Untargeted approaches in isotopic studies of metabolism and application to ab initio metabolic reconstruction

Noemie Butin

▶ **To cite this version:**

Noemie Butin. Untargeted approaches in isotopic studies of metabolism and application to ab initio metabolic reconstruction. Microbiology and Parasitology. Université Paul Sabatier - Toulouse III, 2023. English. NNT : 2023TOU30136 . tel-04356559

HAL Id: tel-04356559

https://theses.hal.science/tel-04356559

Submitted on 20 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Noémie BUTIN

Le 18 juillet 2023

Approches non ciblées de profilage isotopique pour l'analyse du métabolisme et application à la reconstruction métabolique ab initio

---

Ecole doctorale : **SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingenieries**

Spécialité : **Ingénieries microbienne et enzymatique**

Unité de recherche :
TBI - Toulouse Biotechnology Institute, Bio & Chemical Engineering

Thèse dirigée par
Jean-Charles PORTAIS

Jury

M. Benjamin PFEUTY, Rapporteur
Mme Sophie COLOMBIé, Rapporteure
M. Jean-Charles PORTAIS, Directeur de thèse
Mme Stéphanie HEUX, Présidente

Un grand nombre de personnes ont participé à l'aboutissement de cette thèse, par leur aide, leur confiance et leur soutien.

Je tiens tout d'abord à remercier les membres de mon jury, Benjamin Pfeuty et Sophie Colombié pour avoir accepté de lire et d'évaluer mon travail. Merci également à Stéphanie Heux pour avoir présidé mon jury. Je suis ravie d'avoir été ta première thésarde en qualité de présidente et j'espère que nous continuerons à avoir des échanges aussi agréables que ceux que nous avons eu pendant la préparation de la soutenance.

À Jean-Charles Portais, mon directeur de thèse. Lorsque tu m'as proposé ce sujet de thèse, l'aspect multidisciplinaire m'est apparu comme un challenge attractif. Par cette diversité de domaines, ce sujet n'a pas été évident à prendre en main mais j'ai toujours été bien entourée et accompagnée. Merci de m'avoir fait confiance pour travailler sur ce sujet que tu portes depuis plusieurs années et pour mener à bien cette thèse. Comme tu le dis, la recherche a toujours ses bas et ses hauts, et ces derniers valent le coup ! Je suis ravie d'avoir parcouru ce chemin avec toi.

À Floriant Bellvert, I dit it, Merci de m'avoir accompagnée ces dernières années, et ce, même bien avant le début de thèse. J'ai apprécié de t'avoir en support (technique et moral) toutes ces années, notamment pendant ces longues discussions d'ordre sémantique ; de découvrir ta méthode caractéristique de réparation de panne analytique (on touche à tout, on démonte tout) ; et de te voir toujours valoriser ta team. J'ai apprécié être toujours accueillie et écoutée, autant pendant les coups de mou que les coups de peps. Hâte de repartir en congrès avec l'équipe et toi !

À MetaToul FluxoMet, comme on le dit souvent dans le bureau, grand cœur sur vous ! A Cécilia, merci pour implication et ton dynamisme pour la réalisation de mes manips, à la team RMN, Lindsay et en particulier Edern pour tes retours avant les présentations. A Loïc, ravie que tu t'inscrives dans ce projet et qu'on continue à travailler (et plus) ensemble ! Je remercie tout particulièrement la team MS : Lara, Hanna, Amandine, Nina, Maud. Entre nous, j'espère que vous serez indulgentes la prochaine fois que je toucherai un spectro. Merci pour votre disponibilité, votre soutien permanent, la cellule d'écoute, les discussions qui partent dans tous les sens, les fous-rires, les afters, les sorties MHEvent et j'en passe. La dernière ligne droite aurait été bien plus compliquée sans vous tou(te)s.

Au groupe Isomet. À Pierre Millard pour ta cadence impressionnante pour le développement de l'outil, pour avoir pris le temps de m'expliquer et de me réexpliquer, pour ton soutien et pour m'avoir partagé tes tips de méthode de travail (et on peut aussi ramener du chocolat quand ça marche). À l'équipe de l'ETH de Zürich, Patrick et Uwe pour m'avoir rapidement intégrée dans ce projet, et avoir été indulgents face à mon anglais parfois aléatoire. A Clément, pour ta gentillesse, ta disponibilité, et pour avoir réussi à vulgariser et me faire comprendre ton langage codé.

À Sergueï, merci pour ta réactivité, pour ta patience, pour ton aide précieuse sur ces derniers mois et pour avoir sauvé mes environnements.

Merci à mes amis, trop nombreux pour que je les cite tous. Un remerciement tout particulier aux Zalous, aux Fromavins, aux Skirouettes, à la team des crousti pour les verres de soutien surtout en cette fin de rédaction.*

À ma famille, La Tribu, même si c'est loin de vos domaines respectifs, merci d'avoir été présents et de vous être accrochés.

À mon Rayon de soleil,

À tous, merci pour cette soirée magique d'afterthèse, agréable mélange de mondes et d'univers.

Résumé

Cette thèse s'inscrit dans un cadre de biologie des systèmes métaboliques et a pour objectif général l'identification, au sein d'une cellule ou d'un organisme, du réseau métabolique actif dans un contexte physiologique donné. Si la métabolomique permet d'identifier l'ensemble des métabolites d'un organisme - et ainsi de caractériser l'état de son système métabolique – elle ne renseigne pas directement sur la topologie et l'activité du réseau. La fluxomique est quant à elle basée sur des expériences de traçage isotopique qui permettent d'identifier les voies métaboliques actives et, à l'aide de modèles mathématiques adaptées, de déterminer les vitesses réelles des réactions (flux métaboliques) et ainsi d'accéder à la dynamique métabolique des cellules ou des organismes. Les approches de fluxomique ont jusque-là essentiellement reposé sur une connaissance préalable du système métabolique étudié (de façon à le modéliser) et sur des données de marquage acquises sur des composés identifiés du réseau (approches dites ciblées), et sont souvent focalisées sur l'étude du métabolisme carbone central. Le développement récent d'approches non ciblées de mesure d'incorporation isotopique par spectrométrie de masse permet cependant d'élargir considérablement la couverture du réseau métabolique accessible et offre également la possibilité de mettre en place de nouvelles approches de fluxomique.

Dans ce contexte, l'objectif spécifique de cette thèse est d'initier la mise en place d'une approche de reconstruction métabolique *ab initio*, qui vise à reconstruire le réseau actif d'un organisme dans un contexte donné, sur la seule base de données expérimentales de marquage isotopique non-ciblées. Dans un premier temps, des outils de traitement non-ciblé des données de marquage obtenues par spectrométrie de masse haute résolution ont été développés. Cela inclut une méthodologie permettant d'évaluer et d'assurer la qualité des données isotopiques ainsi générées. Dans une deuxième partie, une stratégie de fluxomique *ab initio* a été initiée. Elle repose sur l'obtention de données cinétiques d'incorporation de marquage isotopique et sur le développement d'approches de simulation de ces cinétiques de marquage pour aboutir à la construction de sous-réseaux métaboliques actifs au sein desquels la distribution des flux est résolue.

Abstract

This thesis is included into the framework of metabolic systems biology and aims at the identification, within a cell or an organism, of the metabolic network that is actually active in a given physiological context. If metabolomics makes it possible to identify all the metabolites of an organism - and thus to characterize the state of the metabolic system - it does not directly provide information on the topology and the dynamics of the network. Fluxomics, which is based on isotopic tracing experiments, makes it possible to identify the active metabolic pathways and, using suitable mathematical models, to determine the actual rates of the biochemical reactions (metabolic fluxes) and thus to access the metabolic dynamics of cells or organisms. Fluxomics approaches have until now essentially relied on prior knowledge of the studied metabolic system (in order to model it) and on labeling data collected for identified compounds in the network (so-called targeted approaches), and are often focused on the investigation of central carbon metabolism. The recent development of non-targeted approaches for measuring isotopic incorporation by mass spectrometry, however, considerably expands the coverage of the accessible metabolic network and also offers the possibility of implementing new fluxomics approaches.

In this context, the specific objective of this thesis is to initiate the implementation of an *ab initio* metabolic network reconstructionapproach, which aims to reconstruct the contextual metabolic network of an organism without a priori considerations, i.e. based only on a data-driven approach using non-targeted isotopic labeling experiments. First, tools for the non-targeted processing of the labeling data obtained by high-resolution mass spectrometry were developed. This includes a methodology to assess and ensure the quality of the measured isotopic data. In a second part, an *ab initio* fluxomics strategy was initiated. It is based on dynamic isotope-labeling experiments coupled to mathematical tools for simulating the kinetics of label incorporation into the detected metabolites, lead to the construction of active metabolic sub-networks within which the flux distribution can be established.

Communications

An optimization method for untargeted MS-based isotopic tracing investigations of metabolism. Communication orale, 14èmes JS du RFMF à Aussois, 23 novembre 2021

Development of an untargeted approach based on isotope profiling of metabolic networks by high resolution mass spectrometry, Butin N, Heuillet M, Bergès C, Guionnet M, Peyriga L, Portais JC, Bellvert F. Poster présenté lors du congrès Metabomeeting à Toulouse, du 22 au 24 janvier 2020

Development of an untargeted approach based on isotope profiling of metabolic networks by high resolution mass spectrometry, Butin N, Heuillet M, Bergès C, Guionnet M, Peyriga L, Portais JC, Bellvert F. Poster présenté lors des 12èmes JS du RFMF à Clermont-Ferrand, du 21 au 23 mai 2019 - Prix du poster toutes catégories

A combined metabolomics and lipidomics approach enable the stratification of acute-on-chronic liver failure patients according to their severity, Communication orale, 12èmes JS du RFMF à Clermont-Ferrand, du 21 au 23 mai 2019

Publications scientifiques

** Publications en lien avec le travail de thèse*

*** Data-driven ^{13}C-fluxomics towards *ab initio* reconstruction of metabolic networks.** **Butin N**, Millard P, Frainay C, Legregam L, Jourdan F, Schmitt U, Bellvert F, Kiefer P, Portais JC. *Manuscript in preparation*

Mosquito sex and mycobiota contribute to fructose metabolism in the Asian tiger mosquito Aedes albopictu. Guégan M, Martin E, Tran Van V, Fel B, Hay AE, Simon L, **Butin N**, Bellvert F, Haichar FEZ, Valiente Moro C.. Microbiome. 2022 Aug 30;10(1):138.. PMID: 36038937; PMCID: PMC9425969.s doi: 10.1186/s40168-022-01325-9

*** An optimization method for untargeted MS-based isotopic tracing investigations of metabolism.** **Butin N**, Bergès C, Portais JC, Bellvert F. Metabolomics. **2022** Jun 16;18(7):41. PMID: 35713733; PMCID: PMC9205802. doi: 10.1007/s11306-022-01897-5.

*** Exploring the Glucose Fluxotype of the E. Coli y-Ome Using High-Resolution Fluxomics.** Bergès, C. & Cahoreau, E.; Millard, P.; Enjalbert, B.; Dinclaux, M.; Heuillet, M.; Kulyk, H.; Gales, L.; **Butin, N.**; Chazalviel, M.; et al. *Metabolites* **2021**, *11*, 271, doi:10.3390/metabo11050271.

Blood metabolomics uncovers inflammation-associated mitochondrial dysfunction as a potential mechanism underlying ACLF. Moreau R, Clària J, Aguilar F, Fenaille F, Lozano JJ, Junot C, Colsch B, Caraceni P, Trebicka J, Pavesi M, Alessandria C, Nevens F, Saliba F, Welzel TM, Albillos A, Gustot T, Fernández J, Moreno C, Baldassarre M, Zaccherini G, Piano S, Montagnese S, Vargas V, Genescà J, Solà E, Bernal W, **Butin N**, Hautbergue T, Cholet S, Castelli F, Jansen C, Steib C, Campion D, Mookerjee R, Rodríguez-Gandía M, Soriano G, Durand F, Benten D, Bañares R, Stauber RE, Gronbaek H, Coenraad MJ, Ginès P, Gerbes A, Jalan R, Bernardi M, Arroyo V, Angeli P; CANONIC Study Investigators of the EASL Clif Consortium; Grifols Chair; European Foundation for the Study of Chronic Liver Failure (EF Clif)**..** J Hepatol. **2020** Apr;72(4):688-701. Epub 2019 Nov 25. Erratum in: J Hepatol. 2020 Jun;72(6):1218-1220. PMID: 31778751. doi: 10.1016/j.jhep.2019.11.009.

Orchestration of Tryptophan-Kynurenine Pathway, Acute Decompensation, and Acute-on-Chronic Liver Failure in Cirrhosis. Clària J, Moreau R, Fenaille F, Amorós A, Junot C,

Gronbaek H, Coenraad MJ, Pruvost A, Ghettas A, Chu-Van E, López-Vicario C, Oettl K, Caraceni P, Alessandria C, Trebicka J, Pavesi M, Deulofeu C, Albillos A, Gustot T, Welzel TM, Fernández J, Stauber RE, Saliba F, **Butin N**, Colsch B, Moreno C, Durand F, Nevens F, Bañares R, Benten D, Ginès P, Gerbes A, Jalan R, Angeli P, Bernardi M, Arroyo V; CANONIC Study Investigators of the EASL Clif Consortium, Grifols Chair and the European Foundation for the Study of Chronic Liver Failure (EF Clif). Hepatology. 2019 Apr;69(4):1686-1701. Epub 2019 Mar 19. PMID: 30521097. doi: 10.1002/hep.30363.

Inhibition of central de novo ceramide synthesis restores insulin signaling in hypothalamus and enhances β-cell function of obese Zucker rats. Campana M, Bellini L, Rouch C, Rachdi L, Coant N, **Butin N**, Bandet CL, Philippe E, Meneyrol K, Kassis N, Dairou J, Hajduch E, Colsch B, Magnan C, Le Stunff H. Mol Metab. **2018** Feb;8:23-36.. Epub 2017 Nov 7. PMID: 29233519; PMCID: PMC5985020. doi: 10.1016/j.molmet.2017.10.013

Lipoprotein lipase in hypothalamus is a key regulator of body weight gain and glucose homeostasis in mice. Laperrousaz E, Moullé VS, Denis RG, Kassis N, Berland C, Colsch B, Fioramonti X, Philippe E, Lacombe A, Vanacker C, **Butin N**, Bruce KD, Wang H, Wang Y, Gao Y, Garcia-Caceres C, Prévot V, Tschöp MH, Eckel RH, Le Stunff H, Luquet S, Magnan C, Cruciani-Guglielmacci C.. Diabetologia. **2017** Jul;60(7):1314-1324.. Epub 2017 Apr 29. PMID: 28456865. doi: 10.1007/s00125-017-4282-7

Table des matières

Table des illustrations 11

Index des tableaux 13

Abréviations 14

Chapitre 1 : Introduction générale

1. Préambule 18

2. Un cadre de biologie des systèmes métaboliques 19

PARTIE I : Utilisation des approches isotopiques pour l'étude des systèmes métaboliques 22

3. Importance des isotopes pour l'analyse du métabolisme 22

3.1. Les isotopes 22

3.2. Espèces isotopiques d'une molécule 24

3.3. Notation des espèces isotopiques 25

3.4. Exploitation des isotopes pour l'étude du métabolisme 26

3.5. Mesure des profils de marquage isotopique 27

4. Métabolomique assistée par les isotopes stables 45

4.1. Les standards isotopiques pour la métabolomique 45

4.2. Stratégies d'utilisation des standards isotopiques 48

5. Etude du métabolisme par traçage isotopique 54

5.1. Protocole expérimental 55

5.2. Applications des approches de traçage isotopique 68

PARTIE II : Réseaux métaboliques : de la reconstruction à la modélisation pour le calcul de flux 77

6. Les réseaux métaboliques 77

6.1. Introduction aux réseaux métaboliques 77

6.2. Reconstruction des réseaux métaboliques 79

6.3. Visualisation des réseaux métaboliques 91

6.4. Analyse topologique des réseaux métaboliques 92

6.5. Modélisation du métabolisme 94

7. Fluxomique par marquage isotopique 104

7.1. Principe du calcul de flux 106

7.2. Stratégies d'analyses de flux 108

7.3. Avantages et limites 114

7.4. Couplage de modèles 114

8. Objectifs de recherche 115

References 116

Chapitre 2 : An optimization method for untargeted MS-based isotopic tracing investigations of metabolism.

