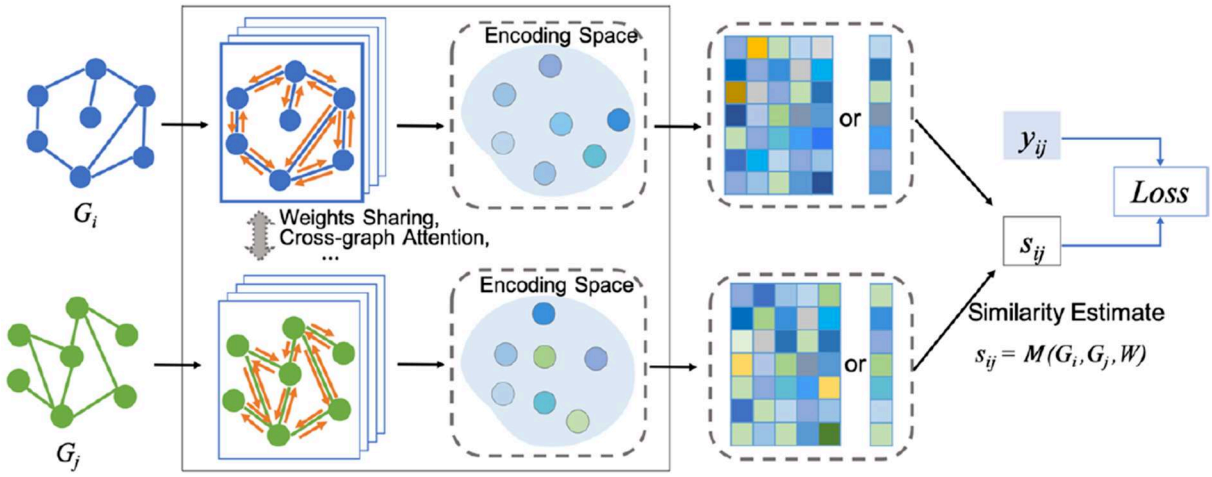

Figure 53 : Schéma de l'architecture d'un protocole d'estimation de la similarité entre des graphs basé sur des GNNs. *Figure provenant de* [261]

Il est cependant intéressant de garder à l'esprit qu'il sera important de prendre en compte tous les facteurs confondants possibles afin de ne pas biaiser l'apprentissage du modèle. En effet, ce type d'approche à vocation à être entrainé sur de grandes quantités de données donc le risque de facteurs confondants (effet lot de cellule, effet liés à la combinaison d'études ou de bases différentes) tels que ceux décrit pour la base de données Open TG-GATEs est important et devra être pris en compte.

2.3. Comparaison d'empreintes métaboliques

Au cours de ces travaux de thèse, nous avons eu l'opportunité de faire des points réguliers avec les utilisateurs potentiels de la stratégie développée. Grâce à ces points réguliers, nous nous sommes aperçus que bien que la visualisation des sous-graphes minimaux soit pertinente pour étudier plus en détail les potentiels mMoA des conditions modélisées, ce type de représentation nécessite tout de même une certaine expertise biochimique et une étude bibliographique importante afin de pouvoir identifier les mécanismes d'actions et les potentiels effets hépatotoxiques associés aux réactions prédites comme perturbées par notre stratégie. Il serait donc intéressant de pouvoir proposer une vision plus macroscopique de la perturbation métabolique induite par l'exposition à un xénobiotique, que l'on pourrait considérer comme une visualisation de l'empreinte métabolique du composé. L'objectif de ces empreintes métaboliques serait de permettre l'identification rapide des zones perturbées mais également de pouvoir comparer visuellement différentes empreintes métaboliques. Afin réaliser cette

comparaison visuelle, il est nécessaire d'avoir une carte métabolique qui soit lisible mais également fixe (les nœuds ne changent pas de place d'une visualisation à une autre).

Ce type de représentation existe pour les réseaux Recon2 [72] et Recon3D [217], il s'agit de ReconMap [262]. Ces cartes permettent une visualisation simplifiée du métabolisme humain avec une organisation similaire aux cartes disponibles dans la base de données KEGG. Il est également possible de modifier la couleur des nœuds correspondant à des réactions ou des métabolites d'intérêt afin de tendre vers la vision plus macroscopique que nous cherchons à obtenir. Cependant, dans une carte KEGG ou ReconMap, les nœuds ne sont pas placés par proximité dans le réseau métabolique, ce qui peut induire en erreur l'utilisateur lors de l'interprétation de ces cartes. A noter que toutes les connexions ne sont pas représentées afin de faciliter la visualisation.

L'objectif serait donc dans un premier temps de proposer une méthode de visualisation simplifiée du réseau métabolique mais n'induisant pas de biais de représentation des distances métaboliques. Nous avons eu l'occasion de débuter la réflexion à ce sujet et avons proposé une méthode permettant de représenter un GSMN sous la forme d'une matrice. Le nombre de cases dans cette matrice permet de déterminer le nombre de partitions (constitués de métabolites, réactions ou la combinaison des deux) identifié par l'approche PAM (Partition Around Medoids). Cette méthode permet de partitionner l'ensemble des réactions à partir de la matrice de distance correspondant au graphe métabolique construit à partir du GSMN. Pour rappel, cette matrice contient les distances métaboliques entre toutes les paires de réactions du réseau, calculées par la méthode du « ManytoMany Shortest Path » dans le graphe métabolique. Il est possible de réaliser ce partitionnement sur un graphe des réactions, un graphe des composés mais également un graphe bipartite. Les cases de la matrice sont ordonnées selon le nombre d'arêtes qu'elles partagent via un clustering hiérarchique, ce qui signifie que les cases proches les unes des autres dans la grille (Fig 54) contiennent des réactions (et/ou métabolites selon le type de graphe représenté) connectées par des arêtes et donc proches dans le réseau.

Utiliser une partition par médoïdes au lieu d'une partition par centroïdes permet d'identifier, à l'intérieur de chaque partition, un élément central, qui peut être utilisé pour proposer une première annotation du contenu de chaque case de la matrice (Fig 54).

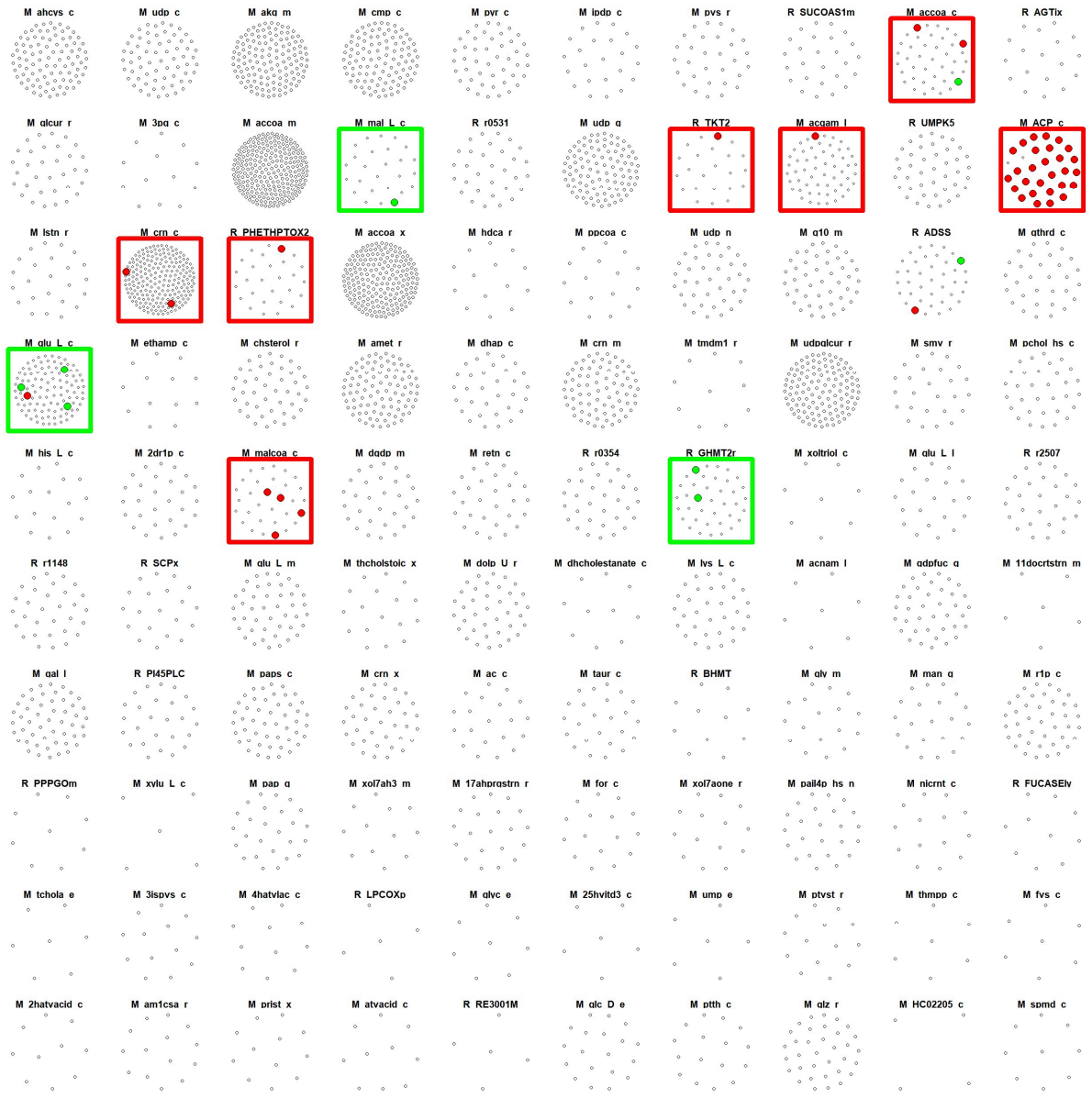

Figure 54: Exemple de la représentation sous forme de grille du réseau Recon2.2 et des réactions perturbées suite à une exposition à 7µM d'amiodarone pendant 24 heures. *Le contenu de chaque case correspond à une partition calculée par la méthode PAM. L'ordre des cases dans la grille est défini à partir d'un clustering hiérarchique sur la matrice des arêtes partagées entre les cases de la grille*

Cette première ébauche de visualisation peut être encore améliorée. Il serait notamment intéressant de proposer une annotation automatique des cases de la matrice afin de pouvoir identifier les fonctions métaboliques principales impactées par un composé. On pourrait par exemple chercher à identifier les fonctions enzymatiques principales de chaque case (via les EC numbers) ou encore réaliser une analyse de sur-représentation à partir des réactions de chacune des cases. Pour faciliter la comparaison de ces empreintes métaboliques, il serait intéressant de rendre ces visualisations intéractives et plus ergonomiques. Afin de répondre à la problématique de comparaison d'empreintes, le développement d'approches permettant de visualiser l'intersection de cases perturbées entre deux conditions ou au contraire les cases n'étant perturbées que dans l'une des deux conditions serait pertinent.

La comparaison visuelle d'empreintes métaboliques bénéficierait également d'une comparaison moins subjective, notamment par le biais de métriques qui permettraient de guider cette comparaison visuelle.

Bibliographie

1. 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. 2018;46:W495–502.

2. Eupati - Toxicité systémique. https://toolbox.eupati.eu/glossary/toxicite-systemique/?lang=fr#:~:text=Cela%20fait%20r%C3%A9f%C3%A9rence%20%C3%A0%20des,de%20son%20point%20d'entr%C3%A9e. 2023.

3. Mellor CL, Marchese Robinson RL, Benigni R, Ebbrell D, Enoch SJ, Firman JW, et al. Molecular fingerprint-derived similarity measures for toxicological read-across: Recommendations for optimal use. Regulatory Toxicology and Pharmacology. 2019;101:121–34.

4. Su R, Wu H, Liu X, Wei L. Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies. Brief Bioinform. 2019;

5. De Abrew KN, Overmann GJ, Adams RL, Tiesman JP, Dunavent J, Shan YK, et al. A novel transcriptomics based in vitro method to compare and predict hepatotoxicity based on mode of action. Toxicology. 2015;328:29–39.

6. Li T, Tong W, Roberts R, Liu Z, Thakkar S. Deep Learning on High-Throughput Transcriptomics to Predict Drug-Induced Liver Injury [Internet]. Frontiers in Bioengineering and Biotechnology . 2020. p. 1366. Available from: https://www.frontiersin.org/article/10.3389/fbioe.2020.562677

7. Guijas C, Montenegro-Burke JR, Warth B, Spilker ME, Siuzdak G. Metabolomics activity screening for identifying metabolites that modulate phenotype. Nat Biotechnol. Nature Publishing Group; 2018. p. 316–20.

8. Metallo CM, Vander Heiden MG. Understanding Metabolic Regulation and Its Influence on Cell Physiology. Mol Cell. 2013. p. 388–98.

9. Scott C. Misconceptions about Aerobic and Anaerobic Energy Expenditure [Internet]. Journal of the International Society of Sports Nutrition©. A National Library of Congress Indexed Journal. ISSN. 2005. Available from: www.sportsnutritionsociety.org

10. Sever R, Glass CK. Signaling by nuclear receptors. Cold Spring Harb Perspect Biol. 2013;5.

11. Bjork JA, Butenhoff JL, Wallace KB. Multiplicity of nuclear receptor activation by PFOA and PFOS in primary human and rodent hepatocytes. Toxicology. 2011;288:8–17.

12. Kokkonen P, Beier A, Mazurenko S, Damborsky J, Bednar D, Prokop Z. Substrate inhibition by the blockage of product release and its control by tunnel engineering. RSC Chem Biol. 2021;2:645–55.

13. Schmidt ND, Peschon JJ, Segel IH. Kinetics of Enzymes Subject to Very Strong Product Inhibition: Analysis Using Simplified Integrated Rate Equations and Average Velocities. J. theor. Biol. 1983.

14. Rui L. Energy metabolism in the liver. Compr Physiol. 2014;4:177–97.

15. Soars MG, McGinnity DF, Grime K, Riley RJ. The pivotal role of hepatocytes in drug discovery. Chem Biol Interact. 2007. p. 2–15.

16. Gómez-Lechón MJ, Castell J V., Donato MT. The use of hepatocytes to investigate drug toxicity. Methods Mol Biol. 2010;640:389–415.

17. Guguen-Guillouzo C, Guillouzo A. General Review on In Vitro Hepatocyte Models and Their Applications BT - Hepatocytes: Methods and Protocols. In: Maurel P, editor. Totowa, NJ: Humana Press; 2010. p. 1–40. Available from: https://doi.org/10.1007/978-1-60761-688-7_1

18. Lee SML, Schelcher C, Laubender RP, Fröse N, Thasler RMK, Schiergens TS, et al. An algorithm that predicts the viability and the yield of human hepatocytes isolated from remnant liver pieces obtained from liver resections. PLoS One. 2014;9:e107567.

19. Ingelman-Sundberg M. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. Trends Pharmacol Sci. 2004;25:193–200.

20. Zárybnický T, Matoušková P, Lancošová B, Šubrt Z, Skálová L, Boušová I. Inter-Individual Variability in Acute Toxicity of R-Pulegone and R-Menthofuran in Human Liver Slices and Their Influence on miRNA Expression Changes in Comparison to Acetaminophen. Int J Mol Sci. 2018;19.

21. Roden DM. Mechanisms underlying variability in response to drug therapy: implications for amiodarone use. Am J Cardiol. 1999;84:29R-36R.

22. Knowles BB, Howe CC, Aden DP. Human Hepatocellular Carcinoma Cell Lines Secrete the Major Plasma Proteins and Hepatitis B Surface Antigen. Science (1979). 1980;209:497–9.

23. Lőrincz T, Deák V, Makk-Merczel K, Varga D, Hajdinák P, Szarka A. The performance of hepg2 and heparg systems through the glass of acetaminophen-induced toxicity. Life. 2021;11.

24. Jetten MJA. Toxicogenomics responses in the in vitro liver : a view on human interindividual variation [Internet]. maastricht university; 2014. Available from: https://cris.maastrichtuniversity.nl/en/publications/24e8a187-465f-492c-8f0f-f71c121a19cf

25. Jossé R, Aninat C, Glaise D, Dumont J, Fessard V, Morel F, et al. Long-term functional stability of human HepaRG hepatocytes and use for chronic toxicity and genotoxicity studies. Drug Metabolism and Disposition. 2008;36:1111–8.

26. Madorran E, Stožer A, Bevc S, Maver U. In vitro toxicity model: Upgrades to bridge the gap between preclinical and clinical research. Bosn J Basic Med Sci. 2020;20:157–68.

27. Van Norman GA. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink Our Current Approach? JACC Basic Transl Sci. 2019;4:845–54.

28. Knight A. Systematic Reviews of Animal Experiments Demonstrate Poor Human Clinical and Toxicological Utility Human Clinical and Toxicological Utility [Internet]. 2007. Available from: https://www.wellbeingintlstudiesrepository.org/acwp_arte

29. Satya P. Yadav. the wholeness in suffix -omics -omes and the word om.

30. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51:D933–41.

31. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res. 2013;41.

32. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas - A tool for pathology. Journal of Pathology. 2008. p. 387–93.

33. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: A resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020;48:D440–4.

34. Ramachandran R, Bugbee K, Murphy K. From Open Data to Open Science. Earth and Space Science. 2021;8.

35. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res [Internet]. 2014/10/13. 2015;43:D921–7. Available from: https://pubmed.ncbi.nlm.nih.gov/25313160

36. Ganter B, Snyder RD, Halbert DN, Lee MD. Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. Pharmacogenomics. 2006;7:1025–44.

37. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science (1979) [Internet]. 2006 [cited 2020 Feb 13];313:1929–35. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1132939

38. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171:1437-1452.e17.

39. Lim N, Pavlidis P. Evaluation of connectivity map shows limited reproducibility in drug repositioning. Sci Rep. 2021;11.

40. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44:D1075–9.

41. Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res. 2023;51:D1257–62.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nat Genet. 2000. p. 25–9.

43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences [Internet]. 2005;102:15545–50. Available from: https://doi.org/10.1073/pnas.0506580102

44. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool [Internet]. 2013. Available from: http://amp.pharm.mssm.edu/Enrichr.

45. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.

46. Karp PD, Midford PE, Caspi R, Khodursky A. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. BMC Genomics [Internet]. 2021;22:191. Available from: https://doi.org/10.1186/s12864-021-07502-8

47. Wieder C, Frainay C, Poupin N, Rodríguez-Mier P, Vinson F, Cooke J, et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol. 2021;17:e1009105.

48. Wassenaar PNH, Rorije E, Vijver MG, Peijnenburg WJGM. Evaluating chemical similarity as a measure to identify potential substances of very high concern. Regulatory Toxicology and Pharmacology. 2021;119.

49. Stuard SB, Heinonen T. Relevance and Application of Read-Across – Mini Review of European Consensus Platform for Alternatives and Scandinavian Society for Cell Toxicology 2017 Workshop Session. Basic Clin Pharmacol Toxicol. Blackwell Publishing Ltd; 2018. p. 37–41.

50. Smith SW. Chiral toxicology: It's the same thing only different. Toxicological Sciences. 2009. p. 4–30.

51. Rehman W, Arfons LM, Lazarus HM. The rise, fall and subsequent triumph of thalidomide: Lessons learned in drug development. Ther Adv Hematol. 2011. p. 291–308.

52. Zhao L, Russo DP, Wang W, Aleksunes LM, Zhu H. Mechanism-Driven Read-Across of Chemical Hepatotoxicants Based on Chemical Structures and Biological Data. Toxicol Sci. 2020;174:178–88.

53. Zhu H, Bouhifd M, Donley E, Egnash L, Kleinstreuer N, Kroese ED, et al. Supporting read-across using biological data. ALTEX. 2016;33:167–82.