1. Introduction 135

2. Experimental section 137

2.1. Preparation of biological samples 137

2.2. LC/MS measurements 138

2.3. Data processing 139

2.4. Evaluation criteria for processing optimization 141

3. Results and discussion 143

3.1. Overall strategy and case study 143

3.2. Definition of reference sets for optimization 145

3.3. Optimization of isotopologue extraction 147

3.4. Optimization of isotopologue clustering 152

3.5. Application to the case study 154

4. Conclusion 157

References 158

Supplementary Data 162

Chapitre 3 : Data-driven ^{13}C-fluxomics towards *ab initio* reconstruction of metabolic networks

1. Introduction 186

2. Results 187

2.1. General strategy 187

2.2. Construction of minimal subnetworks 190

2.3. Implementation of IsoMet : Isotopic driven Metabolic reconstruction 193

2.4. In silico validation of the Network Analyzer 196

3. Discussion 200

4. Methods 201

4.1. Generation of the theoretical dataset 201

4.2. GSAM 201

4.3. Network Analyzer 202

References 203

Supplementary Data 205

Chapitre 4 : Exploring the Glucose Fluxotype of the E.coli y-ome Using High-Resolution Fluxomics.

1. Introduction 222

2. Results 223

2.1. Selection of E.coli y-ome strains 223

2.2. Integrated workflow for high-throughput collection ofhigh-resolution fluxotypes 225

2.3. Design of fluxomics experiments 225

2.4. High-Resolution fluxotyping workflow validation 226

2.5. High-Resolution glucose fluxotyping of 180 selected y-gene mutant strains 228

2.6. Glucose fluxotypes of ΔybjP and ΔydcS strains 230

2.7. Cofactor usage in the ΔybjP and ΔydcS strains 232

2.8. Scope and quality of high-throughput fluxomics investigations 233

3. Discussion 235

4. Materials and Methods 238

References 244

Supplementary Data 248

Chapitre 5 : Conclusions et perspectives 250

Annexes 254

Table des illustrations

Figure 1.1 : Cascade des sciences "-omiques" 20
Figure 1.2 : Distribution naturelle des isotopes stables 23
Figure 1.3 : Formes isotopiques du carbone pour une molécule contenant quatre atomes de carbone..25
Figure 1.4 : Stratégies d'utilisation des isotopes stables pour l'étude du métabolisme.....26
Figure 1.5 : Spectre du CO_2 dans l'air obtenu à l'aide du spectromètre infrarouge isotopique Delta Ray..
.....31
Figure 1.6 : Principe de la spectroscopie RAMAN appliquée aux approches isotopiques.. 32
Figure 1.7 : Informations isotopiques obtenues par analyse RMN 33
Figure 1.8 : Principe schématisé du fonctionnement d'un spectromètre de masse. 33
Figure 1.9 : Principe de la source électrospray.....37
Figure 1.10 : Principe de fonctionnement de la source DESI. 38
Figure 1.11 : Imagerie cellulaire par NanoSIMS pour mesurer les flux de carbone 42
Figure 1.12 :Spectres de masse du fumarate 43
Figure 1.13 : Impact de la contribution des isotopes présents à l'abondance naturelle sur les spectres de masse haute et basse résolution. 44
Figure 1.14 : Stratégies d'utilisation du Triangle de Pascal pour l'évaluation des données isotopiques
.....49
Figure 1.15 : Principe de détermination du nombre d'atomes de carbone dans une molécule 50
Figure 1.16 : Principe du logiciel FragExtract 51
Figure 1.17 : Quantification des métabolites par la méthode IDMS 53
Figure 1.18 : Principe général des approches de traçage isotopique. 55
Figure 1.19 : Protocole des expériences de marquage isotopique C13 par spectrométrie de masse 56
Figure 1.20 : Exemples de systèmes de culture pour les cellules en suspension.....59
Figure 1.21 : Illustration du mode de fragmentation MRM.. 62
Figure 1.22 : Ordre de grandeur des données générées en LC/MS. 63
Figure 1.23 : Comparaison des étapes de traitement de données HRMS issues d'expériences de marquage via des analyses ciblées et non-ciblées. 64
Figure 1.24 : Spectres de masse d'un composé non marqué et marqué au C13 et principe du groupement isotopique 66
Figure 1.25 : Intérêt du traçage isotopique pour la mesure de la contribution de plusieurs sources de carbone à la formation d'un métabolite spécifique. 70
Figure 1.26 : Identification des 3 principales voies glycolytiques par traçage isotopique. 71
Figure 1.27 : Intérêt du principe de dilution isotopique pour mesurer des partitions de flux.....73
Figure 1.28 : Exemples d'applications de la partition de flux.....74
Figure 1.29 : Réseau métabolique et principales voies du métabolisme central et energétique.....78

Figure 1.30 : Processus de reconstruction de réseau métabolique à l'échelle du génome 81
Figure 1.31 : État des réseaux métaboliques humains génériques et contextuels 84
Figure 1.32 : Méthodes de reconstruction expérimentales. 85
Figure 1.33 : Schématisation du concept des réseaux basés sur les différences de masse 87
Figure 1.34 : Les principaux types de graphes représentant les réseaux métaboliques.....93
Figure 1.35 : Exemple de décomposition d'un réseau en modes élémentaires de flux. 94
Figure 1.36 : Approches de modélisation métabolique. 97
Figure 1.37 : Exemples de motifs de réseau métabolique non accessibles par une analyse MFA 101
Figure 1.38 : Illustration de la modélisation basée sur les contraintes 102
Figure 1.39 : Protocole des approches de fluxomique.....104
Figure 1.40 : Principe général de l'analyse ^{13}C-MFA. 105
Figure 1.41 : Carte de flux.....108
Figure 1.42 : Comparaison des différentes approches d'analyse des flux par marquage isotopique ...109
Figure 1.43 : Stratégies de marquage en parallèle (analyse COMPLETE-MFA) 113
Figure 2.1 : Strategy for software parameter optimization in untargeted MS-based isotopic profiling using a reference labelled material. 144
Figure 2.2 : Proposed strategy for the two-step optimization of data processing in untargeted MS-based isotopic tracing studies 147
Figure 2.3 : Impact of parameter optimization on the measurement of isotopologue abundances 151
Figure 2.4 : Quality of isotopologue clustering.. 153
Figure 2.5 : Comparison of isotopic clusters using X13CMS and geoRge. 156
Figure 3.1 : General strategy for ab initio metabolic subnetwork construction based on untargeted ^{13}C-labeling dynamic experiments. 189
Figure 3.2 : Construction and evaluation of minimal subnetworks. 192
Figure 3.3 : Computational workflow of IsoMet approach 195
Figure 3.4 : Validation of the Network Analyzer 199
Figure 4.1 : Strategy for high-resolution fluxotyping of the E. coli y-ome. 224
Figure 4.2 : Metabolic fluxes measured for E. coli WT strains and a Δzwf mutant strain. 227
Figure 4.3 : Comparison of the growth parameters and fluxotypes of y-ome strains with those of their parental strain 229
Figure 4.4 : Distribution of metabolic fluxes in ΔydcS and ΔybjP strains.....231
Figure 4.5 : Production of NADPH, NADH/$FADH_2$ and ATP in the central carbon metabolism. 232

Index des tableaux

Tableau 1.1 : Performance des principaux analyseurs de masse utilisés en métabolomiques.....41
Table 2.1 : Cluster precision and recall for X13CMS and geoRge 154
Table 2.2 : Impact of parameter optimization on the extraction of data from the E. coli WT samples.. 155
Table 3.1 : Results of validation process.....197
Table 4.1 : Evaluation of data size and quality in HT fluxomics investigations 234

Abréviations

1-9

2/3-PG : 2/3 Phosphoglycérate
6PGL : 6-phosphogluconolactone
6PGC : 6-phosphogluconate

A

ADP : Adenosine diphosphate
AKG : Alpha-ketoglutarate
ATP : Adenosine triphosphate

B

BM : Biomasse

C

CBM : Constraint Based Model
CID : Carbon Isotopologue Distribution
COMPLETE-MFA : complementary parallel labeling experiments technique for metabolic flux analysis

D

DHAP : Dyhydroxyacétone phosphate
DIMS : Direct Injection Mass Spectrometry

E

E4P : Erythrose-4-phosphate
E. coli : Escherichia coli
ESI : Electrospray ionisation

F

F6P : Fructose 6-phosphate
(r)FBA : (regulatory) Flux Balance Analysis
FBP : Fructose 1,6-bisphosphate
FVA : Flux Variability Analysis

G

G6P : Glucose 6-phosphate
GC : Chromatographie gazeuse
GSM : Genome Scale Model
GSMN : Genome Scale Network

I

IDMS : Isotopic Dilution Mass Spectrometry
IRIS : Isotope Ratio Infrared Spectroscopy
IROA : Isotopic Ratio Outlier Analysis

L

LC : Chromatographie liquide

M

(^{13}C)-MFA : Metabolic Flux Analysis
MRM : Multiple Reaction Monitoring
MS(/MS) : (Tandem) Mass Spectrometry

N

NAD(H) : Nicotinamide adénine dinucléotide phosphate
NADP(H) : Nicotinamide adénine dinucléotide phosphate
NanoSIMS : Nanoscale Secondary Ion Mass Spectrometry

P

PEP : Phosphoenolpyruvate
PT : Triangle de Pascal

R

R5P : Ribose-5-phosphate
RMN : Résonance Magnétique Nucléaire
RU5P : Ribulose-5-phosphate
RT : Temps de rétention

S

S7P : Sedoheptulose
SUCCOA : Succinyl-CoA

X

X5P : Xylulose-5-phosphate

Chapitre 1

Introduction générale

1. Préambule

Cette thèse s'inscrit dans un cadre de biologie des systèmes métaboliques et s'appuie sur l'utilisation d'approches de métabolomique et de fluxomique pour appréhender l'étude des réseaux métaboliques. Ces deux approches « omiques » permettent de caractériser de manière très détaillée l'état (métabolomique) et la dynamique (fluxomique) du métabolisme à l'échelle cellulaire, tissulaire ou de l'organisme entier. Elles sont utilisées de façon croissante dans un très grand nombre de domaines des sciences de la vie, mais aussi de la santé, de l'agronomie et des biotechnologies. Toutefois, elles restent toutes deux encore très difficiles à mettre en œuvre car elles reposent chacune sur la combinaison d'approches expérimentales et analytiques complexes et d'outils de traitement et d'interprétation des données qui le sont tout autant. Ce constat est particulièrement vrai pour la fluxomique, qui repose sur des expériences de traçage isotopique associées à la mesure d'incorporation de l'isotope dans les métabolites par spectrométrie de masse ou par RMN et au développement de modèles mathématiques sophistiqués. Dans ce contexte, l'objectif général de cette thèse est d'étendre la capacité actuelle d'investigation des sytèmes métaboliques par le développement de méthodes de fluxomique originales. Les travaux développés concernent en premier lieu l'acquisition et le traitement non ciblé de données isotopiques de façon à étendre la couverture du réseau métabolique accessible à l'analyse des flux, puis le développement d'une approche originale de fluxomique, dite fluxomique *ab initio*, qui a pour objectif final la reconstruction de réseaux métaboliques sur la base de données de marquage isotopique. Enfin, j'ai également participé à un travail portant sur le développement de la fluxomique haut-débit.

Ce travail a été réalisé au sein du laboratoire RESTORE (Université de Toulouse, Inserm U1031, CNRS 5070, UPS, EFS) et du plateau MetaToul-FluxoMet (Toulouse Biotechnology Institute, Université de Toulouse - CNRS 5504 - INRA 792 - INSA Toulouse), en collaboration avec l'équipe MetaSys (Toulouse Biotechnology Institute), l'UMR 1331 Toxalim (INRAE/ENVT/INPT-EI Purpan/UPS) à Toulouse et l'ETH Zürich (équipe J. Vorholt) en Suisse.

2. Un cadre de biologie des systèmes métaboliques

Le métabolisme – défini de manière très large comme l'ensemble des processus de conversion de matière et d'énergie au sein d'un système vivant – joue un rôle fondamental dans les systèmes vivants en supportant l'ensemble des besoins énergétiques et synthétiques nécessaires à toutes les fonctions biologiques. L'étude du métabolisme présente un intérêt majeur pour la compréhension de son fonctionnement, de son rôle dans la physiologie et la physiopathologie, ainsi que dans l'adaptation des systèmes vivants.

L'étude du métabolisme est ancienne et a réposé très longtemps sur des approches réductionnistes visant à identifier et caractériser chacun des éléments (métabolites, enzymes, coenzymes, etc) qui le composent, à identifier et caractériser les différentes voies métaboliques, ou encore à identifier les éléments de régulation de ces voies. En opposition – ou plutôt en complémentarité – avec l'approche réductionniste, une vision systémique (ou holistique) de la biologie – et du métabolisme – s'est développée depuis deux décennies. Cette approche vise à appréhender le système dans son ensemble en essayant de comprendre comment tous les éléments du système interagissent entre eux pour conduire aux comportements (aux phénotypes) observés [Jamshidi and Palsson, 2006]. Dans cette vision le système (ici un organisme vivant) est représenté comme un ensemble d'éléments (gènes, protéines, métabolites …) qui interagissent entre eux de manière structurée (organisation en structures fonctionnelles) pour satisfaire une ou plusieurs fonctions (satisfaire un ou plusieurs besoins). L'organisation de chaque élément au sein de ces structures permet à l'organisme d'assurer l'accomplissement des fonctions vitales (survie, croissance, développement, etc) et de s'adapter à son environnement.

L'un des principaux défis de la biologie des systèmes est d'identifier le lien entre la structure du système (dans toute sa complexité) et ses capacités fonctionnelles. Pour cela elle s'appuie sur la combinaison d'approches expérimentales, notamment les approches dites « -omiques » (génomique, transcriptomique, protéomique, métabolomique, fluxomique) qui fournissent une analyse à large échelle des différents éléments du système et de leurs interactions (moléculaires, fonctionnelles, régulatrices) (Figure 1.1). Des approches d'analyse *in silico* de plus en plus élaborées permettent ensuite d'exploiter ces masses de données pour analyser, comprendre voire prédire le fonctionnement des processus biologiques.