54. Jacques C, Jamin EL, Jouanin I, Canlet C, Tremblay-Franco M, Martin J-F, et al. Safety assessment of cosmetics by read across applied to metabolomics data of in vitro skin and liver models. Arch Toxicol [Internet]. 2021; Available from: https://doi.org/10.1007/s00204-021-03136-7

55. Pawar G, Madden JC, Ebbrell D, Firman JW, Cronin MTD. In Silico Toxicology Data Resources to Support Read-Across and (Q)SAR [Internet]. Frontiers in Pharmacology . 2019. p. 561. Available from: https://www.frontiersin.org/article/10.3389/fphar.2019.00561

56. Escher SE, Kamp H, Bennekou SH, Bitsch A, Fisher C, Graepel R, et al. Towards grouping concepts based on new approach methodologies in chemical hazard assessment: the read-across approach of the EU-ToxRisk project. Arch Toxicol. Springer; 2019. p. 3643–67.

57. Berndt N, Holzhütter HG. Mathematical modeling of cellular metabolism. Recent Results in Cancer Research. 2016;207:221–32.

58. Qiu S, Cai Y, Yao H, Lin C, Xie Y, Tang S, et al. Small molecule metabolites: discovery of biomarkers and therapeutic targets. Signal Transduct Target Ther. Springer Nature; 2023.

59. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. Anal Chim Acta. Elsevier B.V.; 2015. p. 10–23.

60. Ruiz-Perez D, Guan H, Madhivanan P, Mathee K, Narasimhan G. So you think you can PLS-DA? BMC Bioinformatics. 2020;21.

61. Lu H, Chen Y, Nielsen J, Kerkhoven EJ. Kinetic Models of Metabolism. Metab Eng [Internet]. 2021. p. 153–70. Available from: https://doi.org/10.1002/9783527823468.ch5

62. Saa PA, Nielsen LK. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. Biotechnol Adv. Elsevier Inc.; 2017. p. 981–1003.

63. Curien G, Bastien O, Robert-Genthon M, Cornish-Bowden A, Cárdenas ML, Dumas R. Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. Mol Syst Biol. 2009;5.

64. König M, Bulik S, Holzhütter HG. Quantifying the contribution of the liver to glucose homeostasis: A detailed kinetic model of human hepatic glucose metabolism. PLoS Comput Biol. 2012;8.

65. Khodayari A, Maranas CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat Commun. 2016;7.

66. Maeda K, Hatae A, Sakai Y, Boogerd FC, Kurata H. MLAGO: machine learning-aided global optimization for Michaelis constant estimation of kinetic modeling. BMC Bioinformatics. 2022;23.

67. Foster CJ, Wang L, Dinh H V, Suthers PF, Maranas CD. Building kinetic models for metabolic engineering [Internet]. 2020. Available from: https://www.sciencedirect.com/science/article/pii/S0958166920301774

68. Strutz J, Martin J, Greene J, Broadbelt L, Tyo K. Metabolic kinetic modeling provides insight into complex biological questions, but hurdles remain. Curr Opin Biotechnol. Elsevier Ltd; 2019. p. 24–30.

69. Marin de Mas I, Herand H, Carrasco J, Nielsen LK, Johansson PI. A Protocol for the Automatic Construction of Highly Curated Genome-Scale Models of Human Metabolism. Bioengineering. 2023;10.

70. Chazalviel M, Frainay C, Poupin N, Vinson F, Merlet B, Gloaguen Y, et al. MetExploreViz: web component for interactive metabolic network visualization. Bioinformatics [Internet]. 2018;34:312–3. Available from: https://pubmed.ncbi.nlm.nih.gov/28968733

71. Jensen PA, Lutz KA, Papin JA. TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. BMC Syst Biol. 2011;5.

72. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol [Internet]. 2013/03/03. 2013;31:419–25. Available from: https://pubmed.ncbi.nlm.nih.gov/23455439

73. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Nucleic Acids Res. 2000. Available from: http://www.genome.ad.jp/kegg/

74. Henry CS, Dejongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010;28:977–82.

75. 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. 2018;14.

76. 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. PLoS Comput Biol [Internet]. 2013;9:e1002980. Available from: https://doi.org/10.1371/journal.pcbi.1002980

77. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States department of energy systems biology knowledgebase. Nat Biotechnol. Nature Publishing Group; 2018. p. 566–9.

78. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018;46:7542–53.

79. Wang H, Robinson JL, Kocabas P, Gustafsson J, Anton M, Cholley P-E, et al. Genome-scale metabolic network reconstruction of model animals as a platform for translational research. 2023;118.

80. Thiele I, Palsson B. Reconstruction annotation jamborees: A community approach to systems biology. Mol Syst Biol. 2010.

81. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003;5:264–76.

82. Haggart CR, Bartell JA, Saucerman JJ, Papin JA. Whole-genome metabolic network reconstruction and constraint-based modeling. Methods Enzymol. 2011;500:411–33.

83. Haggart CR, Bartell JA, Saucerman JJ, Papin JA. Whole-genome metabolic network reconstruction and constraint-based modeling. Methods Enzymol. 2011;500:411–33.

84. Gerlin L, Frainay C, Jourdan F, Baroukh C, Prigent S. Plant genome-scale metabolic networks. Metabolomics : Practical Guide to Design and Analysis. Academic Press Inc.; 2019. p. 237–70.

85. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41:36–42.

86. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. Nucleic Acids Res. 2021;49:D92–6.

87. Hotaling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: Where are we now? 2021;118.

88. Glont M, Nguyen TVN, Graesslin M, Hälke R, Ali R, Schramm J, et al. BioModels: expanding horizons to include more modelling approaches and formats. Nucleic Acids Res [Internet]. 2018;46:D1248–53. Available from: https://doi.org/10.1093/nar/gkx1023

89. 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 [Internet]. 2018;46:W495–502. Available from: https://doi.org/10.1093/nar/gky301

90. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44:D515–22.

91. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51:D1373–80.

92. Pantziri MDA, Klapa MI. Standardization of Human Metabolic Stoichiometric Models : Challenges and Directions. 2022;2:1–7.

93. Ravikrishnan A, Raman K. Critical assessment of genome-scale metabolic networks: The need for a unified standard. Brief Bioinform. 2015;16:1057–68.

94. Cook DJ, Nielsen J. Genome-scale metabolic models applied to human health and disease. Wiley Interdiscip Rev Syst Biol Med. 2017;9:1–18.

95. Robinson JL, Kocabaş P, Wang H, Cholley P-E, Cook D, Nilsson A, et al. An atlas of human metabolism. Sci Signal. 2020;13.

96. Moulin C, Tournier L, Peres S. Combining kinetic and constraint-based modelling to better understand metabolism dynamics. Processes. 2021;9.

97. Yasemi M, Jolicoeur M. Modelling cell metabolism: A review on constraint-based steady-state and kinetic approaches. Processes. 2021;9:1–38.

98. Rodríguez-Mier P, Poupin N, de Blasio C, Le Cam L, Jourdan F. DEXOM: Diversity-based enumeration of optimal context-specific metabolic networks. PLoS Comput Biol [Internet]. 2021;17:e1008730. Available from: https://doi.org/10.1371/journal.pcbi.1008730

99. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26:3140–2.

100. Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 2008;26:1003–10.

101. Poupin N, Corlu A, Cabaton NJ, Dubois-Pot-Schneider H, Canlet C, Person E, et al. Large-Scale Modeling Approach Reveals Functional Metabolic Shifts during Hepatic Differentiation. J Proteome Res [Internet]. 2019;18:204–16. Available from: https://doi.org/10.1021/acs.jproteome.8b00524

102. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics. 2010;26.

103. Aurich MK, Paglia G, Rolfsson Ó, Hrafnsdóttir S, Magnúsdóttir M, Stefaniak MM, et al. Prediction of intracellular metabolic states from extracellular metabolomic data. Metabolomics. 2015;11:603–19.

104. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol [Internet]. 2012;10:291–305. Available from: https://doi.org/10.1038/nrmicro2737

105. Varma A, Palsson BO. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Nat Biotechnol. 1994;12:994–8.

106. Bordbar A, Monk J, King Z, Palsson B. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet. 2014;15:107–20.

107. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes Escherichia coli traits. Nat Biotechnol. Nature Publishing Group; 2017. p. 904–8.

108. Burgard AP, Pharkya P, Maranas CD. OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. Biotechnol Bioeng. 2003;84:647–57.

109. Varma A, Palsson B 0. Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild-Type Escherichia coli W3110. Appl Environ Microbiol. 1994.

110. Becker SA, Palsson BO. Context-Specific Metabolic Networks Are Consistent with Experiments. PLoS Comput Biol [Internet]. 2008;4:e1000082. Available from: https://doi.org/10.1371/journal.pcbi.1000082

111. Roumans KHM, Sagarminaga JB, Peters HPF, Schrauwen P, Schrauwen-Hinderling VB. Liver fat storage pathways: Methodologies and dietary effects. Curr Opin Lipidol. Lippincott Williams and Wilkins; 2021. p. 9–15.

112. Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol [Internet]. 2010;6:401. Available from: https://doi.org/10.1038/msb.2010.56

113. Vlassis N, Pacheco MP, Sauter T. Fast Reconstruction of Compact Context-Specific Metabolic Network Models. PLoS Comput Biol [Internet]. 2014;10:e1003424. Available from: https://doi.org/10.1371/journal.pcbi.1003424

114. Wang Y, Eddy JA, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol. 2012;6:153.

115. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Syst. 2017;4:318-329.e6.

116. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. PLoS Comput Biol [Internet]. 2012;8:e1002518. Available from: https://doi.org/10.1371/journal.pcbi.1002518

117. Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 2008;26:1003–10.

118. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C. Elementary flux modes in a nutshell: Properties, calculation and applications. Biotechnol J. 2013. p. 1009–16.