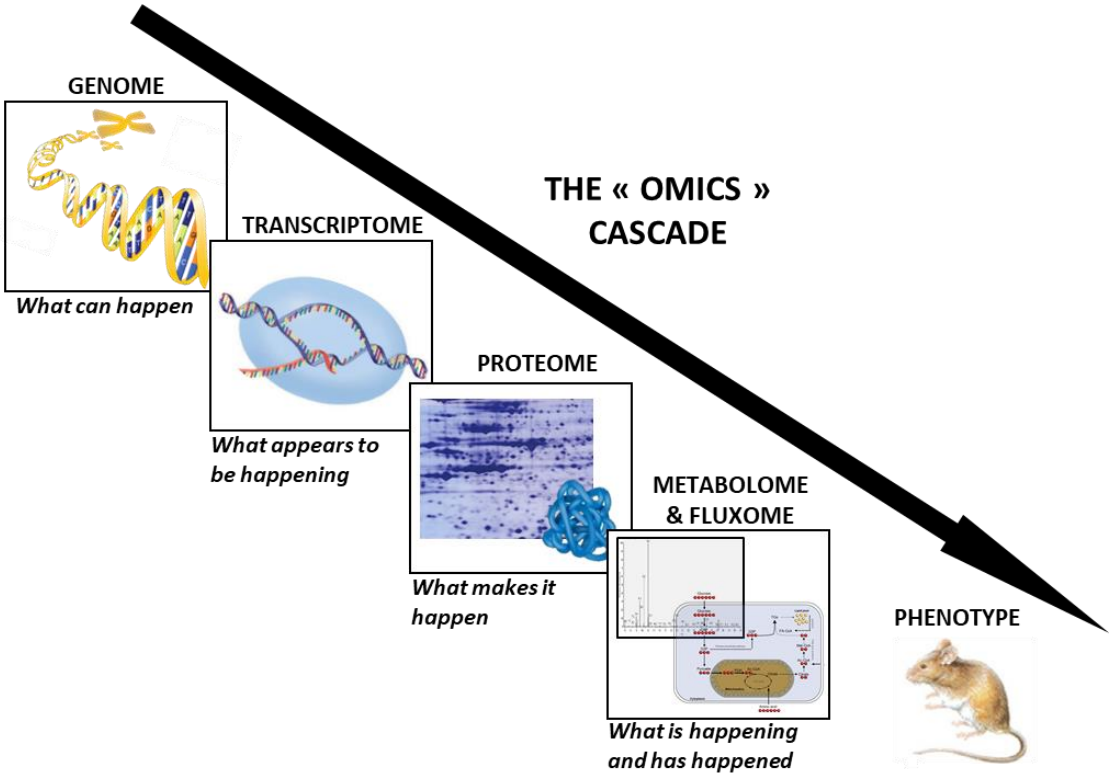


Figure 1.1 : Cascade des sciences "-omiques". Figure adaptée de Garcia-sevillano et al, 2014.

Pour en venir au cas particulier du métabolisme, on définira ici un système métabolique comme l'ensemble de tous les acteurs du métabolisme (métabolites, enzymes, coenzymes, effecteurs, régulateurs, etc) et l'ensemble des interactions entre eux qui déterminent le comportement métabolique d'une cellule, d'un tissu ou d'un organisme. Le système métabolique est organisé en réseau de réactions qui représente l'ensemble des interactions possibles entre les éléments du système. De manière globale, le réseau peut intégrer tout type d'interactions entre les acteurs du métabolisme, par exemple les interactions physiques (interactions protéines-protéines, protéines-ligands, etc), les interactions régulatrices (interactions enzymes-effecteurs, etc), etc. Cependant, l'objectif premier de l'étude des systèmes métaboliques est de comprendre dans toute sa complexité l'ensemble des processus biochimiques sur lequels reposent toutes les fonctions énergétiques et synthétiques nécessaires à l'organisme. Dans ce contexte, le réseau métabolique doit représenter avant tout l'ensemble des réactions métaboliques de l'organisme et toutes leurs interactions potentielles. La reconstruction du réseau métabolique, c'est-à-dire l'établissement de la carte métabolique la plus complète possible, et l'analyse aussi bien expérimentale qu'*in silico* de ce réseau, sont au cœur de la biologie des systèmes métaboliques. Les objectifs sont d'identifier et de comprendre la nature du réseau, son organisation et ses propriétés fonctionnelles (incluant les capacités d'adaptation) dans le contexte d'études fondamentales ou appliquées.

Dans ce cadre général de biologie des systèmes métaboliques, ce travail de thèse se situe à un double niveau. Il vise en premier lieu à développer des approches analytiques permettant d'accéder à l'information la plus large possible sur les réseaux métaboliques. Si toutes les approches omiques sont utilisées pour appréhender le métabolisme au niveau systémique, la métabolomique et la fluxomique constituent les approches les plus directes d'analyse du métabolisme et les plus proches du phénotype métabolique. La métabolomique vise à identifier et quantifier le métabolome, c'est-à-dire l'ensemble des petites molécules (<1500 Da) présentes dans la cellule, le tissu ou l'organisme, et permet de caractériser l'état du système métabolique dans un contexte particulier et à un moment donné. Elle permet d'établir le catalogue des métabolites présents et qui correspond à la partie des éléments du réseau actif dans la condition étudiée, mais ne permet pas d'identifier les réactions elle-mêmes. De plus, un même métabolite peut participer à plusieurs réactions et voies métaboliques. Le glutamate par exemple est lié à plus de 200 réactions biochimiques [Fan et al, 2012]. Il est donc impossible de discerner les contributions de chaque réaction ou chaque voie sur la seule base de l'analyse des métabolites. La fluxomique repose quant à elle sur des expériences de traçage isotopique qui permettent d'identifier les réactions ou les voies métaboliques impliquées dans l'utilisation des composés. Elle donne donc un accès plus direct aux réactions du réseau et à leurs interconnexions. Elle va au-delà en permettant de convertir l'information de marquage en valeurs de flux métaboliques grâce à des modèles mathématiques adaptés, donnant ainsi accès à la dynamique du métabolisme. Les deux approches sont très complémentaires l'une de l'autre, que ce soit pour identifier le réseau, ou pour caractériser l'état (métabolomique) et la dynamique (fluxomique) du système métabolique.

Dans le cadre de ce travail de thèse, nous avons travaillé sur le développement d'approches de profilage isotopique non ciblé par spectrométrie de masse, de façon à étendre la couverture du réseau métabolique pouvant être étudié en fluxomique. Puis, nous avons initié la mise en place d'une méthode de recontruction de fluxomique *ab initio* basée sur des données aquises par ce type d'approche expérimentale, dont l'objectif est de reconstruire des réseaux métaboliques et de mesurer les flux métaboliques dans ces réseaux à partir des données expérimentales. La suite de ce chapitre 1 comprend deux grandes parties en phase avec ces objectifs. La première est consacrée à l'utilisation des approches isotopiques pour l'étude des systèmes métaboliques, et la seconde aux réseaux métaboliques, de la reconstruction à la modélisation pour le calcul des flux.

PARTIE I : Utilisation des approches isotopiques pour l'étude des systèmes métaboliques

3. Importance des isotopes pour l'analyse du métabolisme

Nous allons voir dans cette partie de l'introduction comment l'utilisation du marquage isotopique se révèle être un outil puissant pour l'étude des sytèmes métaboliques en général et pour l'analyse fonctionnelle des réseaux métaboliques en particulier.

3.1. Les isotopes

Les isotopes sont des atomes d'un même élément chimique qui possèdent le même nombre de protons mais un nombre différent de neutrons. Les différents isotopes d'un même élément possèdent des propriétés chimiques identiques mais des propriétés physiques différentes. On distingue notamment les isotopes stables des isotopes instables (ou radioactifs), qui subissent une désintégration radioactive pour se transformer en élément plus stable, c'est à dire en atome différent. L'isotope qui se transforme et émet un rayonnement porte le nom de radio-isotope. A titre d'exemple, le carbone possède trois isotopes principaux: le carbone-12 et le carbone-13 qui sont des isotopes stables, et le carbone-14 qui est un isotope radioactif principalement connu pour son utilisation en datation radiométrique.

L'utilisation d'isotopes pour l'analyse du métabolisme a été introduite pour la première fois en 1935 lorsque Schoenheimer et Rittenberg ont utilisé du deutérium pour suivre le devenir d'acides gras chez la souris [Schoenheimer and Rittenberg, 1935 ; Klein and Heinzle al, 2012]. A la fin de la seconde guerre mondiale, l'utilisation des isotopes stables a diminué pour laisser la place aux radio-isotopes. Ceci s'explique par la disponibilité accrue des radio-isotopes et la facilité avec laquelle la radioactivité peut être détectée, permettant des analyses très sensibles de l'incorporation de l'isotope dans les molécules [Lee et al, 2010 ; Wilkinson, 2018]. L'incorporation de l'isotope dans les métabolites était suivie par l'apparition puis la disparition de la radioactivité dans les intermédiaires métaboliques au cours du temps. Cette technique est encore utilisée, notamment en médecine.

Avec l'essor des techniques analytiques de plus en plus fines, les isotopes radioactifs ont été remplacés par des isotopes stables. Par rapport à la radioactivité, ces nouvelles techniques ont offert l'avantage de fournir une information intramoléculaire sur la position (RMN) ou le nombre (spectrométrie de masse) d'isotopes incorporés dans chaque composé

détecté. Bien qu'ils soient principalement utilisés pour étudier les systèmes biologiques, les isotopes stables sont également exploités pour un large éventail d'applications comme l'environnement ou la géologie. Parmi les avantages qu'ils présentent, les isotopes stables ne présentent pas de danger pour le chercheur et l'environnement ce qui simplifie les procédures expérimentales (plus de purification ni de dégradation). Parce qu'ils sont désormais prédominants dans la métabolomique et la fluxomique contemporaines [Kim et al, 2016 ; Klein & Heinzle, 2012], seuls les isotopes stables sont considérés dans la suite de ce travail de thèse. On retrouve parmi les isotopes stables les plus communément utilisés en métabolomique et fluxomique les atomes suivants : ^{2}H, ^{13}C, ^{15}N, ^{18}O (Figure 1.2). L'isotope stable le plus couramment utilisé est le ^{13}C, qui permet de tracer le devenir des atomes de carbone dans le métabolisme [Wiechert, 2001]. La suite du travail de thèse se concentrera sur l'utilisation du carbone 13 comme isotope stable pour l'étude du métabolisme. Le principe d'utilisation des autres isotopes est similaire.

Element	Isotope	mass	Mass difference	Abundance (%)
Hydrogen	^{1}H	1.007825		99.985
	^{2}H	2.014102	+1.006277	0.015
Carbon	^{12}C	12.0		98.890
	^{13}C	13.003355	+1.003355	1.110
Nitrogen	^{14}N	14.003074		99.634
	^{15}N	15.000109	+0.997035	0.366
Oxygen	^{16}O	15.994915		99.762
	^{17}O	16.999132	+1.004217	0.038
	^{18}O	17.999161	+2.004246	0.200
Phosphor	^{31}P	30.973762		100
Sulfur	^{32}S	31.972071		95.020
	^{33}S	32.971459	+0.999388	0.750
	^{34}S	33.967867	+1.995796	4.210
	^{36}S	35.967081	+3.995010	0.020

Figure 1.2 : Distribution naturelle des isotopes stables : abondance relative et leur masses en Dalton

3.2. Espèces isotopiques d'une molécule

Une molécule marquée isotopiquement présente plusieurs formes ou espèces isotopiques. Les espèces isotopiques sont des molécules chimiquement identiques, ne différant que par leur composition isotopique. Dans les expériences de marquage au carbone 13, chaque molécule comportant *n* atomes de carbone possède un total de 2^n espèces isotopiques du carbone (Figure 1.3). Plusieurs termes sont utilisés pour décrire ces différentes espèces.

Selon les recommandations de l'IUPAC, le terme ***isotopomère*** (isomère isotopique) réfère aux isomères ayant le même nombre d'isotopes mais dont les positions diffèrent. Ainsi, la molécule $^{13}CH_3$-$^{12}COO^-$ est un isotopomère de la forme $^{12}CH_3$-$^{13}COO^-$.

De même, le terme IUPAC ***isotopologue*** renvoie à des entités moléculaires qui ne diffèrent que par leur composition isotopique (ici en carbone), c'est à dire en fonction du nombre de substitution isotopique. Les deux isotopomères $^{13}CH_3$-$^{12}COO^-$ et $^{13}CH_3$-$^{12}COO^-$ (qui possèdent 1 subtitution isotopique) sont des isotopologues de la forme $^{13}CH_3$-$^{13}COO^-$ (2 substitutions isotopiques). Une molécule comportant *n* atomes de carbone contient *n*+1 isotopologues du carbone notés du M0 (molécule non marquée) au Mn en fonction du nombre d'atomes ^{13}C incorporés dans la molécule.

A noter que les deux termes isotopomère et isotopologue s'excluent entre eux (deux isotopologues ne peuvent pas être isotopomères l'un de l'autre, et inversement). Il n'existe donc pas de terme conventionnel permettant de décrire l'ensemble des formes isotopiques d'une même molécule.

Dans la suite du manuscrit, le terme espèce (ou forme) isotopique sera utilisé pour décrire toute forme isotopique d'une molécule. Nous verrons que le terme isotopologue est pertinent pour les données obtenues par spectrométrie de masse (deux isotopologues n'ayant pas la même masse moléculaire) alors que le terme isotopomère est utilisé pour les données de RMN, qui donne accès à la position du marquage dans les molécules. Une molécule contenant quatre atomes de carbone présente $2^4 = 16$ espèces isotopiques et 4+1 = 5 isotopologues du carbone (Figure 1.3). Le nombre d'isotopomères varie suivant le nombre d'atome de carbone, mais obéit à la loi de distribution binomiale et correspond donc aux coefficients du triangle de Pascal.

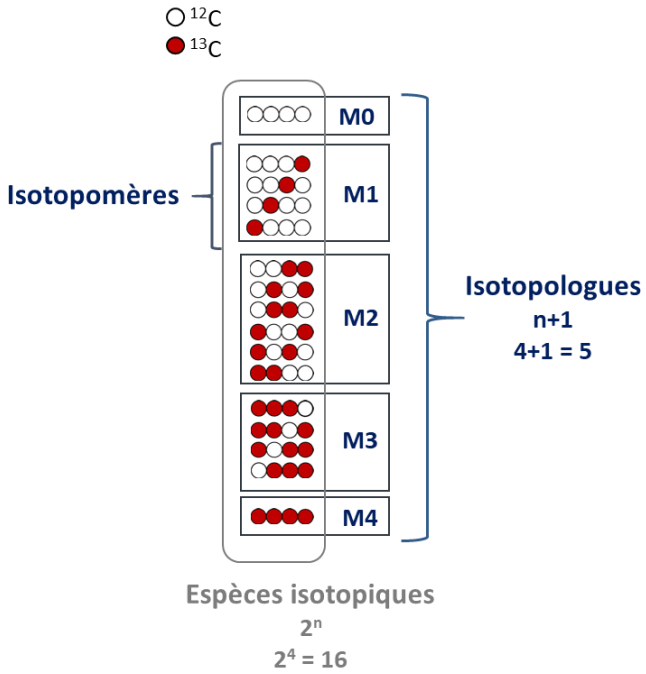


Figure 1.3 : Formes isotopiques du carbone pour une molécule contenant quatre atomes de carbone.

3.3.Notation des espèces isotopiques

Il existe un format conventionnel d'écriture pour une espèce isotopique donnée d'une molécule, dans lequel les espèces isotopiques et leur positions sont fournies entre crochets, par exemple [1-$^{13}C_1$]-glucose pour spécifier qu'une molécule de glucose contient en position 1 un atome de ^{13}C. Le symbole U est utilisé si tous les atomes d'un élément sont substitués isotopiquement (exemple : [U-$^{13}C_6$]-glucose). Enfin, cette notation est aussi utilisée lorsque plusieurs éléments sont substitués isotopiquement, par exemple [1-$^{13}C_1$, 2-$^{15}N_1$]-alanine pour une molécule d'alanine marquée au carbone-13 en position 1 et à l'azote-15 en position 2. L'écriture ci-dessus n'est valide que pour une seule espèce isotopique donnée.

Une notation alphanumérique est également utilisée, dans laquelle on écrit le contenu de chaque position carbonée par 0 (^{12}C) ou 1 (^{13}C) la substitution isotopique. Un symbole # est mis au début pour spécifier cette notation isotopique. Par exemple le [1-$^{13}C_1$]-glucose s'écrit glucose#100000 suivant cette notation. Celle-ci permet de regrouper les différentes espèces isotopiques par sous-familles, grâce au symbole X qui signifie ^{12}C ou ^{13}C. Par exemple glucose#X00000 pour les espèces de glucose non marquées sur les positions 2 à 6, ou encore glucose#1XXXXX pour regrouper toutes les espèces de glucose marquées en 1 quelque soit leur marquage sur les autres positions carbonées. Cette notation est très utilisée en RMN

(données positionnelles) mais aussi dans les outils de calcul des flux car elle est adaptée à un traitement informatique des données.

Les isotopologues sont généralement écrits simplement par le symbole Mi, i étant le nombre d'atome de ^{13}C incorporé dans la molécule (ci-dessus).

3.4.Exploitation des isotopes pour l'étude du métabolisme

Les isotopes stables peuvent être utilisés en appui à la métabolomique ou dans le cadre d'études de traçage isotopique (Figure 1.4).

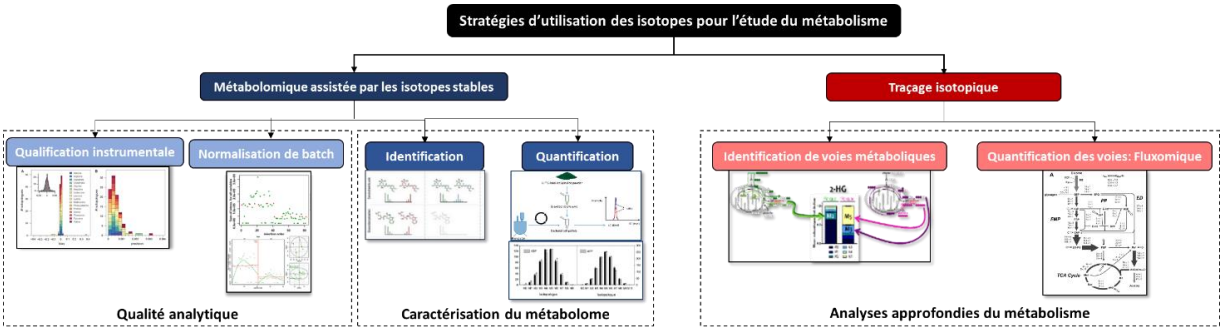


Figure 1.4 : Stratégies d'utilisation des isotopes stables pour l'étude du métabolisme.

La première application consiste à utiliser des standards marqués isotopiquement dans les approches de métabolomique. Ces standards marqués ont tout d'abord une valeur qualitative permettant d'assurer la qualité des données analytiques. Ils peuvent être utilisés à différents moments de l'étude : (i) en amont de l'expérience pour qualifier les instruments et méthodes analytiques [Heuillet et al, 2018] ou (ii) une fois les données acquises pour standardiser les batch analytiques et supprimer d'éventuels biais expérimentaux. Les standards isotopiques sont également utilisés pour caractériser plus finement le métabolome, en apportant une dimension isotopique supplémentaire permettant l'identification plus précise et/ou la quantification absolue des métabolites détectés [Chokkathukalam et al, 2014 ; Creek et al, 2012].

Le deuxième axe consiste à utiliser les isotopes stables comme traceurs isotopiques. Le traçage isotopique est une approche permettant de mesurer l'utilisation des voies métaboliques intracellulaires. Elle consiste à remplacer un (ou plusieurs) atomes (ex : ^{12}C) d'une molécule d'intérêt par leur équivalent isotopique stable (ex : ^{13}C) et de suivre le devenir de ces composés marqués au sein du métabolisme.

La production des standards et des traceurs isotopiques et leurs différentes stratégies d'utilisation sont détaillées dans les sections 4 et 5 de cette introduction.

3.5. Mesure des profils de marquage isotopique

La mesure des profils de marquage isotopique - ou profilage isotopique - qui seront utilisés dans les études du métabolisme requiert l'utilisation d'outils analytiques sensibles et d'outils de traitement de données robustes et automatiques. Ces différents outils doivent permettre d'analyser la diversité du métabolome (et du fluxome) dans toute sa complexité.

- ***Complexité du métabolome***

Le métabolome correspond à l'ensemble des métabolites présents dans un biofluide, un tissu ou une cellule d'un organisme vivant, ou plus généralement d'un échantillon biologique. Les métabolites interviennent dans de nombreuses réactions biochimiques et participent à des fonctions biologiques variées. L'analyse du métabolome est rendue complexe par différents facteurs liés à sa nature.

La première difficulté est liée à la diversité physico-chimique des métabolites. En effet les métabolites appartiennent à une large variété de classes chimiques telles que les acides aminés, les acides organiques, les nucléotides ou les lipides. Ces composés ont des structures et des propriétés physico-chimiques très diverses et ne peuvent par exemple pas être extraits ni analysés par une seule méthode. Cette diversité ne permet pas l'étude de l'ensemble des métabolites de manière simultanée. Il n'existe pas de méthode analytique universelle permettant d'explorer la totalité du métabolome. Pour couvrir l'ensemble du métabolome, il est donc nécessaire de combiner plusieurs méthodes d'analyse et plusieurs techniques analytiques.

De même, les métabolites varient également très fortement en termes de gamme de concentration. Cela s'étend du picomolaire (ex : vitamines) à quelques dizaines de millimolaire (ex : sucres ou lipides), soit une gamme dynamique de concentrations s'étalant sur 9-10 décades. Les techniques analytiques ayant généralement des gammes dynamiques de l'ordre de 3-4 décades, il est donc là aussi nécessaire d'utiliser différentes méthodes pour couvrir l'ensemble sur le plan quantitatif.

Une autre difficulté est directement liée au nombre de métabolites existants dans les organismes vivants. Les bases de données publiques recensent les métabolites identifiés, telles que la base de données « The Human Metabolome Database » (HMDB version 5.0 https://hmdb.ca) qui dénombre actuellement plus de 250 000 métabolites chez l'Homme [Wishart et al, 2022]. Ce nombre continue d'évoluer avec les récents développements des techniques de détection et d'annotation, par exemple le contenu d'HMDB a évolué de 2180 métabolites à 217 920 en une dizaine d'années. Cependant beaucoup de composés ne sont pas

encore clairement identifiés, ou ne sont pas détectés par les méthodes analytiques actuelles. A l'inverse, les méthodes analytiques actuelles permettent la détection d'un très grand nombre de signaux qui ne représentent pas forcément le nombre de métabolites, en raison de signaux parasites ou redondants (artefacts, adduits, produits de fragmentations). La détection de plusieurs centaines, voire milliers de signaux dans un échantillon biologique ne signifie pas la présence du même nombre de métabolites correspondants. Le nombre total de métabolites existants dans la nature ou dans un organisme, ainsi que le nombre de métabolites présent dans un échantillon donné, sont donc difficiles à estimer.

Une difficulté supplémentaire réside dans l'échelle du temps du métabolisme, au sein de laquelle se déroule des processus extrèmement rapides (de l'ordre de la micro-seconde ou moins pour certaines voies du métabolisme central) et des processus nettement plus lents (biosynthèse de certains coenzymes par exemple). Il est donc nécessaire de mettre en place des protocoles d'échantillonnage très rapides pour éviter une évolution non physiologique du métabolome avant son analyse. Lors de la préparation des échantillons, le protocole d'extraction peut nécessiter l'emploi de procédures particulières (ex : « quenching » à froid) permettant de stopper le métabolisme le plus rapidement possible. A titre de comparaison, l'échelle de temps de la protéomique est de plusieurs minutes à plusieurs heures, ce qui est moins stringent en termes de contraintes d'échantillonnage.

L'ensemble des spécificités énoncées ci-dessus rend l'analyse métabolomique délicate. De plus, le contenu en métabolites dépend également très fortement de l'organisme et des cellules ou tissus étudiés (cellules ou tissus mammifères, plantes, bactéries, levures, etc) et il est nécessaire de développer des stratégies expérimentales spécifiques à chaque type d'organisme – ou d'échantillon – envisagé.

Un très grand nombre de facteurs peuvent influencer cet état métabolique du système. Parmi ceux-ci figurent des facteurs intrinsèques (tels que l'âge, le sexe, … si on considère l'Homme) et des facteurs extrinsèques (alimentation, exposition à des stress, …). Comme l'état métabolique global va dépendre de l'ensemble de ces facteurs, la métabolomique est un bon indicateur des différentes influences (et perturbations) que subit l'organisme. Il est indispensable de planifier scrupuleusement le plan expérimental en amont de l'analyse métabolomique pour adapter l'ensemble des approches et techniques analytiques en fonction de la question biologique, de l'organisme et des composés d'intérêt.

- ***Complexité du fluxome***

Le fluxome représente l'ensemble des flux métaboliques dans une cellule, un tissu ou un organisme. Contrairement à la mesure du métabolome, qui représente un instantané dans le temps du contenu en métabolites d'un organisme, l'analyse du fluxome permet d'accéder à la dynamique du système (l'activité métabolique) en mesurant les vitesses réelles des réactions métaboliques dans le système biologique. Elle représente le résultat final de toutes les interactions métaboliques et régulatrices qui déterminent le comportement du métabolisme à l'échelle d'une cellule, d'un tissu ou d'un organisme.

La méthode la plus précise de mesure des flux métaboliques est basée sur la combinaison d'expériences de traçage isotopique et d'outils mathématiques permettant de calculer les vitesses des réactions à partir du marquage isotopique des métabolites. La fluxomique est ainsi basée sur des expériences de traçage isotopique. En terme analytique, l'ajout d'une dimension isotopique apporte une difficulté supplémentaire pour l'acquisition et la mesure des données qui seront exploitées pour analyser le fluxome. En effet, la présence d'isotopes va impacter les spectres recueillis en termes de nombre et d'intensité des signaux analytiques. Il est donc nécessaire de disposer des techniques analytiques performantes et sensibles ainsi que d'outils de calculs et de traitement de données puissants et robustes. Ces aspects seront détaillés plus loin dans l'introduction.

De manière globale, les approches intégratives de l'étude des "omiques" sont rendues compliquées par l'annotation fonctionnelle insuffisante et l'influence des facteurs environnementaux. De plus en plus, les approches dites multi-omiques sont utilisées dans divers domaines d'application afin d'obtenir des informations complémentaires sur le fonctionnement d'un organisme et son interaction avec son environnement. Par exemple, la combinaison des approches métabolomique et fluxomique va permettre de fournir des données complémentaires en apportant des informations sur la composition et l'abondance des métabolites ainsi que sur l'activité des voies métaboliques [Adebiyi et al., 2015 ; Jang et al, 2018].

3.5.1. Principales méthodes de mesure

Les techniques analytiques les plus couramment utilisées pour mesurer les isotopes stables ou suivre leur incorporation au sein des métabolites sont la Résonnance Magnétique Nucléaire (RMN) et la spectrométrie de masse (MS) mais d'autres techniques analytiques existent telles que l'imagerie ou la spectroscopie optique (IR, RAMAN) [Babele and Young, 2019]. Ces différentes méthodes d'analyse vont être introduites dans ce qui suit. La spectrométrie de masse, qui fait l'objet du travail, sera plus largement développée que les autres techniques.

3.5.1.1. Méthodes optiques

- ***Spectroscopie Infrarouge***

La spectroscopie infrarouge permet d'identifier des groupes fonctionnels au sein de molécules organiques. Elle consiste à mesurer l'absorption d'un rayonnement infrarouge (IR) envoyé à travers un échantillon. Ce rayonnement va être absorbé par les liaisons entre les molécules à différentes longueurs d'ondes. Le spectre résultant contient des pics qui représentent la vibration provoquée par l'absorption du rayonnement IR. Les liaisons entre molécules ont des fréquences de vibration uniques et caractéristiques. L'intensité de cette vibration dépend de trois facteurs principaux : la force de la liaison, la masse de l'atome et la longueur de la liaison. De ce fait les spectres sont sensibles à la subsitution isotopique dans la molécule, le déplacement isotopique étant plus grand proportionnellement aux écarts de masse atomiques entre les différents isotopes. Cette sensibilité peut alors être exploitée pour mesurer les rapports d'abondance des isotopes (Figure 1.5). La spectroscopie infrarouge à rapport isotopique (IRIS) est notamment utilisée pour mesurer la composition isotopique en ^{2}H, ^{18}O des eaux végétales et des eaux du sol [Martin-Gomez et al, 2015]. Elle est également utilisée en médecine pour diagnostiquer les maladies respiratoires en mesurant en temps réel les isotopes du CO_2 ($^{13}CO_2$, $^{12}CO_2$ et $^{18}OC^{16}O$) dans l'air expiré [Zhou et al, 2020 ; Sutehal et al, 2021]. L'IRIS a l'avantage d'être une technique peu onéreuse et qui nécessite peu de préparation d'échantillons.

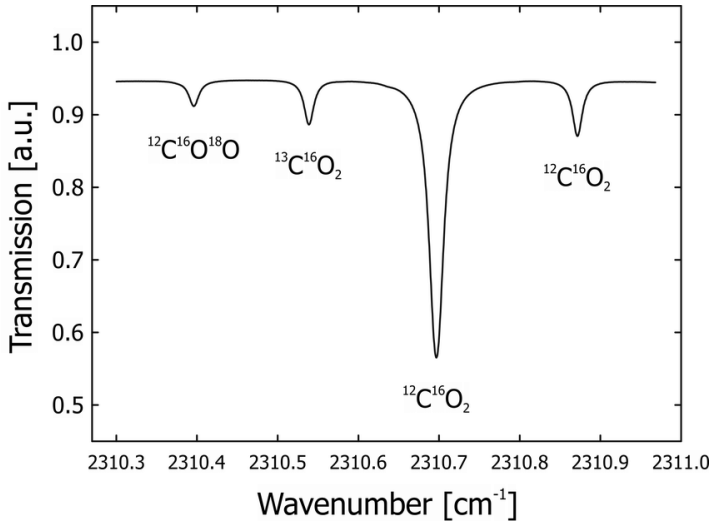


Figure 1.5 : Spectre du CO_2 dans l'air obtenu à l'aide du spectromètre infrarouge isotopique Delta Ray. Illustration tirée de Töchterle et al, 2017.