119. Encyclopedia of Systems Biology. Encyclopedia of Systems Biology. Springer New York; 2013.

120. Terzer M, Stelling J. Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics. 2008;24:2229–35.

121. von Kamp A, Schuster S. Metatool 5.0: Fast and flexible elementary modes analysis. Bioinformatics. 2006;22:1930–1.

122. Orth JD, Fleming RMT, Palsson BØ. Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide . EcoSal Plus. 2010;4.

123. Yilmaz LS, Li X, Nanda S, Fox B, Schroeder F, Walhout AJ. Modeling tissue-relevant *Caenorhabditis elegans* metabolism at network, pathway, reaction, and metabolite levels. Mol Syst Biol [Internet]. 2020 [cited 2020 Oct 15];16. Available from: https://onlinelibrary.wiley.com/doi/10.15252/msb.20209649

124. 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. 2019;5.

125. Robaina-Estévez S, Nikoloski Z. On the effects of alternative optima in context-specific metabolic model predictions. PLoS Comput Biol [Internet]. 2017;13:e1005568. Available from: https://doi.org/10.1371/journal.pcbi.1005568

126. Robaina Estévez S, Nikoloski Z. Generalized framework for context-specific metabolic model extraction methods. Front Plant Sci [Internet]. 2014;5. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2014.00491

127. Machado D, Herrgård M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Comput Biol [Internet]. 2014;10:e1003580. Available from: https://doi.org/10.1371/journal.pcbi.1003580

128. Safak Yilmaz L, Walhout AJM. A Caenorhabditis elegans Genome-Scale Metabolic Network Model. Cell Syst. 2016;2:297–311.

129. Yang C, Hua Q, Shimizu K. Integration of the information from gene expression and metabolic fluxes for the analysis of the regulatory mechanisms in Synechocystis. Appl Microbiol Biotechnol. 2002;58:813–22.

130. Rossell S, van der Weijden CC, Lindenbergh A, van Tuijl A, Francke C, Bakker BM, et al. Unraveling the complexity of flux regulation: A new method demonstrated for nutrient starvation in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences [Internet]. 2006;103:2166–71. Available from: www.pnas.orgcgidoi10.1073pnas.0509831103

131. Moxley JF, Jewett MC, Antoniewicz MR, Villas-Boas SG, Alper H, Wheeler RT, et al. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. Proceedings of the National Academy of Sciences. 2009;106:6477–82.

132. Rossell S, Huynen MA, Notebaart RA. Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements. PLoS Comput Biol [Internet]. 2013;9:e1002988. Available from: https://doi.org/10.1371/journal.pcbi.1002988

133. Tsai PY, Lee MS, Jadhav U, Naqvi I, Madha S, Adler A, et al. Adaptation of pancreatic cancer cells to nutrient deprivation is reversible and requires glutamine synthetase stabilization by mTORC1. Proc Natl Acad Sci U S A. 2021;118.

134. Espada L, Dakhovnik A, Chaudhari P, Martirosyan A, Miek L, Poliezhaieva T, et al. Loss of metabolic plasticity underlies metformin toxicity in aged Caenorhabditis elegans. Nat Metab. 2020;2:1316–31.

135. Klamt S, Haus U-U, Theis F. Hypergraphs and Cellular Networks. PLoS Comput Biol. 2009;5:e1000385.

136. Shen T, Zhang Z, Chen Z, Gu D, Liang S, Xu Y, et al. A genome-scale metabolic network alignment method within a hypergraph-based framework using a rotational tensor-vector product. Sci Rep [Internet]. 2018;8:16376. Available from: https://doi.org/10.1038/s41598-018-34692-1

137. Mithani A, Preston GM, Hein J. Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. Bioinformatics. 2009;25:1831–2.

138. Pearcy N, Chuzhanova N, Crofts JJ. Complexity and robustness in hypernetwork models of metabolism. J Theor Biol. 2016;406:99–104.

139. Pearcy N, Crofts JJ, Chuzhanova N. Hypergraph models of metabolism. International Journal of Bioengineering and Life Sciences. 2014;8.

140. Frainay C, Jourdan F. Computational methods to identify metabolic sub-networks based on metabolomic profiles. Brief Bioinform. 2017;18:43–56.

141. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math (Heidelb) [Internet]. 1959;1:269–71. Available from: https://doi.org/10.1007/BF01386390

142. Bellman R. ON A ROUTING PROBLEM*.

143. Floyd RW. Algorithm 97: Shortest Path. Commun ACM [Internet]. 1962;5:345. Available from: https://doi.org/10.1145/367766.368168

144. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep [Internet]. 2019;9:5233. Available from: https://doi.org/10.1038/s41598-019-41695-z

145. Von Luxburg U. A Tutorial on Spectral Clustering [Internet]. Stat Comput. 2007. Available from: www.springer.com.

146. Zhang Q, Zhang ZD. SubNet: A Java application for subnetwork extraction. Bioinformatics. 2013;29:2509–11.

147. Duesbury, Holliday, JD, Willett. Maximum Common Subgraph Isomorphism Algorithms. Communications in Mathematical and in Computer Chemistry [Internet]. 2017;77:213–32. Available from: http://eprints.whiterose.ac.uk/102232/

148. Weisstein E. "Subgraph." MathWorld--A Wolfram Web Resource. 2023.

149. R. L. Graham, P. Hell. on the history of the minimum spanning tree problem. Ann Hist Comput. 1985;7.

150. Eppstein D. FINDING THE k SHORTEST PATHS *. Society for Industrial and Applied Mathematics [Internet]. 1998;28:652–73. Available from: http://www.siam.org/journals/ojsa.php

151. Vitter JS, Zaroliagis CD. Algorithm engineering : 3rd International Workshop, WAE '99 London, UK, July 19-21, 1999 : proceedings. Springer; 1999.

152. Dupont P, Callut J, Dooms G, Monette J-N, Deville Y. "Relevant subgraph extraction from random walks in a graph." 2006;1–30. Available from: http://hdl.handle.net/2078.1/109752

153. Fooshee D, Andronico A, Baldi P. ReactionMap: An efficient atom-mapping algorithm for chemical reactions. J Chem Inf Model. 2013;53:2812–9.

154. Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS. SAR maps: A new SAR visualization technique for medicinal chemists. J Med Chem. 2007;50:5926–37.

155. Espejo R, Mestre G, Postigo F, Lumbreras S, Ramos A, Huang T, et al. Exploiting graphlet decomposition to explain the structure of complex networks: the GHuST framework. Sci Rep. 2020;10.

156. Finotelli P, Piccardi C, Miglio E, Dulio P. A Graphlet-Based Topological Characterization of the Resting-State Network in Healthy People. Front Neurosci. 2021;15.

157. Tu K, Li J, Towsley D, Braines D, Turner LD. Gl2vec: Learning feature representation using graphlets for directed networks. Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019. Association for Computing Machinery, Inc; 2019. p. 216–21.

158. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. Science (1979). 2002;298:824–7.

159. Pržulj N. Biological network comparison using graphlet degree distribution. Bioinformatics. Oxford University Press; 2007.

160. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res [Internet]. 2014/10/13. 2015;43:D921–7. Available from: https://pubmed.ncbi.nlm.nih.gov/25313160

161. Urushidani T. Prediction of Hepatotoxicity Based on the Toxicogenomics Database. Hepatotoxicity. 2007. p. 507–29.

162. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20:307–15.

163. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249–64.

164. Zeilinger K, Freyer N, Damm G, Seehofer D, Knöspel F. Cell sources for in vitro human liver cell culture models. Exp Biol Med (Maywood). 2016;241:1684–98.

165. Heusinkveld HJ, Wackers PFK, Schoonen WG, van der Ven L, Pennings JLA, Luijten M. Application of the comparison approach to open TG-GATEs: A useful toxicogenomics tool for detecting modes of action in chemical risk assessment. Food Chem Toxicol. 2018;121:115–23.

166. Liu Z, Zhu L, Thakkar S, Roberts R, Tong W. Can Transcriptomic Profiles from Cancer Cell Lines Be Used for Toxicity Assessment? Chem Res Toxicol. 2020;33:271–80.

167. Taškova K, Fontaine J-F, Mrowka R, Andrade-Navarro MA. Evaluation of in vivo and in vitro models of toxicity by comparison of toxicogenomics data with the literature. Methods. 2018;132:57–65.

168. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012/01/17. 2012;28:882–3.

169. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discov Today. 2016;21:648–53.

170. Thakkar S, Chen M, Fang H, Liu Z, Roberts R, Tong W. The Liver Toxicity Knowledge Base (LKTB) and drug-induced liver injury (DILI) classification for assessment of human liver injury. Expert Rev Gastroenterol Hepatol. 2018;12:31–8.

171. Hoofnagle JH, Serrano J, Knoben JE, Navarro VJ. LiverTox: a website on drug-induced liver injury. Hepatology. 2013. p. 873–4.

172. Biour M, Ben Salem C, Chazouillères O, Grangé J-D, Serfati L, Poupon R. Hépatotoxicité des médicaments 14e mise à jour du fichier bibliographique des atteintes hépatiques et des médicaments responsables. Gastroenterol Clin Biol. 2004;28:720–59.

173. Biour M, Slim R, Elouni B, Chaker ben salem, Chaillet P. Hépatox®. Présentation du fichier bibliographique des atteintes hépatiques et des médicaments responsables. EMC - Hépatologie. 2010;5:1–3.

174. Binder H, Preibisch S. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. Biophys J. 2005;89:337–52.

175. Schneider S, Smith T, Hansen U. SCOREM: statistical consolidation of redundant expression measures. Nucleic Acids Res [Internet]. 2012;40:e46–e46. Available from: https://doi.org/10.1093/nar/gkr1270

176. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23:257–8.

177. Tuna S, Niranjan M. Reducing the algorithmic variability in transcriptome-based inference. Bioinformatics. 2010;26:1185–91.

178. Richelle A, Joshi C, Lewis NE. Assessing key decisions for transcriptomic data integration in biochemical networks. PLoS Comput Biol. 2019;15:e1007185.