- ***Spectroscopie RAMAN***

La spectroscopie RAMAN est une technique analytique de spectroscopie vibrationnelle non destructive et non invasive qui exploite le phénomène physique selon lequel un milieu modifie légèrement la fréquence de la lumière y circulant. Cette technique est utilisée pour déterminer la structure de composés chimiques et a la capacité de détecter le marquage des cellules par des isotopes stables. Dans le cadre d'expériences de marquage, lorsque les atomes sont substitués par leur équivalent isotopique stable, le changement de la structure chimique de la molécule est négligeable et l'intensité des bandes de vibration correspondantes reste la même. Cependant la fréquence de vibration des liaisons concernées est affectée de manière significative. Dans le cadre d'expériences de marquage au carbone 13, la bande RAMAN correspondante sera décalée vers le rouge dans le spectre (Figure 1.6). La valeur de décalage de cette bande dépend de la valeur exacte de la variation de masse observée. Cette technique a été appliquée à divers domaines en microbiologie [Kubryk et al, 2015] et utilisée pour explorer et déterminer l'activité de voies métaboliques in vivo (ex : chez la levure [Noothalapati and Shigeto, 2014]) et apporter des informations sur la structure globale du réseau métabolique d'un organisme [Noothalapati and Shigeto, 2014]. Bien que de nombreux développements voient le jour : microspectroscopie [Kubryk et al, 2015], la spectroscopie RAMAN reste une méthode peu sensible avec des temps d'acquisition longs.

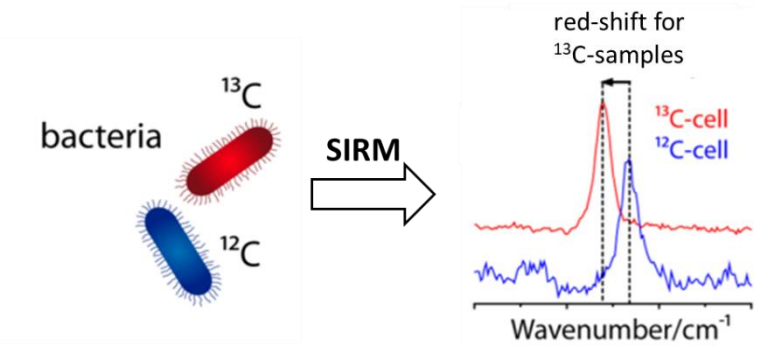


Figure 1.6 : Principe de la spectroscopie RAMAN appliquée aux approches isotopiques. Spectres Raman d'isotopes stables illustrant le décalage de la bande RAMAN à partir d'échantillon E. coli cultivées dans du glucose C12 et C13. SIRM signifie Stable Isotope Raman Microspectroscopy. Illustration tirée de Kubryk et al, 2015.

3.5.1.2. Résonnance Magnétique Nucléaire

La RMN est l'une des premières techniques analytiques utilisées pour l'analyse du métabolome [Nicholson and Wilson, 1989]. Cette technique est basée sur les propriétés magnétiques de certains noyaux atomiques. Pour les applications en métabolomique, on retrouve la RMN du ^{1}H, du ^{31}P, du ^{13}C ou du ^{15}N. Cette technique a largement été utilisée pour de l'identification moléculaire et l'élucidation structurelle [Emwas et al, 2019]. Elle est également largement utilisée en isotopie pour le traçage isotopique (^{13}C mais aussi ^{15}N par exemple), la quantification du marquage isotopique des molécules et l'analyse de flux [Lane and Fane, 2017]. La RMN offre la spécificité d'obtenir directement une information sur la position du marquage, puisque chaque atome de carbone d'une molécule donne un signal spécifique. Deux types d'informations isotopiques complémentaires peuvent ainsi être obtenues par une analyse RMN : les isotopomères (à noter que le terme est utilisé de manière abusive ici par rapport à sa définition IUPAC, pour décrire différentes combinaisons de marquage sur tout ou partie du squelette carboné) et l'enrichissement spécifique, qui est le pourcentage de ^{13}C incorporé dans une position carbonée particulière de la molécule (Figure 1.7). Hormis cet aspect, les principaux avantages de la RMN reposent sur le fait qu'il s'agit d'une technique non destructive, quantitative [Markley et al, 2017] et robuste [Bingol and Brüschweiler, 2015] nécessitant peu de préparation d'échantillon. Son caractère non destructif est à la base de son exploitation pour des études *in vivo* ou *in situ* du métabolisme sur des cellules, des tissus ou des organismes vivants (y compris l'Homme) au moment de l'analyse.

Même si la RMN offre beaucoup d'avantages, sa faible sensibilité en comparaison à la spectrométrie de masse reste une de ses principales limites pour les approches de fluxomique [Dunn et al, 2013 ; Giraudeau, 2020]. Par ce fait, elle est souvent restreinte à l'analyse de

composés majoritaires s'accumulant en quantité suffisante pour obtenir des signaux exploitables. Elle reste ainsi principalement utilisée pour mesurer la consommation des nutriments et la production de composés terminaux (ex : produits métaboliques tels que lactate, acétate, etc) et pour la mesure des flux extracellulaires. Néanmoins, les avancées technologiques (cryosondes, microsondes) ont permis d'augmenter la sensibilité de l'analyse RMN. Le développement de techniques 2D dans lesquelles les signaux sont éclatés dans deux dimensions spectrales ont également permis de simplifier l'interprétation de spectres d'échantillons complexes.

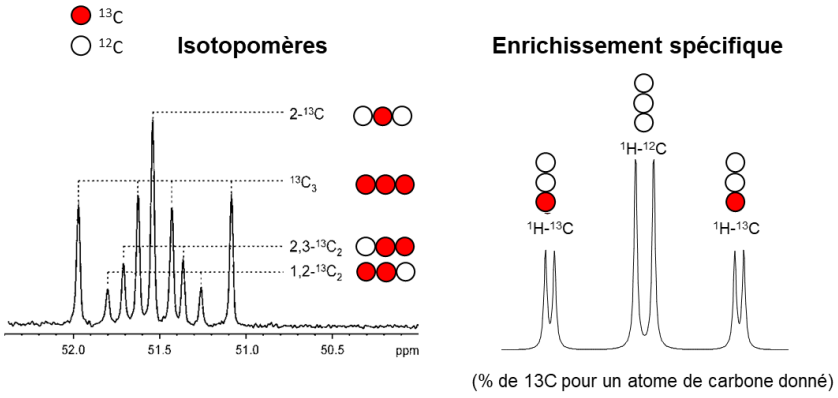


Figure 1.7 : Informations isotopiques obtenues par analyse RMN

3.5.1.3. Spectrométrie de masse

La spectrométrie de masse est une technique analytique qui permet de séparer les molécules contenues dans un échantillon biologique en fonction de leurs masses moléculaires, après ionisation. Le spectromètre de masse est composé d'une source d'ionisation, d'un analyseur qui sépare les ions formés selon leur rapport masse sur charge (m/z), et d'un détecteur comptabilisant l'abondance des ions (Figure 1.8). Un traitement informatique permet de visualiser les résultats sous forme de spectres de masse représentant l'abondance des ions en fonction de leur rapport m/z.

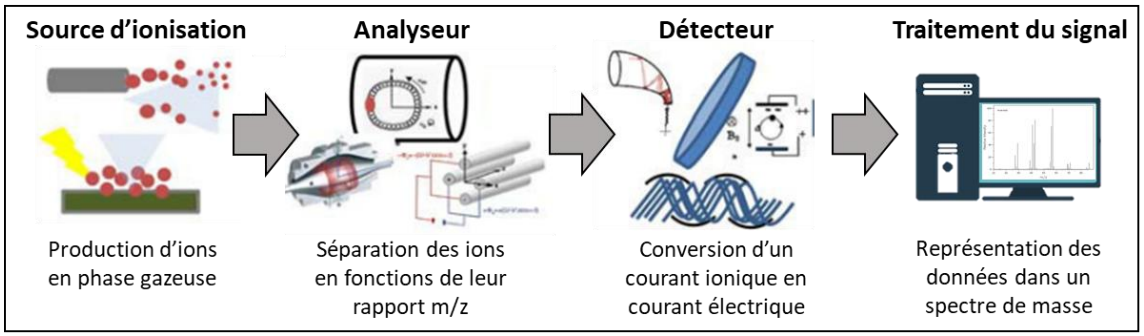


Figure 1.8 : Principe schématisé du fonctionnement d'un spectromètre de masse.

La spectrométrie de masse fait partie des techniques analytiques les plus utilisées pour mesurer le marquage dans des molécules organiques grâce à la différence de masse observée après l'incorporation d'un ou plusieurs isotopes. Le spectromètre de masse peut être couplé à une technique séparative telle que la chromatographie (gazeuse, liquide, ionique), l'électrophorèse capillaire ou la mobilité ionique qui permettent de séparer en amont les composés d'un mélange en fonction de leurs propriétés physico-chimiques et de leur affinité avec la colonne utilisée. Ce type de couplage combiné au développement de la spectrométrie de masse haute résolution (HRMS) permet de mesurer clairement l'ensemble du profil isotopique des molécules et de faciliter ainsi l'intégration des données. La haute sensibilité des spectromètres de masse actuels permet également de mesurer des niveaux faibles de marquage.

Pendant longtemps les analyses isotopiques par spectrométrie de masse ont été réalisées avec des méthodes ciblées (c'est-à-dire focalisées sur des composés identifiés dès le départ). Toutefois, des méthodes non ciblées de profilage isotopique, permettant d'élargir considérablement la dimension des réseaux métaboliques accessibles et d'identifier de nouveaux métabolites ou de nouvelles voies métaboliques, ont émergé récemment. La spectrométrie de masse peut également être utilisée en tandem (MS/MS ou MS^n), principalement dans une optique d'identification de molécules car elle permet une élucidation structurelle par fragmentation des composés, ou encore pour apporter une information positionnelle sur le marquage des composés.

Pour les raisons citées ci-dessus et notamment la meilleure couverture métabolique apportée par la spectrométrie de masse, le couplage chromatographie-spectrométrie de masse sera la technique analytique principalement utilisée au cours de ce travail de thèse.

3.5.2. Spécificité de la spectrométrie de masse

3.5.2.1. Principales techniques séparatives en couplage MS

Les techniques séparatives du type chromatographie sont les principales méthodes utilisées pour séparer les molécules dans des échantillons complexes. Il existe différentes méthodes de chromatographie comme la chromatographie liquide (LC), la chromatographie gazeuse (GC) ou la chromatographie ionique (IC). Ces méthodes reposent sur le même principe : l'échantillon est entrainé par une phase mobile à travers une phase stationnaire. Les métabolites contenus dans l'échantillon migrent à une vitesse dépendant de leurs propriétés physico-chimiques et en fonction de leur affinité avec les deux phases.

- ***Chromatographie gazeuse (GC)***

La GC permet de séparer les molécules sous forme gazeuse. Elle est couramment utilisée dans les analyses isotopiques mais n'est applicable qu'aux molécules volatiles ou volatilisables par dérivatisation. Le mélange à analyser est entrainé par un gaz au travers d'une phase stationnaire solide. La GC offre une haute capacité de séparation des pics et une forte sensibilité [Haggarty and Burgess, 2017] et est également plus reproductive et plus rapide que d'autre techniques chromatographiques. Cependant, contrairement à la LC, la GC nécessite généralement en amont, une dérivation chimique des espèces métaboliques non volatiles, telles que les sucres, les acides aminés ou des lipides présentant des fonctions polaires. L'aspect positif de cette dérivation est qu'elle offre un bruit de fond plus faible et augmente sensiblement le signal rapport/bruit. Mais cette étape reste contraignante car elle ajoute une étape à la préparation d'échantillons et peut induire des dégradations de composés. Au bilan, il est estimé que la GC permet l'analyse d'environ 20% des molécules connues.

- ***Chromatographie liquide (LC)***

La LC, comme la GC, est basée sur une différence d'affinité entre deux phases (mobile et stationnaire). Elle possède l'avantage de permettre la séparation simultanée de métabolites ayant des propriétés physico-chimiques différentes sans qu'il soit nécessaire de les modifier chimiquement. L'essor de la chromatographie liquide haute performance et ultra-haute performance (HPLC et UHPLC) permet également de travailler dans des conditions plus stringentes (pressions plus élevées et volumes plus faibles). Ces conditions permettent un gain de temps et de solvants et permettent de travailler avec des volumes d'échantillons plus faibles. Cela conduit à une meilleure sensibilité dûe à une diminution de la largeur des pics chromatographiques et une augmentation de leur intensité, ce qui améliore le rapport signal sur bruit. La variété des colonnes proposées en LC (HILIC, C18, …) augmente également la diversité des composés analysables ce qui est un critère non négligeable dans le cadre d'analyses non-ciblées [Harrieder et al, 2022 ; Pezzati et al, 2020]. On estime qu'environ 80% des composés connus peuvent être analysés en LC-MS. De plus, contrairement à la GC-MS, l'ionisation par électrospray utilisée en LC-MS est une technique d'ionisation douce qui réduit la fragmentation à la source.

- ***Chromatographie ionique (IC)***

La chromatographie ionique (IC) permet la séparation de composés chimiques en solution selon leur charge ionique. Cette technique est particulièrement bien adaptée à l'analyse des composés de forte polarité associés au métabolome central (acides organiques, sucres phosphorylés, nucléotides phosphorylés). La présence d'un suppresseur chimique et d'un suppresseur CO_2 permettent une très bonne sensibilité du signal. En effet l'étape de suppression diminue la conductivité de fond de l'éluant, minimise le bruit de fond, optimise le rapport signal sur bruit et augmente la sensibilité du système de mesure. Ceci présente l'avantage de permettre l'utilisation de plus faibles concentrations pour obtenir une sensibilité acceptable.

3.5.2.2. Source d'ionisation électrospray

Les ions, dits ions moléculaires, sont formés à partir de la molécule à analyser dans une source d'ionisation. La source la plus couramment utilisée en métabolomique pour ioniser la molécule est la source électrospray (ESI), qui est une technique d'ionisation dite « douce » car elle ne provoque pas de fragmentation significative des ions moléculaires formés. Elle consiste à appliquer une haute tension à l'échantillon liquide, le transformant ainsi en minuscules gouttelettes (spray) chargées selon le mode d'ionisation (positif ou négatif). Par action d'un gaz de nébulisation et un enchainement de processus de désolvatation, ces microgouttelettes sont transformées en gaz qui est injecté dans le détecteur de masse (Figure 1.9) [Fenn et al, 1989 ; Gowda and Djukovi, 2014]. Mais d'autres sources d'ionisation existent telles que l'ionisation électrique (EI) et l'ionisation chimique (CI), plus utilisées en présence d'un couplage chromatographie gazeuse [Gowda and Djukovic, 2014], ou encore l'ionisation chimique à pression atmosphérique (APCI) [Ayala-Cabrera et al, 2023] et l'ionisation par désorption laser (MALDI) [Enomoto et al, 2018]. L'ESI présente cependant quelques inconvénients, le plus notable étant les effets de suppression d'ions importants lors de l'analyse de mélanges moléculaires complexes [Gowda and Djukovic, 2014].

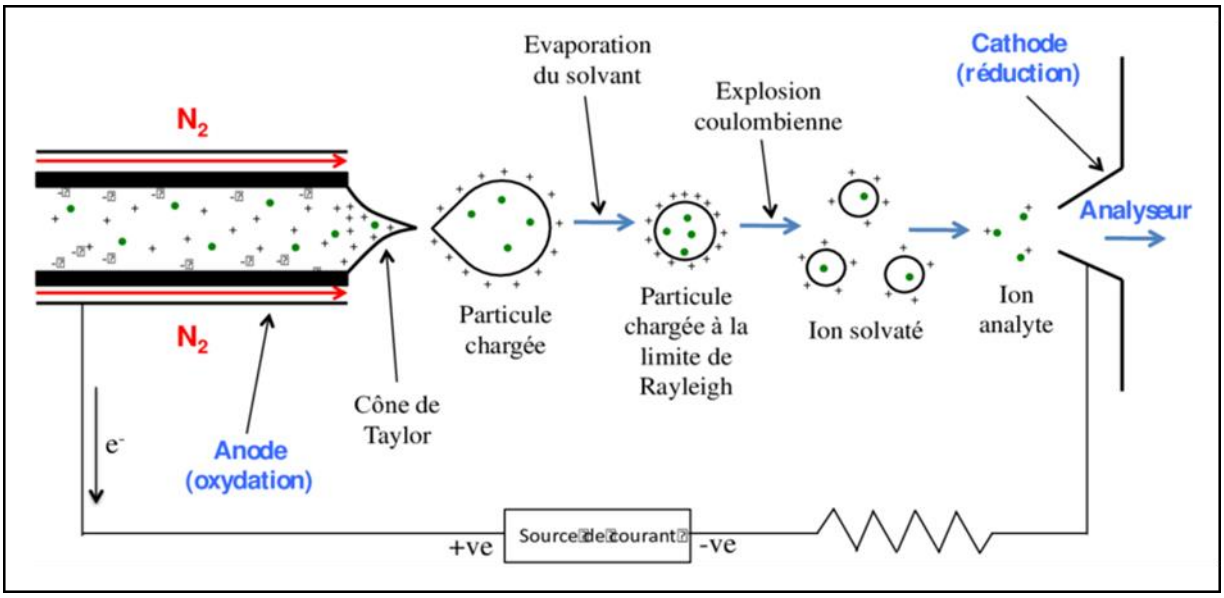


Figure 1.9 : Principe de la source électrospray.

- ***Analyse sans couplage***

Il faut noter qu'il est également possible d'utiliser la spéctrométrie de masse sans séparation préalable. L'injection directe (DIMS) est l'approche la plus simple et la plus rapide en spectrométrie de masse. L'analyse est effectuée sur des échantillons non modifiés en milieu ambiant. Parmi les sources permettant d'éffectuer de l'injection directe en MS, on trouve la LAESI (laser ablation electrospray ionization) et la DART (Direct analysis in real time mass spectrometry ionization). Cependant, la source la plus couramment utilisée est la source DESI [Takats et al, 2005] (Figure 1.10). Cette méthode d'ionisation utilise des microgouttelettes chargées pour extraire des analytes d'échantillons complexes sur une surface. L'injection directe a l'avantage de s'affranchir d'une étape de préparation d'échantillons et permet un criblage haut-débit. Elle est régulièrement utilisée en imagerie. Son utilisation est toutefois limitée en termes de quantification et d'identification. En effet, la mesure simultanée d'un grand nombre de composés dans une matrice complexe, entraîne des effets de matrice, comme la suppression d'ions.

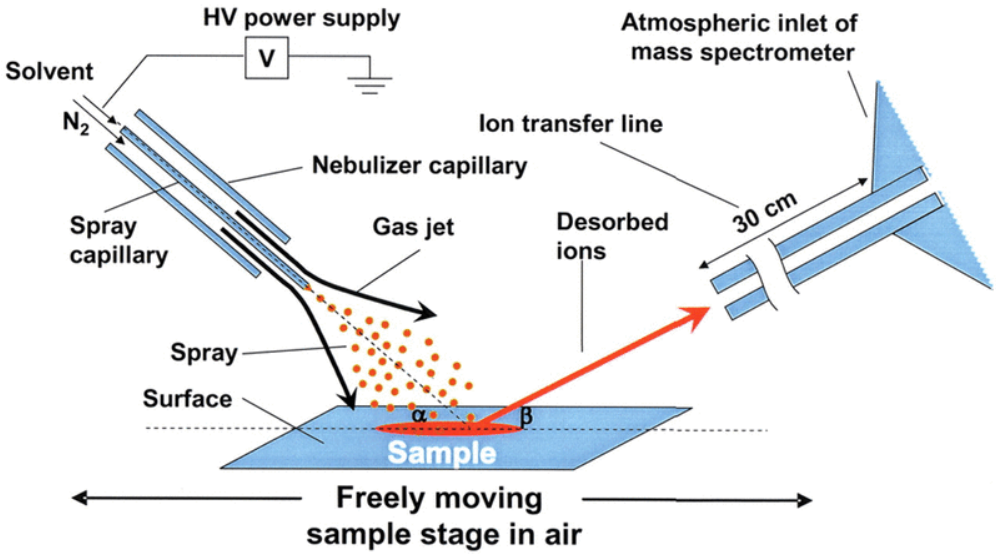


Figure 1.10 : Principe de fonctionnement de la source DESI.

3.5.2.3. Principaux analyseurs de masse pour études isotopiques

Les ions produits dans la source d'ionisation sont ensuite séparés au niveau de l'analyseur en fonction de leur rapport m/z. Les performances analytiques des analyseurs sont évaluées à partir de cinq paramètres principaux :

(i) La résolution (R) : la résolution d'un spectromètre de masse définit sa capacité à séparer des ions ayant des rapports m/z proches. Ce paramètre est l'un des plus importants pour les expériences de traçage isotopique.

$$R = \frac{m}{\Delta m}$$

Où m correspond au rapport m/z mesuré et Δm à la largeur à mi-hauteur de ce même pic.
On distingue les appareils à basse résolution (R < 1000) des appareils à haute résolution (R > 10 000), voire à très haute résolution (R > 100 000).

(ii) La précision correspond à la capacité de l'analyseur à mesurer avec justesse la masse de l'ion par rapport à sa masse théorique. Elle s'exprime sous la forme d'une erreur entre la masse mesurée et la masse théorique. Elle est exprimée en ppm (partie par million)

$$\Delta ppm = \frac{(m/z\ \exp - m/z\ th)}{m/z\ th} \times 10^6$$

La précision est un élément déterminant pour l'extraction et l'identification des signaux d'intérêt.

(iii) La gamme dynamique correspond à l'étendue des concentrations détectables par l'instrument.

(iv) La gamme m/z qui détermine les limites des rapports m/z détectables par l'instrument.

(v) La vitesse d'acquisition qui correspond au temps nécessaire à l'enregistrement d'un spectre de masse. Elle s'exprime sous la forme d'une fréquence (nombre de spectres par seconde ou Hertz) ou sous la forme d'un temps d'acquisition en millisecondes.

Pour l'étude du métabolome, une haute résolution et une haute sensibilité sont souhaitées. Cependant il est difficile de pouvoir combiner les deux car une sensibilité élevée induit généralement une baisse de résolution et inversement. On distingue plusieurs types d'analyseurs de masse, les analyseurs uniques ou hybrides. Parmi la première catégorie, on trouve des analyseurs à basse résolution (BR) et haute résolution (HR). Les performances des principaux analyseurs de masse sont présentées dans le tableau 1.1 [Kaklamanos et al, 2020]. Les différents types d'analyseurs sont illustrés en annexe (Annexe 1).

- ***Analyseurs basse résolution***

Les analyseurs basse résolution (BR) sont des quadripôles (Q) ou des pièges à ions linéaires ou non (ion trap).

L'analyseur quadripôlaire est le spectromètre de masse le plus simple et le plus abordable. Il est constitué de quatre électrodes parallèles. Des potentiels de courant alternatif à radiofréquence variable sont appliqués à chaque paire d'électrodes (positive et négative), créant un champ quadripolaire. Au sein de ce champ, seuls les ions présentant une trajectoire stable sont conduits jusqu'à l'extrémité de l'analyseur, les autres étant arrêtés lorsque leur trajectoire instable les conduit à entrer en collision avec les électrodes. Les limites des analyseurs de masse quadripolaires sont leur faible résolution et la mauvaise précision de la masse.

Les pièges à ions linéaires sont des analyseurs constitués de quatre électrodes à section hyperbolique. Lors de leur introduction dans l'analyseur, les ions sont décélérés par collision avec un gaz inerte présent à faible pression. Ils sont piégés dans un mouvement d'oscillation entre les électrodes puis éjectés du piège vers le détecteur par ordre croissant de m/z. Globalement, la résolution des pièges à ions est légèrement supérieure à celle des quadripôles. Cependant, ils sont de moins en moins utilisés comme seul analyseur en raison du développement plus rapide de la technologie du triple quadripôle et de la haute résolution. Les ions trap seront préférentiellement utilisés en combinaison avec d'autres analyseurs.

- ***Analyseurs haute résolution***

Les analyseurs haute résolution (HR) sont les analyseur à temps de vol (TOF), à Résonance Cyclonique Ionique (FT-ICR Fourrier Transform – Ion Cyclotron Resonance) et les spectromètres de masse de technologie Orbitrap.

L'analyseur de masse TOF est basée sur le fait que les ions ayant la même énergie mais des masses différentes se déplacent à des vitesses différentes. Dans ce type d'analyseur, les ions sont accélérés par un champ électrostatique pour obtenir la même énergie cinétique et se déplacent ensuite jusqu'au détecteur avec une vitesse inversement proportionnelle au carré de leur rapport m/z. La mesure du temps de vol de chaque ion permet de déterminer leur rapport m/z. Les TOF sont les moins résolutifs mais ont des temps d'acquisition relativement courts.

À l'opposé, les FT-ICR permettent d'atteindre la plus haute résolution et la meilleure précision de mesure mais nécessitent plus de temps pour acquérir un spectre. Les ions sont placés dans un champ magnétique très puissant, qui les fait se déplacer sur des trajectoires circulaires. Ils sont piégés par le champ magnétique et une excitation à large bande de fréquence est ensuite effectuée. Le mouvement des ions induit un courant qui est enregistré, mesuré et transformé par la méthode de Fourier pour produire des spectres de masse.

Les analyseurs de type Orbitrap quant à eux présentent des performances intermédiaires. L'orbitrap est essentiellement un piège à ions électrostatique : les ions sont piégés, puis ils se déplacent de façon circulaire avant d'être éjectés en fonction de leur valeur m/z. Les Orbitrap sont de plus développés et leur performance évolue rapidement. Par ce fait, ils sont de plus en plus utilisés pour les analyses métabolomiques.

Les spectromètres de masse hybrides combinent plusieurs analyseurs, généralement un analyseur BR (ex : quadripôle ou piège à ions linéaire) couplé à un analyseur HR tels que le LTQ-Orbitrap et le QTOF. Associer plusieurs analyseurs permet d'améliorer leurs performances. Ces spectromètres hybrides peuvent être utilisés pour du profilage ou en tandem pour la sélection et la fragmentation d'un nombre réduit d'ions. La spectrométrie de masse en tandem (MS/MS ou MS^n) consiste à sélectionner un ion par un premier analyseur, à le fragmenter, puis à effectuer une deuxième sélection sur les fragments ainsi générés. L'obtention d'ions de générations supérieures est possible par simple renouvellement du processus (sélection d'un ion produit, fragmentation, sélection d'un ion produit de 2^e génération, fragmentation, etc...). Cette séquence est appelée MS^n, n étant le nombre de générations d'ions successives. On trouve également des triples quadripôles (QQQ) et des quadripôles couplés à un piège à ion (QTrap) qui sont souvent associés à des études ciblées car plus sélectifs.

Analyseur	Résolution	Précision en masse (ppm)	Gamme dynamique	Gamme de m/z	Temps d'acquisition (Hz)
Quadripôle	< 3000	> 100	Très bonne	> 3000	0.5-4
Ion trap	< 5000	< 30	Faible	> 20000	1-10
TOF	5000-40000	4-100	Bonne	> 20000	> 2000
Orbitrap	> 140000	< 1-2	Medium	> 6000	> 12
FT-ICR	$2.5.10^6$	< 0.6	Medium	> 30000	1

Tableau 1.1 : Performance des principaux analyseurs de masse utilisés en métabolomiques. Données constructeurs issues de Kaklamanos et al, 2020.

- ***Spectromètre de masse à rapport isotopique***

Le spectromètre de masse à rapport isotopique (IRMS) est un appareil spécifique qui permet l'étude d'échantillons en fonction de leurs rapports isotopiques (mesure l'abondance relative des différents isotopes d'un même élément chimique). Il peut également être couplée à des techniques chromatographiques (LC/GC). Les molécules sont ensuite oxydées par une étape de combustion en sortie de colonne pour les convertir en gaz simple (ex : CO_2, H_2). Ce gaz est ensuite analysé par le spectromètre de masse pour déterminer l'abondance des isotopes. Dans certaines études sur le traçage isotopique stable du métabolisme, l'IRMS est utilisé pour sa capacité à détecter et quantifier des niveaux d'enrichissements isotopiques très faibles (inférieurs à 5%) [Fan et al, 2012].

3.5.2.1. Imagerie par spectrométrie de masse

Le principe de l'imagerie par spectrométrie de masse est de réaliser une cartographie de la surface d'une coupe de tissu, en enregistrant à chaque point un spectre de masse. Elle est basée sur l'utilisation d'une source d'ionisation principalement de type MALDI (désorption laser) ou DESI (désorption par électrospray) et d'un analyseur de masse à haute résolution. L'imagerie par spectrométrie de masse est une méthode destructive pour laquelle les approches de quantification sont compliquées dû à l'hétérogénéité des tissus. Récemment une nouvelle approche, l'imagerie NanoSIMS (Secondary Ion Mass Spectrometry) a été développée pour permettre de visualiser les composés dans les cellules et tissus de manière quantitative, avec une résolution spatiale élevée et une grande sensibilité. Dans la plupart des analyses NanoSIMS, le marquage par isotopes stables est utilisé pour permettre la détection de composés d'intérêts ou révéler l'activité métabolique dans des échantillons biologiques, en identifiant l'endroit où l'enrichissement isotopique est localisé. En règle générale, cela se fait en mesurant la valeur du

rapport $^{13}C/^{12}C$ (ou $^{14}N/^{15}N$) et en la comparant à un échantillon standard (Figure 1.11). Cette nouvelle technologie a été utilisée pour tracer les composés volatiles dérivés du métabolisme intracellulaire et révéler leur relation avec les voies métaboliques du cancer [Lee et al, 2018].

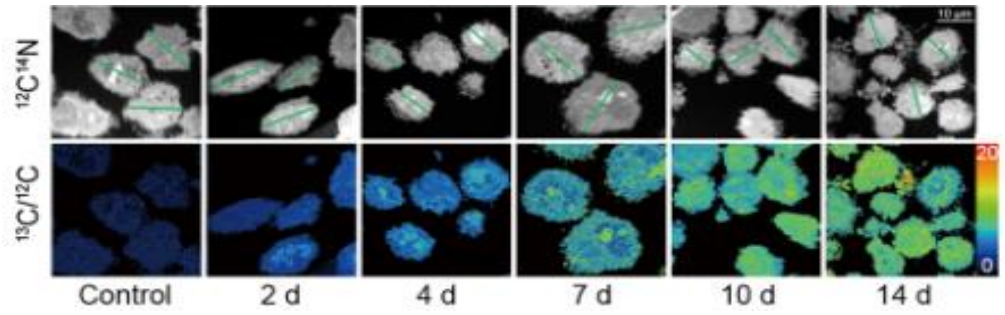


Figure 1.11 : Imagerie cellulaire par NanoSIMS pour mesurer les flux de carbone. Les formes celulaires ($^{12}C^{14}N$) et les rapports $^{13}C/^{12}C$ correspondants après marquage isotopique sont représentés. L'intensité des pixels est mesurée pour calculer le rapport moyen $^{13}C/^{12}C$ de chaque cellule. Illustration tirée de Lee et al, 2018.