179. Zhou X, Wang X, Dougherty ER. Binarization of microarray data on the basis of a mixture model. Mol Cancer Ther. 2003;2:679–84.

180. Schneider S, Smith T, Hansen U. SCOREM: statistical consolidation of redundant expression measures. Nucleic Acids Res. 2012;40:e46–e46.

181. Joshi CJ, Schinn S-M, Richelle A, Shamie I, O'Rourke EJ, Lewis NE. StanDep: Capturing transcriptomic variability improves context-specific metabolic models. PLoS Comput Biol. 2020;16:e1007764.

182. Richelle A, Joshi C, Lewis NE. Assessing key decisions for transcriptomic data integration in biochemical networks. PLoS Comput Biol. 2019;15:e1007185.

183. Zilliox MJ, Irizarry RA. A gene expression bar code for microarray data. Nat Methods. 2007;4:911–3.

184. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res. 2011;39:D1011-5.

185. McCall MN, Jaffee HA, Zelisko SJ, Sinha N, Hooiveld G, Irizarry RA, et al. The Gene Expression Barcode 3.0: improved data processing and mining tools. Nucleic Acids Res. 2014;42:D938-43.

186. 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 [Internet]. 2020;48:D407–15. Available from: https://doi.org/10.1093/nar/gkz1055

187. Glont M, Nguyen TVN, Graesslin M, Hälke R, Ali R, Schramm J, et al. BioModels: expanding horizons to include more modelling approaches and formats. Nucleic Acids Res [Internet]. 2018;46:D1248–53. Available from: https://doi.org/10.1093/nar/gkx1023

188. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol [Internet]. 2013;7:74. Available from: https://doi.org/10.1186/1752-0509-7-74

189. Heindel JJ, Blumberg B, Cave M, Machtinger R, Mantovani A, Mendez MA, et al. Metabolism disrupting chemicals and metabolic disorders. Reprod Toxicol. 2017;68:3–33.

190. Sarni ROS, Kochi C, Suano-Souza FI. Childhood obesity: an ecological perspective. J Pediatr (Rio J). 2022;98 Suppl 1:S38–46.

191. Gore AC, Chappell VA, Fenton SE, Flaws JA, Nadal A, Prins GS, et al. EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. Endocr Rev. 2015;36:E1–150.

192. Heal DJ, Gosden J, Jackson HC, Cheetham SC, Smith SL. Metabolic consequences of antipsychotic therapy: preclinical and clinical perspectives on diabetes, diabetic ketoacidosis, and obesity. Handb Exp Pharmacol. 2012;135–64.

193. Miranda RA, Silva BS, de Moura EG, Lisboa PC. Pesticides as endocrine disruptors: programming for obesity and diabetes. Endocrine. 2023;79:437–47.

194. PEARSON K. On the criterion that a given system of derivations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen . Philosophical Magazine. 1900;50:157–75.

195. Fisher RA. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. urnal of the Royal Statistical Society. 1922;85:87–94.

196. Lin M, Lucas HC, Shmueli G. Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research [Internet]. 2013;24:906–17. Available from: https://doi.org/10.1287/isre.2013.0480

197. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4:279–82.

198. Karpen SC. P Value Problems. Am J Pharm Educ. 2017. p. 6570.

199. Efthimiou O. Practical guide to the meta-analysis of rare events. Evidence Based Mental Health [Internet]. 2018;21:72 LP – 76. Available from: http://ebmh.bmj.com/content/21/2/72.abstract

200. Jeanmonod R, Asuka E, Jeanmonod D. inborn errors of metabolism. Treasure Island (FL); 2023.

201. equation-of-circle-when-three-points-on-the-circle-are-given. https://www.geeksforgeeks.org/equation-of-circle-when-three-points-on-the-circle-are-given/.

202. Ferk P, Dariš B. The influence of dimethyl sulfoxide (DMSO) on metabolic activity and morphology of melanoma cell line WM-266-4. Cell Mol Biol. 2018;64:41–3.

203. Kim H-Y. Statistical notes for clinical researchers: Sample size calculation 2. Comparison of two independent proportions. Restor Dent Endod. 2016;41:154.

204. Lin SJ, Lu TP, Yu QY, Hsiao CK. Probabilistic prioritization of candidate pathway association with pathway score. BMC Bioinformatics. 2018;19.

205. Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019;10.

206. Göttlicher M, Minucci S, Zhu P, Krämer OH, Schimpf A, Giavara S, et al. Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells. EMBO J. 2001;20:6969–78.

207. Seto E, Yoshida M. Erasers of histone acetylation: The histone deacetylase enzymes. Cold Spring Harb Perspect Biol. 2014;6.

208. Anthérieu S, Rogue A, Fromenty B, Guillouzo A, Robin M-A. Induction of vesicular steatosis by amiodarone and tetracycline is associated with up-regulation of lipogenic genes in heparg cells. Hepatology [Internet]. 2011;53:1895–905. Available from: https://doi.org/10.1002/hep.24290

209. Hubel E, Fishman S, Holopainen M, Käkelä R, Shaffer O, Houri I, et al. Repetitive amiodarone administration causes liver damage via adipose tissue ER stress-dependent lipolysis, leading to hepatotoxic free fatty acid accumulation. American Journal of Physiology-Gastrointestinal and Liver Physiology [Internet]. 2021;321:G298–307. Available from: https://doi.org/10.1152/ajpgi.00458.2020

210. Pessayre D, Mansouri A, Haouzi D, Fromenty B. Hepatotoxicity due to mitochondrial dysfunction. Cell Biol Toxicol. 1999;15:367–73.

211. Allard J, Bucher S, Massart J, Ferron P-J, Le Guillou D, Loyant R, et al. Drug-induced hepatic steatosis in absence of severe mitochondrial dysfunction in HepaRG cells: proof of multiple mechanism-based toxicity. Cell Biol Toxicol [Internet]. 2021;37:151–75. Available from: https://doi.org/10.1007/s10565-020-09537-1

212. Passi A, Tibocha-Bonilla JD, Kumar M, Tec-Campos D, Zengler K, Zuniga C. Genome-scale metabolic modeling enables in-depth understanding of big data. Metabolites. MDPI; 2022.

213. https://www.jpboseret.eu/biologie/index.php?option=com_content&view=article&id=27&Itemid=138.

214. Bar-Peled L, Kory N. Principles and functions of metabolic compartmentalization. Nat Metab. Nature Research; 2022. p. 1232–44.

215. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. PNAS February. 2004.

216. 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. 2016;12:109.

217. Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol. 2018;36:272–81.

218. Smallbone K. Striking a balance with Recon 2.1. ArXiv [Internet]. 2013; Available from: http://arxiv.org/abs/1311.5696

219. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-. L. the large scale organization of metabolic networks.

220. Fell AD, Wagner A. the small world of metabolism.

221. Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings of the Royal Society B: Biological Sciences. 2001;268:1803–10.

222. Tandon R, Ravikumar P. On the Difficulty of Learning Power Law Graphical Models.

223. Arita M. The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A. 2004;101:1543–7.

224. Croes D, Couche F, Wodak SJ, van Helden J. Metabolic PathFinding: Inferring relevant pathways in biochemical networks. Nucleic Acids Res. 2005;33.

225. Croes D, Couche F, Wodak SJ, Van Helden J. Inferring meaningful pathways in weighted metabolic networks. J Mol Biol. 2006;356:222–36.

226. Pertusi DA, Stine AE, Broadbelt LJ, Tyo KEJ. Efficient searching and annotation of metabolic networks using chemical similarity. Bioinformatics. 2015;31:1016–24.

227. Rahman SA, Advani P, Schunk R, Schrader R, Schomburg D. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics. 2005;21:1189–93.

228. Ichida K, Hosoyamada M, Hosoya Tatsuo, Endou H. Primary Metabolic and Renal Hyperuicemia. Genetic Diseases of the Kidney. 2009;651–60.

229. Latendresse M, Malerich JP, Travers M, Karp PD. Accurate atom-mapping computation for biochemical reactions. J Chem Inf Model. 2012;52:2970–82.

230. Kotera M, Yamamoto R, Yabuzaki J, Hattori M, Komeno T, Tonomura K, et al. RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions [Internet]. 2004. Available from: https://www.researchgate.net/publication/228501550

231. Rahman SA, Torrance G, Baldacci L, Martínez Cuesta S, Fenninger F, Gopal N, et al. Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics [Internet]. 2016;32:2065–6. Available from: https://doi.org/10.1093/bioinformatics/btw096

232. Faust K, Croes D, van Helden J. Metabolic Pathfinding Using RPAIR Annotation. J Mol Biol. 2009;388:390–414.

233. Kotera M, Yamamoto R, Yabuzaki J, Hattori M, Komeno T, Tonomura K, et al. RPAIR: A reactant-pair database representing chemical changes in enzymatic reactions RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions Min-A Oh [Internet]. 2004. Available from: https://www.researchgate.net/publication/228501550

234. Faust K, Croes D, van Helden J. Metabolic Pathfinding Using RPAIR Annotation. J Mol Biol. 2009;388:390–414.

235. Faust K, Dupont P, Callut J, Van Helden J. Systems biology Pathway discovery in metabolic networks by subgraph extraction. 2010;26:1211–8. Available from: https://academic.oup.com/bioinformatics/article/26/9/1211/199334

236. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, et al. PathPred: An enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 2010;38.

237. Knopp S, Sanders P, Schultes D, Schulz F, Wagner D. Computing Many-to-Many Shortest Paths Using Highway Hierarchies. 2007 Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX) [Internet]. Society for Industrial and Applied Mathematics; 2007. p. 36–45. Available from: https://doi.org/10.1137/1.9781611972870.4

238. Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining [Internet]. 1996;226–31. Available from: www.aaai.org

239. Fränti P, Sieranoja S. How much can k-means be improved by using better intialization and repeats. Pattern Recognit. 2019;95–112.

240. Karp RM. Reducibility among Combinatorial Problems BT - Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, an. In: Miller RE, Thatcher JW, Bohlinger JD, editors. Boston, MA: Springer US; 1972. p. 85–103. Available from: https://doi.org/10.1007/978-1-4684-2001-2_9

241. Voß S. Steiner's problem in graphs: heuristic methods. Discrete Appl Math (1979) [Internet]. 1992;40:45–72. Available from: https://www.sciencedirect.com/science/article/pii/0166218X92900212

242. Mnif L, Sellami R, Masmoudi J. Valproic Acid and Hepatic Steatosis: A Possible Link? About a Case Report. Psychopharmacol Bull. 2016;46:59–62.

243. Zhou J, Wang X, Wang M, Chang Y, Zhang F, Ban Z, et al. The lysine catabolite saccharopine impairs development by disrupting mitochondrial homeostasis. Journal of Cell Biology [Internet]. 2018;218:580–97. Available from: https://doi.org/10.1083/jcb.201807204

244. Fromenty B, Pessayre D. Inhibition of mitochondrial beta-oxidation as a mechanism of hepatotoxicity. Pharmacol Ther. 1995;67:101–54.

245. Lheureux PER, Penaloza A, Zahir S, Gris M. Science review: Carnitine in the treatment of valproic acid-induced toxicity – what is the evidence? Crit Care [Internet]. 2005;9:431. Available from: https://doi.org/10.1186/cc3742

246. Maciejak P, Szyndler J, Kołosowska K, Turzyńska D, Sobolewska A, Walkowiak J, et al. Valproate disturbs the balance between branched and aromatic amino acids in rats. Neurotox Res. 2014;25:358–68.

247. Shibata K, Kondo R, Sano M, Fukuwatari T. Increased conversion of tryptophan to nicotinamide in rats by dietary valproate. Biosci Biotechnol Biochem. 2013;77:295–300.

248. Lee C, Cunningham P. Community detection: effective evaluation on large social networks. J Complex Netw [Internet]. 2014;2:19–37. Available from: https://doi.org/10.1093/comnet/cnt012

249. Pusa T, Ferrarini MG, Andrade R, Mary A, Marchetti-Spaccamela A, Stougie L, et al. MOOMIN – Mathematical explOration of 'Omics data on a MetabolIc Network. Bioinformatics [Internet]. 2020;36:514–23. Available from: https://doi.org/10.1093/bioinformatics/btz584

250. Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics [Internet]. 2018;34:1009–15. Available from: https://doi.org/10.1093/bioinformatics/btx682

251. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics [Internet]. 2019;35:i501–9. Available from: https://doi.org/10.1093/bioinformatics/btz318

252. Lee B, Zhang S, Poleksic A, Xie L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis [Internet]. Frontiers in Genetics . 2020. p. 1381. Available from: https://www.frontiersin.org/article/10.3389/fgene.2019.01381

253. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform [Internet]. 2018;19:325–40. Available from: https://doi.org/10.1093/bib/bbw113

254. Yang Z-Y, Liu X-Y, Shu J, Zhang H, Ren Y-Q, Xu Z-B, et al. Multi-view based integrative analysis of gene expression data for identifying biomarkers. Sci Rep. 2019;9:13504.

255. Su R, Yang H, Wei L, Chen S, Zou Q. A multi-label learning model for predicting drug-induced pathology in multi-organ based on toxicogenomics data. PLoS Comput Biol [Internet]. 2022;18:e1010402. Available from: https://doi.org/10.1371/journal.pcbi.1010402

256. Li T, Tong W, Roberts R, Liu Z, Thakkar S. DeepDILI: Deep Learning-Powered Drug-Induced Liver Injury Prediction Using Model-Level Representation. Chem Res Toxicol [Internet]. 2021;34:550–65. Available from: https://doi.org/10.1021/acs.chemrestox.0c00374

257. Anand DV, Xu Q, Wee JJ, Xia K, Sum TC. Topological feature engineering for machine learning based halide perovskite materials design. NPJ Comput Mater. 2022;8.

258. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open. Elsevier B.V.; 2020. p. 57–81.

259. Chen J, Si YW, Un CW, Siu SWI. Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. J Cheminform. 2021;13.

260. Cremer J, Medrano Sandonas L, Tkatchenko A, Clevert D-A, De Fabritiis G. Equivariant Graph Neural Networks for Toxicity Prediction. Chem Res Toxicol [Internet]. 2023; Available from: https://pubs.acs.org/doi/10.1021/acs.chemrestox.3c00032

261. Ma G, Ahmed NK, Willke TL, Yu PS. Deep graph similarity learning: a survey. Data Min Knowl Discov. 2021;35:688–725.

262. Noronha A, Daníelsdóttir AD, Gawron P, Jóhannsson F, Jónsdóttir S, Jarlsson S, et al. ReconMap: An interactive visualization of human metabolism. Bioinformatics. 2017;33:605–7.

Annexes

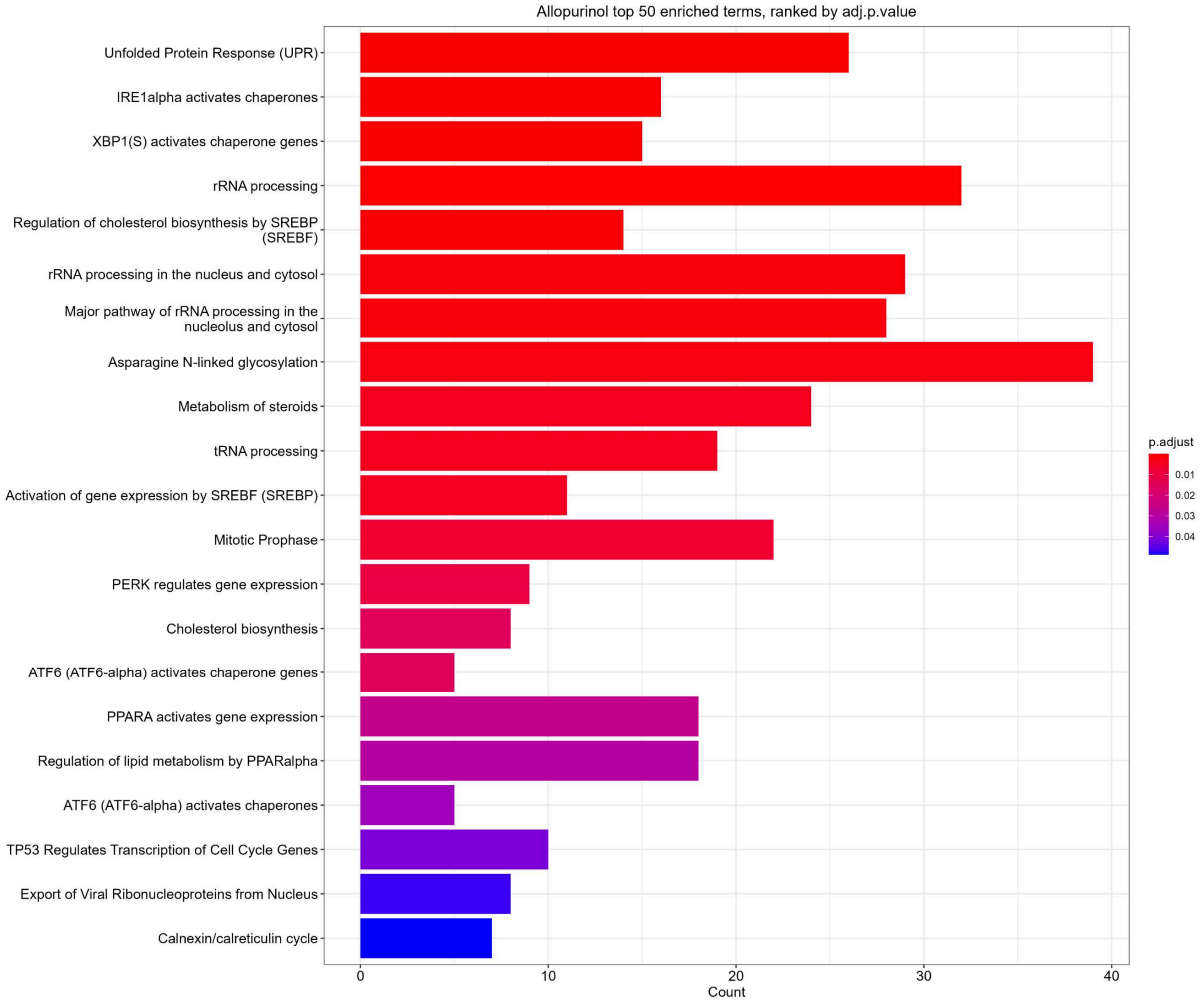

Figure 55 : Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 140µM d'allopurinol pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*

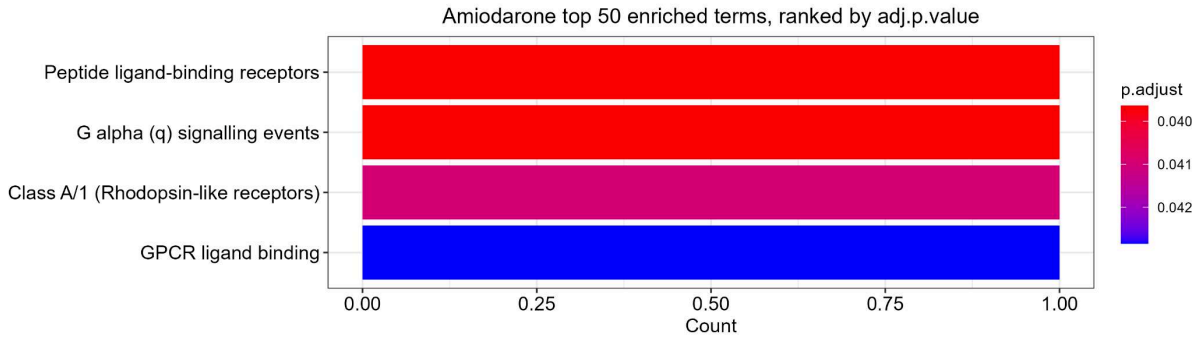

Figure 56 : Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 7µM d'amiodarone pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*