3.5.2.2. Mesure de profils isotopiques par spectrométrie de masse

Les isotopologues possèdent la même formule chimique et la même structure et se comportent donc de manière identique lors de la séparation chromatographique, ce qui résulte sur une coélution des isotopologues à un même temps de rétention. Cependant lors de l'incorporation d'un (ou plusieurs) isotope(s) du carbone dans un métabolite, la masse du métabolite concerné va varier en fonction du nombre d'atomes de ^{13}C incorporés. C'est ce décalage de masse qui est observé lors d'une analyse en spectrométrie de masse. En effet la masse d'un atome de ^{12}C étant de 12 u et celle du ^{13}C de 13.00335 u, les pics de masse du composé d'intérêt seront séparés d'une valeur de 1.00335 m/z correspondant à la variation de masse entre la molécule non marquée et la molécule ayant incorporé un atome de ^{13}C. Le spectre de masse du Fumarate (Figure 1.12) contenant quatre carbones pourra présenter jusqu'à 5 pics de masse en fonction du nombre d'incorporation isotopique.

En métabolomique, les données extraites lors d'une analyse de spectrométrie de masse couplée à la chromatographie sont des données tridimensionnelles (communément appelées features), qui sont caractérisées par leur temps de rétention (RT), leur ratio masse sur charge m/z, et leur intensité (abondance). Dans les études isotopiques, les abondances de chaque isotopologues peuvent être exprimées en valeur relative selon la formule suivante :

$$Mk = \frac{Area\ Mk}{\sum_{i=0}^{n} Mi}$$

n étant le nombre total d'isotopologues de la molécule et $0 \leq Mk \leq 1$

La distribution des isotopologues est caractérisée par le pourcentage de chaque isotopologue, et peut s'exprimer également en fraction. On appelle Distribution des Isotopologues du Carbone (CID) le vecteur représentant la distribution relative de tous les isotopologues du carbone existant pour un composé donné. Le CID est donné par le massif isotopique de l'ion moléculaire analysé après correction de l'abondance naturelle des isotopes des autres éléments chimiques présents dans la molécule (voir ci-dessous).

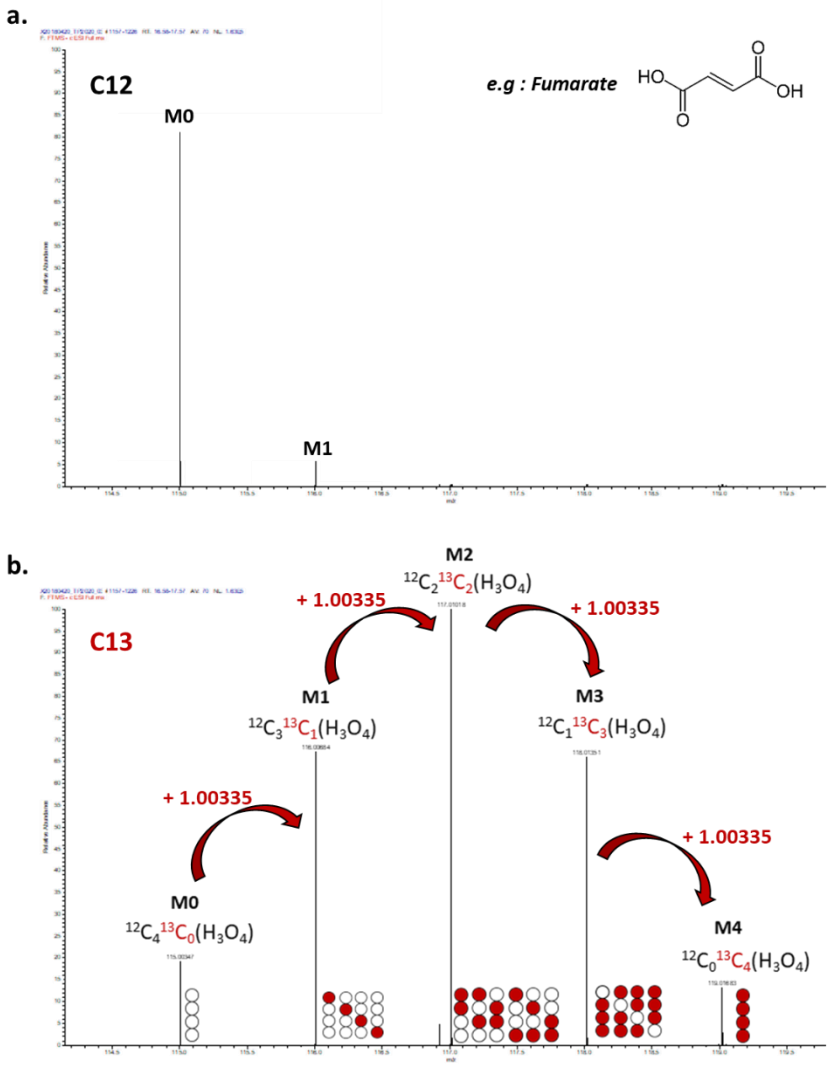


Figure 1.12 :Spectres de masse du fumarate a. Spectre de masse du fumarate non marqué. B. Massif isotopique des isotopologues du fumarate suivant la loi binomiale du Triangle de Pascal mesuré par spectrométrie de masse.

Dans les expériences de marquage, les isotopologues des métabolites sont quantifiés à partir de leurs massifs isotopiques dans les spectres MS. Ces massifs contiennent les informations relatives aux isotopes suivis (ex : ^{13}C) mais également des informations sur les isotopes de tous les éléments chimiques présents dans les molécules, tels que le carbone, l'hydrogène, l'azote, l'oxygène et le soufre [Midani et al, 2017]. La présence de ces isotopes

peut impacter significativement les spectres de masse des métabolites, en fonction de leur abondance naturelle. Les spectromètres de masse à haute résolution permettent de séparer directement certaines formes isotopiques suivant leur résolution de travail. Par exemple, le signal provenant d'une forme marquée ^{13}C peut être séparée de celui provenant d'une espèce contenant un atome de ^{15}N (Figure 1.13). Pour extraire et exploiter correctement les données de marquage, il est donc nécessaire de corriger de l'abondance naturelle tous ces autres isotopes. Plusieurs outils ont été développés pour corriger cela tel que le logiciel IsoCor développé par Millard et al, qui prend en compte la résolution de l'appareil de mesure [Millard et al, 2012].

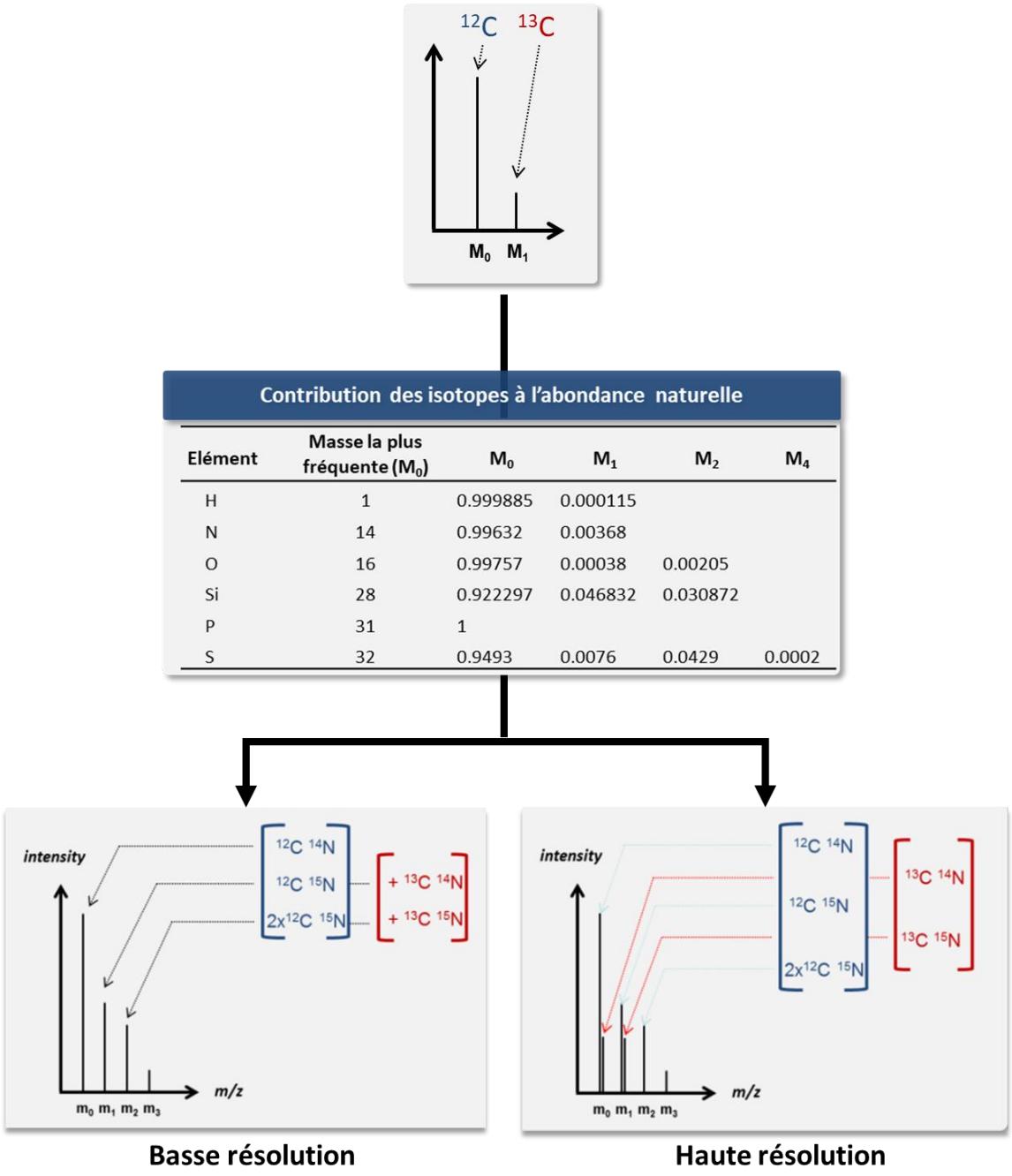


Figure 1.13 : Impact de la contribution des isotopes présents à l'abondance naturelle sur les spectres de masse haute et basse résolution.

Enfin l'enrichissement ^{13}C représente le pourcentage global de carbone 13 de la molécule et est donné par la formule suivante :

$$\%13C = 100 * \frac{0 * Aire\ M0 + 1 * Aire\ M1 + \cdots + n * Aire\ Mn}{n}$$

n étant le nombre d'isotopologues de la molécule.

Dans la suite de cette introduction, nous allons voir comment les outils isotopiques sont exploités dans le cadre d'études du métabolisme par LC/HRMS. Comme énoncé précédemment, on distingue deux grands axes d'utilisation des isotopes pour l'étude du métabolisme : (i) la métabolomique assistée par les isotopes et (ii) les études de traçage isotopique. Le travail de thèse se concentrant particulièrement sur ce deuxième axe, il sera détaillé plus longuement dans la suite de cette introduction.

4. Métabolomique assistée par les isotopes stables

Malgré le développement d'outils analytiques puissants, l'analyse du métabolome est rendue complexe par différents facteurs liés à sa nature (section 3.5). Dans ce contexte, les isotopes sont utilisés de manière croissante comme appui à la métabolomique. Il s'agit d'utiliser des standards (simples ou complexes) isotopiquement marqués pour répondre à différents besoins analytiques, incluant des aspects de qualification des instruments et de validation de méthodes (qualité des données), la normalisation ou la standardisation des analyses (gestion des analyses), s'assurer de la qualité des données analytiques, l'aide à l'annotation du métabolome ou encore pour le développement des approches de métabolomique quantitative (appui à l'interprétation des données).

4.1. Les standards isotopiques pour la métabolomique

4.1.1. Composés purs

Il s'agit de composés disponibles soit commercialement soit obtenus par synthèse spécifique (chimique ou enzymatique), qui peuvent être utilisés comme standards d'identification ou de quantification. Suivant le besoin, le marquage peut être spécifique d'une ou plusieurs positions bien précises au sein de la molécule, ou alors uniforme sur l'ensemble des positions concernées par l'élément chimique considéré. A noter que ce sont principalement des molécules deutérées ou marquées au ^{13}C, et plus récemment au ^{15}N, qui sont principalement

utilisées dans ce cadre. Toutefois, si le marché s'étend depuis quelques années en raison de la croissance des besoins, la disponibilité ou le coût élévé de ces molécules reste très limitant, en particulier pour l'étude globale du métabolome.

De plus, un des problèmes inhérents aux standards marqués aux isotopes stables est qu'aucun isotope n'est pur à 100 % ; il y aura toujours des traces d'autres isotopes. La plupart des isotopes stables enrichis disponibles dans le commerce sont purs à 95-99 % et, bien que leur contenu soit certifié, l'incertitude concernant la contribution des isotopes mineurs peut être importante. C'est un problème pour les expériences isotopiques quantitatives, car on ne peut alors pas supposer qu'un seul isotope (pur) est ajouté au système biologique.

4.1.2. Standards complexes

Les autres standards marqués isotopiquement sont des mélanges de molécules marquées. Ces standards peuvent être soit des mélanges de composés purs, soit issus du marquage chimique d'échantillons biologiques, ou encore être produits biologiquement.

- ***IROA-IS***

Une stratégie de marquage spécifique des standards internes est appliquée dans le cadre de la méthode IROA (« Isotopic Ratio Outlier Analysis »). La particularité de cette méthode est qu'elle s'appuie sur l'utilisation de standards internes possèdant un marquage spécifique conçu à partir de matériels marqués à 95% et 5% en ^{13}C pour améliorer l'annotation du métabolome ainsi que la reproductibilité et la précision de sa quantification. Ces pourcentages d'enrichissement créent une signature isotopique spécifique qui est facilement identifiable par le logiciel ClusterFinder, qui est un logiciel d'analyse de données métabolomiques pour la métabolomique ciblée et non ciblée. Les standards IROA (IROA-IS) sont disponibles commercialement (IROA technologies). Ils peuvent être ajoutés à n'importe quel type d'échantillon (cellules, tissus, biofluides) afin d'identifier et quantifier les métabolites qu'ils contiennent.

- ***Standards produits biologiquement***

Une stratégie alternative consiste à produire biologiquement des standards marqués isotopiquement. Cette méthode est adaptée dans le cas où il n'est pas possible d'avoir accès aux standards spécifiques pour chaque composé. La disponibilité des métabolites marqués aux

isotopes stables étant limitée et leur coût élevé, des techniques de marquage *in vivo* ont été mise en place.

Standard IDMS (Isotopomer Dilution Mass Spectrometry)

La première approche a été décrite par Mashego et al. pour des analyses quantitatives du métabolome par spectrométrie de masse à dilution isotopique (IDMS) [Mashego et al, 2004 ; Wu et al, 2005]. Dans cette approche, des échantillons biologiques (cellules, bactéries) sont cultivés sur des milieux contenant des substrats uniformément marqués, par exemple au ^{13}C. Ces substrats sont ensuite métabolisés par l'organisme, les cellules contenant alors un mélange de métabolites uniformément marqués. L'intérêt de ce standard est la couverture métabolique qu'il propose car il contient l'ensemble des métabolites présents chez un organisme dans une condition donnée. Son utilisation implique certaines contraintes car il doit posséder une composition métabolique proche de l'échantillon analysé. Pour ce faire, il doit idéalement être produit chez le même organisme, dans les mêmes conditions de culture et en parallèle de la culture de l'organisme d'intérêt. Ces conditions peuvent être difficilement respectées car la production du standard IDMS est un protocole expérimental lourd [Wu et al, 2005]. Par exemple, pour analyser le métabolisme de cellules mammifères, le standard IDMS sera préférentiellement produit à partir de levure.