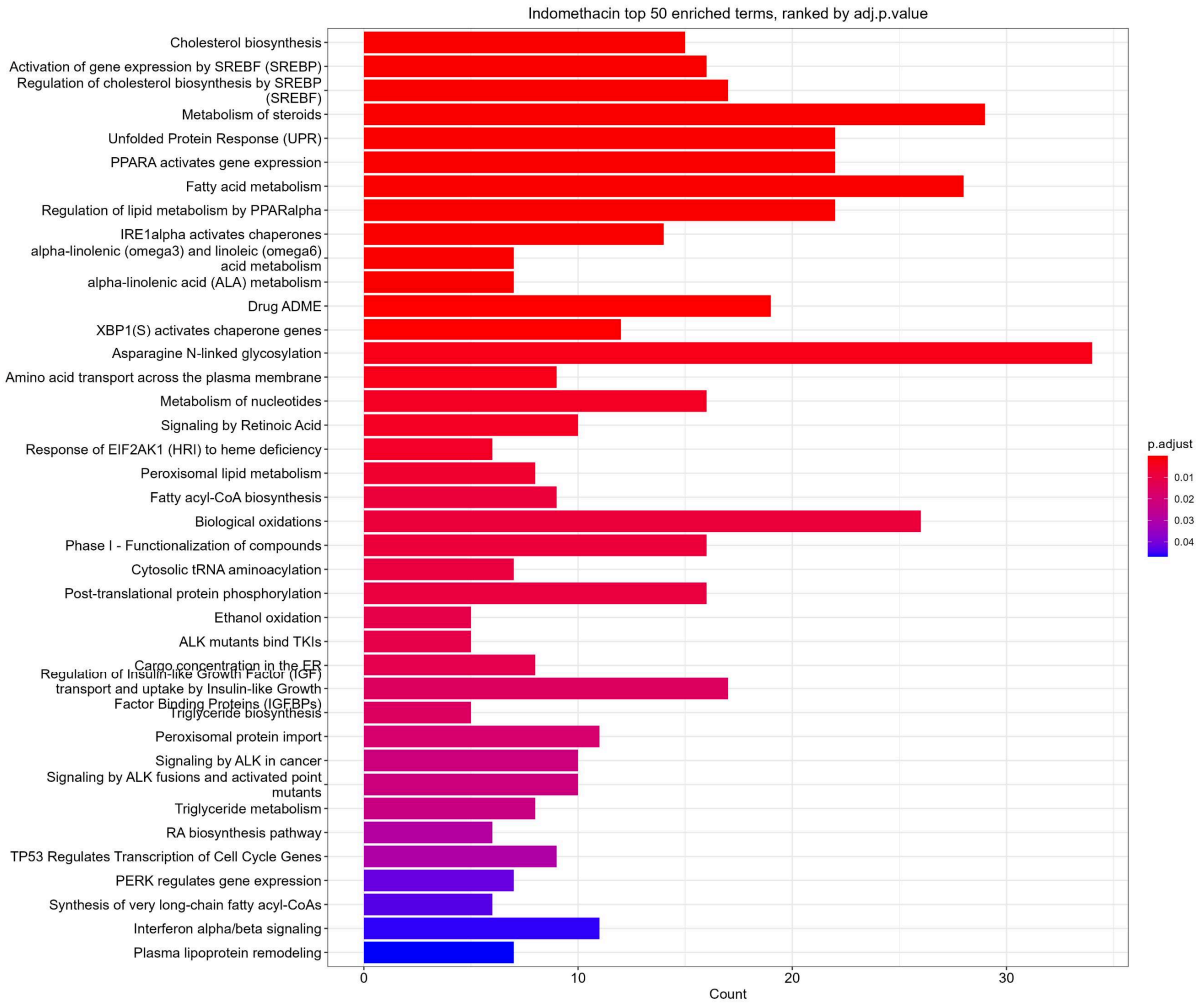

Figure 57: Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 200µM d'indométacine pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*

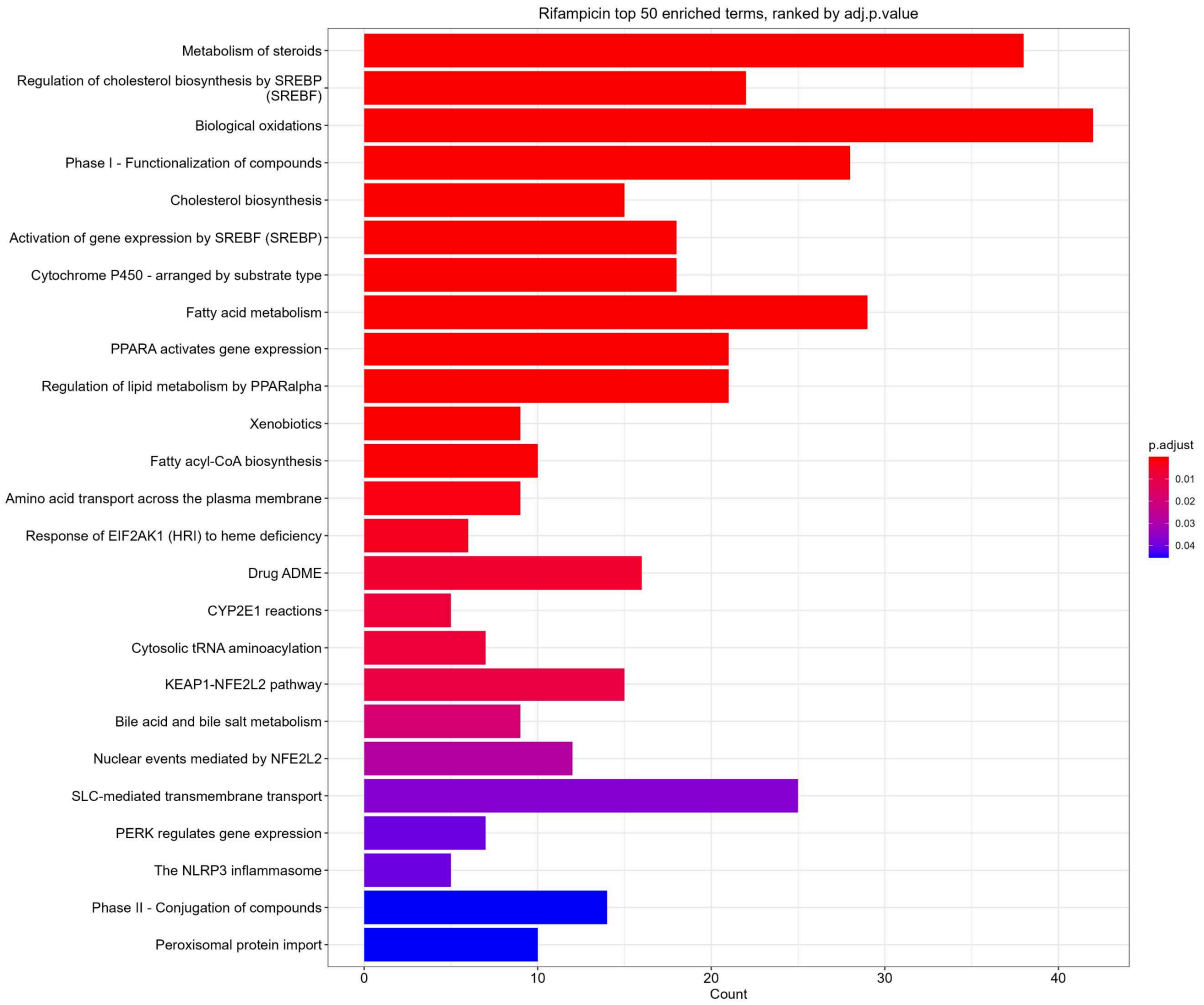

Figure 58: Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 70µM de rifampicine pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*

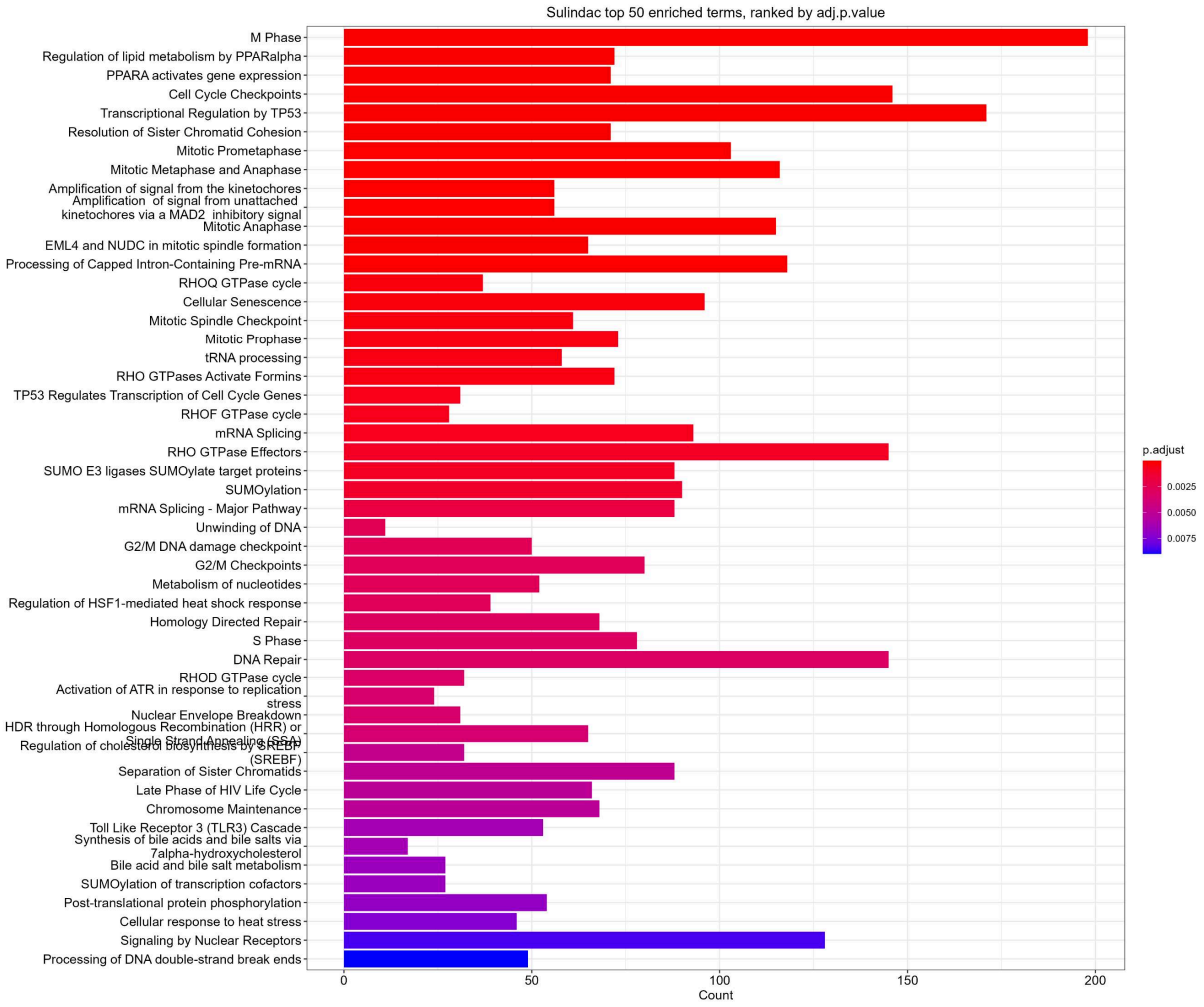

Figure 59 : Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 3000µM de sulindac pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*

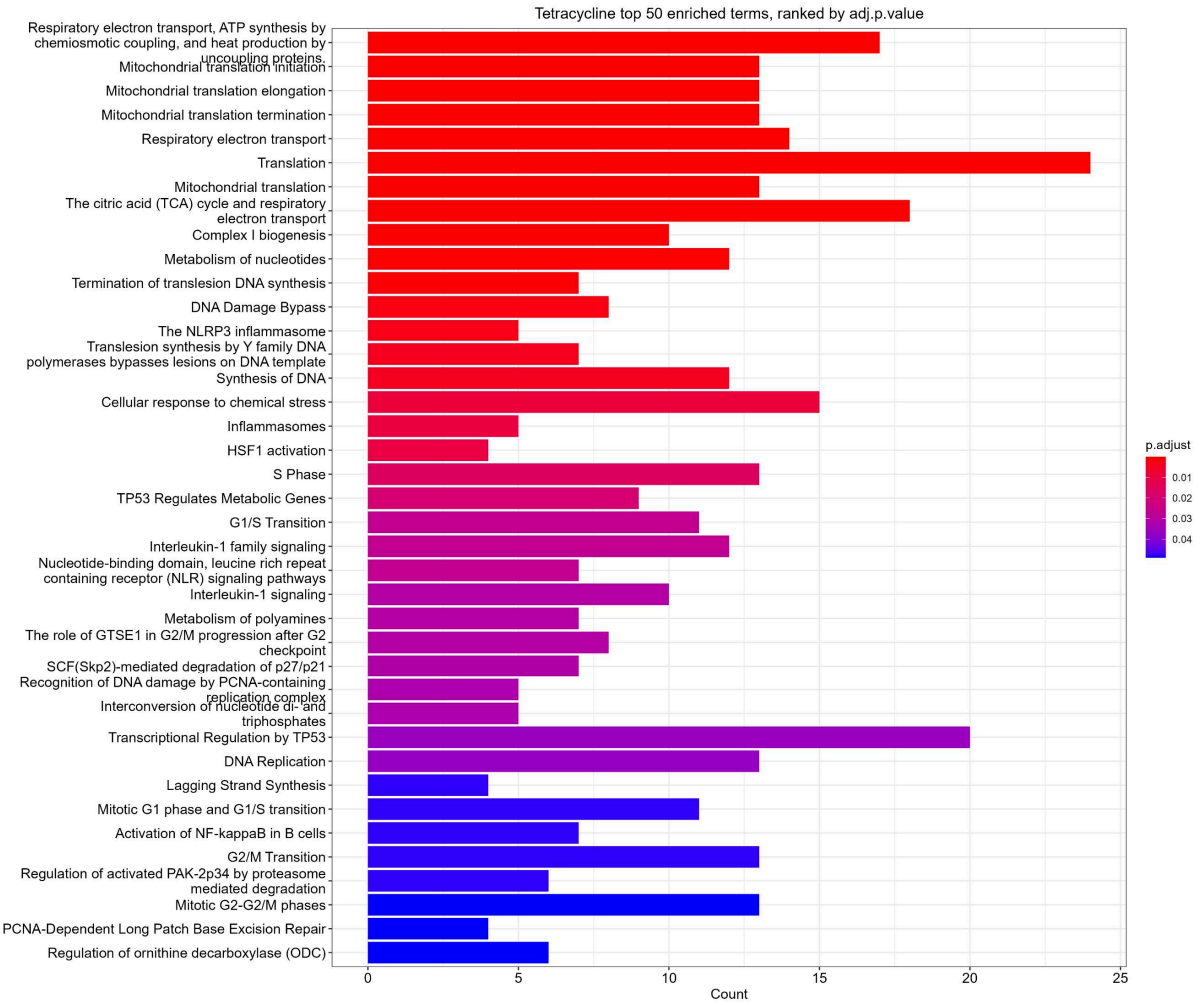

Figure 60 : Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 25µM de tétracycline pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*

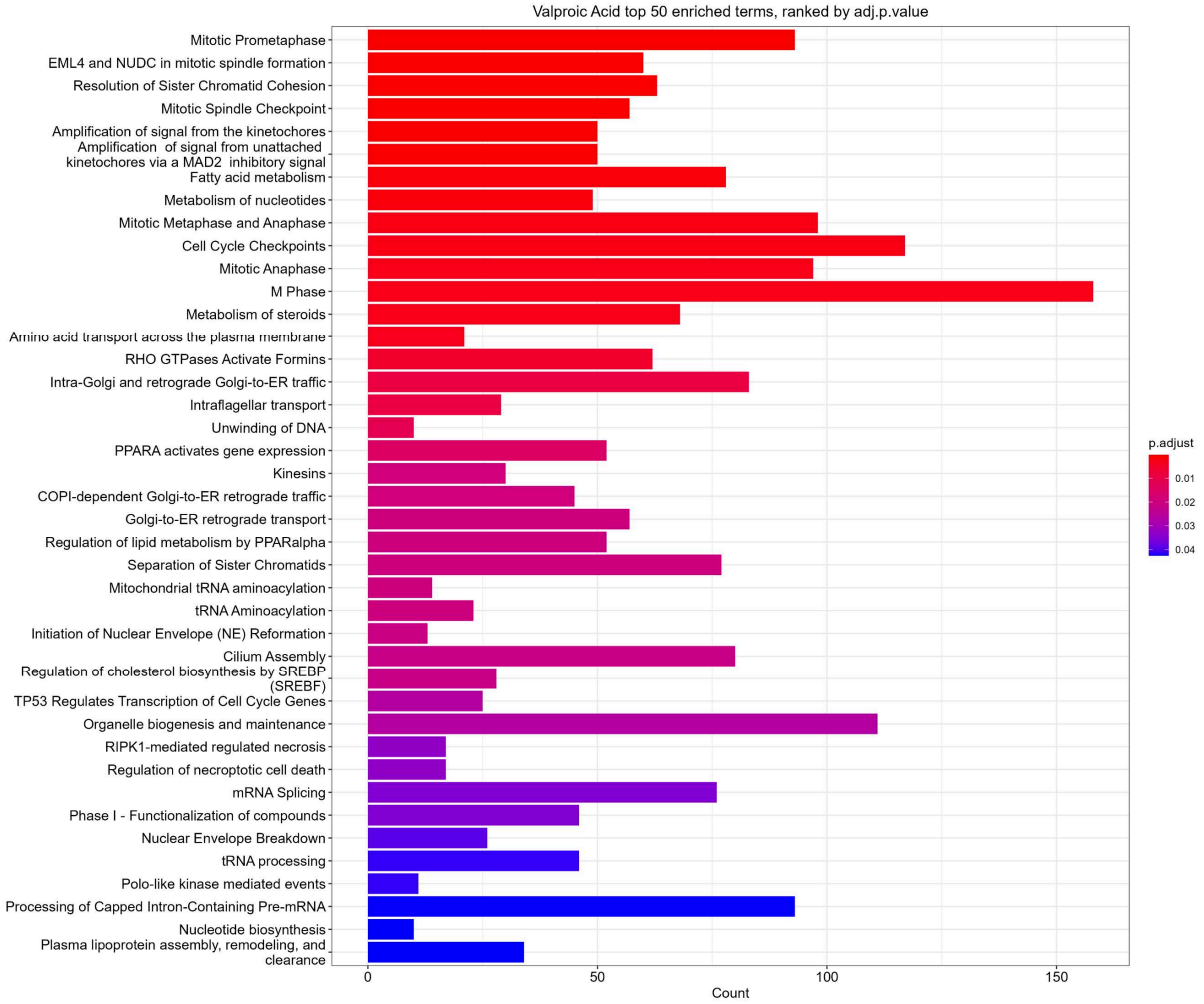

Figure 61: Analyse de sur-représentation pour la liste des DEGs identifiés suite à l'exposition de HPH à 5000µM d'acide valproïque pendant 24 heures avec le package R « ReactomePA ». *L'analyse a été réalisée avec le test exact de Fisher sur la base de données « Reactome 2022 », les p-valeurs ont été corrigées par la méthode de Benjamini-Hochberg. Nous avons considéré comme DEGs les gènes avec un log2(absFC)> 0.26 et une p-valeur corrigée par FDR inférieure à 0.05.*