Échantillon standard « Triangle de Pascal »

Une autre stratégie consiste en la production biologique d'un échantillon de référence marqué contenant toutes les espèces isotopiques de chaque métabolite dans une quantité prévisible et contrôlable. Cette stratégie est valable pour différents isotopes et différents organismes (ex : *E. coli ou P. augusta*). L'organisme en question est cultivé avec un substrat marqué, au ^{13}C par exemple, comme seule source de carbone. La composition isotopique de ce substrat est conçue pour obtenir des échantillons dans lesquels la distribution des isotopologues du carbone de tous les métabolites devrait donner les coefficients binomiaux du triangle de Pascal. Cet échantillon, appelé « Triangle de Pascal », présente l'intégralité du massif isotopique « carboné » d'un métabolite. Le principe et la méthodologie de production de cet échantillon a été décrit par Millard et al, 2014.

4.2.Stratégies d'utilisation des standards isotopiques

4.2.1. Evaluation de la qualité des données

- ***Qualification d'instruments et qualité des données***

L'utilisation des standards marqués est une approche importante pour qualifier des instruments analytiques et donner une indication de la précision des résultats. Les critères communément utilisés pour valider les données analytiques consistent à évaluer la limite de détection de la méthode analytique, la précision et la justesse des mesures, la gamme dynamique et la gamme de linéarité. La spécificité liée aux approches par marquage isotopique est que l'ensemble du massif isotopique est détecté, ce qui augmente le risque de contamination ou de chevauchement des signaux de masse (par exemple si un isotopologue marqué d'un composé est un isobare du M0 d'un autre composé). Les approches isotopiques nécessitent donc la validation des mesures sur l'ensemble des isotopologues des molécules d'intérêts.

Dans cet optique, l'échantillon du Triangle de Pascal (PT) est un atout précieux. Cet échantillon contient tous les isotopologues d'un métabolite en quantités connues (et prévisibles) et suffisantes. Son utilisation au cours d'une expérience métabolomique permet d'évaluer des protocoles expérimentaux complets (de l'échantillonnage au traitement des données), d'identifier l'origine de contaminations ou encore d'évaluer différentes plateformes analytiques [Millard et al, 2014 ; Heuillet et al, 2018]. Pour ce faire, l'échantillon PT est analysé et les profils isotopiques de ses métabolites sont extraits. Par comparaison du profil mesuré (et corrigé) avec le profil isotopique théorique du Triangle de Pascal (Figure 1.14A), il est possible d'évaluer et de valider les performances de la méthode analytique. La validation des mesures isotopiques s'appuie sur deux types de critères de validation : des critères communs à la validation classique et d'autres spécifiques à l'analyse des isotopologues (Figure 1.14B) [Heuillet et al, 2018]. Une méthodologie exploitant ces différentes mesures de qualité a été décrite par Heuillet et al. pour qualifier des techniques instrumentales [Heuillet et al, 2018].

L'utilisation de cet échantillon a été étendu aux approches non-ciblées pour optimiser les paramètres de logiciels d'extraction automatique et valider les données isotopiques générées par ces approches. Cette application est illustrée dans le Chapitre 2 de ce manuscrit de thèse [Butin et al, 2022].

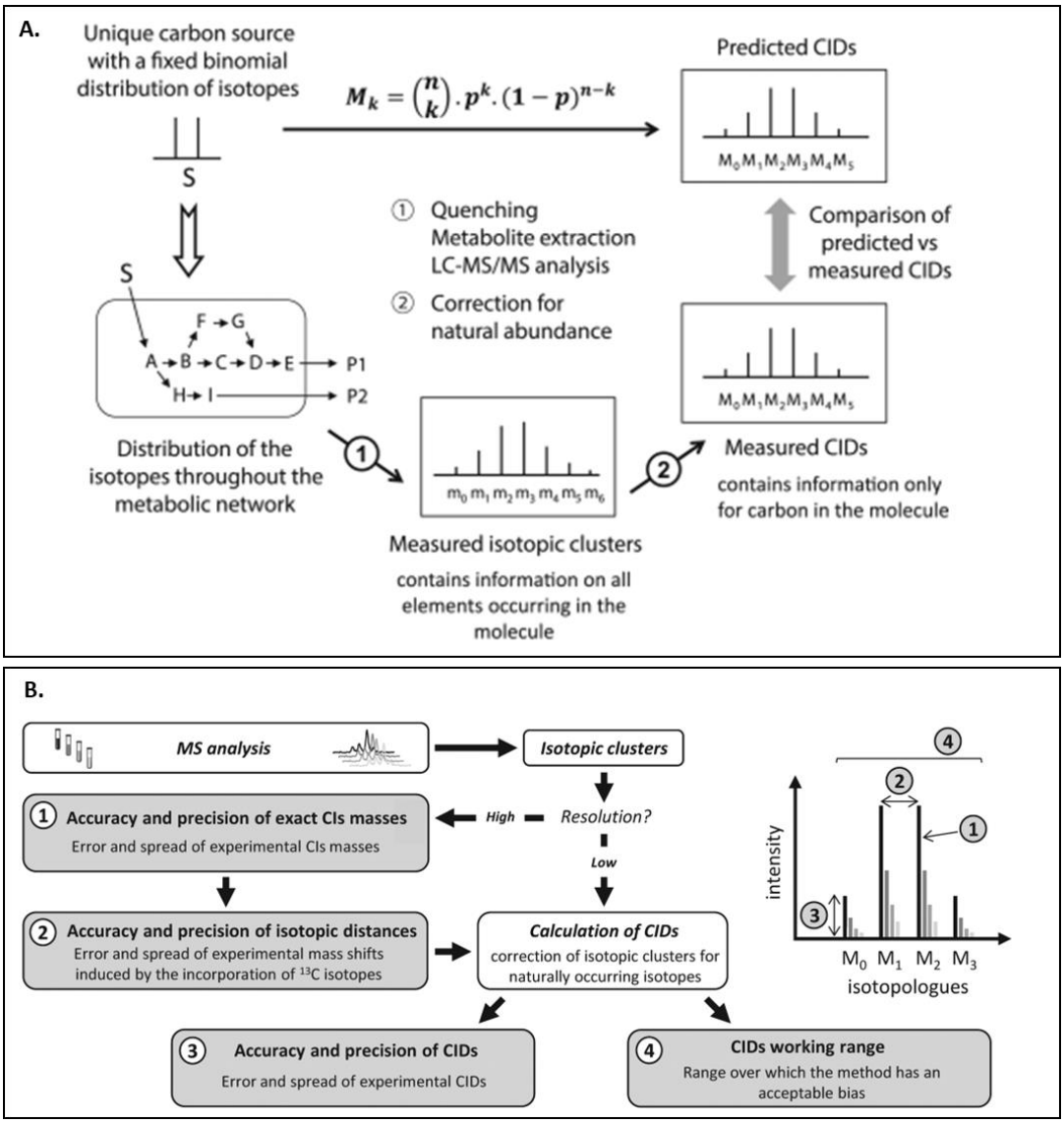


Figure 1.14 : Stratégies d'utilisation de l'échantillon Triangle de Pascal pour l'évaluation des données isotopiques. A. Principe global d'évaluation des données par comparaison des profils isotopiques mesurés (et corrigés) et des profils isotopiques prédits suivant la loi binomiale du triangle de pascal. L'abondance relative M_k de l'isotopologue ayant incorporé k atomes de ^{13}C est donnée par l'équation ci-dessus. n est le nombre total d'atomes de carbone et p l'abondance des isotopes ^{13}C. $\binom{n}{k}$ sont les coefficients binomiaux et le terme "p^k-(1 - p)$^{n-k}$" est la proportion de chaque forme isotopique qui compose l'isotopologue. B. Protocole et critères d'évaluations proposés par Heuillet et al, 2018 pour évaluer les méthodes d'analyses isotopiques par MS.

- ***Normalisation des données***

Les standards isotopiques peuvent également être utilisés pour normaliser les données. Dans le cadre d'analyses LC/HRMS ciblées ou non-ciblées, les intensités et les temps de rétention chromatographiques peuvent subir une dérive temporelle. Idéalement, les données doivent être acquises dans les mêmes conditions et le même jour afin de minimiser ces variations. Malgré tout, il peut être difficile de comparer les mesures entre plusieurs études ou entre différents instruments. L'incorporation en amont de l'analyse d'un standard marqué dans les échantillons permet de suivre cette dérive analytique. Lorsque la variation d'intensité

observée est trop importante, les données sur les métabolites peuvent être normalisées sur la base de cet étalon interne. Cette stratégie permet de normaliser les concentrations de métabolites entre un même ensemble d'échantillons ou entre différents batchs analytiques. Elle est notamment utilisée lors d'analyses de cohortes biologiques, qui se répartissent sur des temps d'étude relativement longs et qui nécessitent d'aligner les différents batchs analytiques pour exploiter les données sur l'ensemble de l'étude [Bongaerts et al, 2021 ; Weindl et al, 2015].

4.2.2. Support pour l'identification de métabolites

- ***Identification de la formule moléculaire***

Les standards marqués isotopiquement peuvent également servir de support pour l'identification et l'annotation de métabolites [Dunn et al ; 2013 ; Wishart; 2011]. L'annotation en spectrométrie de masse commence généralement par la prédiction des formules moléculaires et la comparaison des masses mesurées avec des bases de données de composés telles que PubChem [Kim et al, 2023] ou ChEBI [Hastings et al, 2016], ou de bases de données de voies métaboliques telles que MetaCyc [Karp et al, 2002] ou KEGG [Kanehisa et al, 2000]. Cependant la connaissance de la masse d'un métabolite ne suffit pas pour déterminer sa composition élémentaire.

L'addition de standards uniformément marqués présente plusieurs avantages pour déterminer la composition élémentaire des métabolites concernés. Ils permettent par exemple de déterminer le nombre d'atomes de carbone dans le métabolite. Le spectre du standard marqué présente un pic de masse, l'isotopologue Mn, décalé de la masse du M0 d'une valeur de n*1.00335 Da (Figure 1.15).

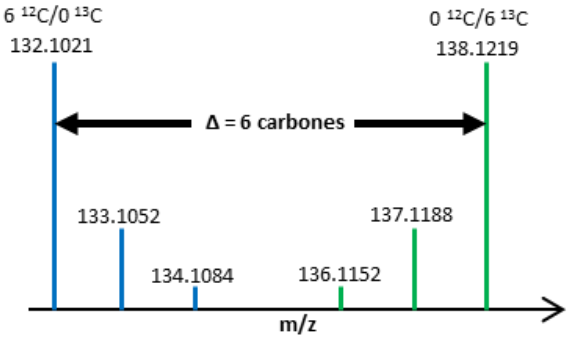


Figure 1.15 : Principe de détermination du nombre d'atomes de carbone dans une molécule

Une autre stratégie consiste à mixer le standard IROA avec l'échantillon et à déterminer le nombre de carbone de la molécule en se basant sur deux paramètres : la distance entre les deux pics de base C12 et C13 et la hauteur relative du M+1 et M-1 respectivement. La stratégie

de marquage caractéristique du standard IROA conduit à des modèles isotopiques spécifiques qui permettent de différencier les signaux biologiques des artefacts et d'obtenir le nombre exact de carbones, ce qui réduit considérablement les formules moléculaires possibles [Clendinen et al, 2015].

- ***Identification structurale***

En plus de la détermination de la formule moléculaire, le marquage isotopique peut aider à l'identification structurale des métabolites [Hegeman et al, 2007, Neumann et al, 2014]. Dans de très nombreux cas, la formule élémentaire n'est pas suffisante pour identifier avec précision les métabolites. Des stratégies de marquage alternatives peuvent être utilisées comme l'utilisation de marquage partiel pour faciliter l'identification des métabolites sur la base des connaissances préalables des voies biochimiques [Creek et al, 2012]. D'autre part, l'identification structurale en spectrométrie de masse étant généralement réalisée via des approches MS^2 ou MS^n, le marquage peut aider à l'interprétation des spectres de fragmentation résultants. Neumann et al, proposent une approche combinant le marquage et la spectrométrie de masse en tandem pour déterminer de manière automatique la composition de fragments d'ions via le logiciel FragExtract (Figure 1.16) [Neumann et al, 2014 ; Bueschl et al, 2017]. Cette approche repose sur l'utilisation d'un mélange d'échantillons non marqués (P) et uniformément marqués au ^{13}C (P'). La différence de masse entre les ions fragments résultant du composé natif et du composé marqué permet d'attribuer le nombre d'atomes de carbone à chaque signal et de générer des formules moléculaires putatives, aidant ainsi à déterminer la composition élémentaire des précurseurs.

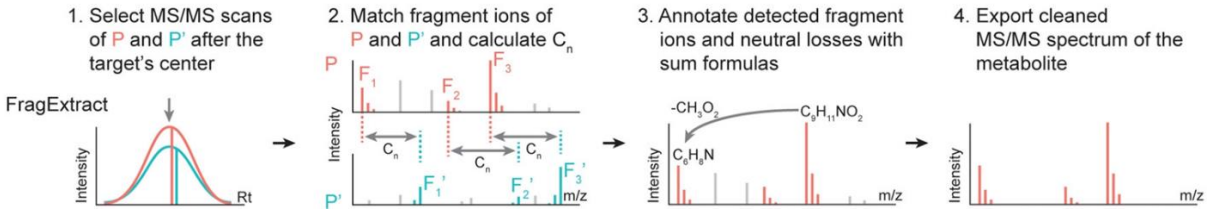


Figure 1.16 : Principe du logiciel FragExtract. Illustration tirée de Bueschl et al ; 2017

4.2.3. Méthodes de quantification

La quantification vise à mesurer les quantités ou les concentrations absolues des métabolites dans les échantillons analysés. Elle peut être réalisée par des méthodes de quantification utilisées classiquement en biochimie analytique (standards externes ou internes), mais dans le cas de la métabolomique qui vise à mesurer plusieurs centaines de composés en une seule analyse, des stratégies plus spécifiques ont été développées.

Un standard externe n'est pas ajouté à l'échantillon mais il est analysé indépendamment avec la même méthode. Il s'agit soit de la molécule d'intérêt, soit d'un composé très proche. Une courbe d'étalonnage permet de mesurer le coefficient de réponse du système analytique pour le composé d'intérêt pour l'appliquer aux échantillons. Cette approche a l'inconvénient de ne pas tenir compte des rendements d'extractions des composés et des effets de matrice particulièrement importants en LC-MS. Pour tenir compte de ces différents effets, une quantité connue de standard interne peut être ajoutée à l'échantillon. Il s'agit soit d'une molécule ayant un comportement proche à l'extraction et à l'analyse, soit de la même molécule marquée isotopiquement ce qui est généralement la meilleure option puisqu'on considère que la présence de l'isotope n'a pas d'effet significatif sur les propriétés physico-chimiques de la molécule. Les deux formes, marquées et non marquées, se comportent de la même manière tout au long du protocole analytique et possèdent le même facteur réponse. La distinction entre les deux se fait sur le spectre de masse, puisque les deux formes n'ont pas la même masse moléculaire. La quantité ajoutée du standard interne étant connue, une simple règle de trois permet d'avoir la quantité du composé dans l'échantillon. Ce principe de quantification par dilution isotopique est assez ancien et très fiable. Sa limite principale pour la métabolomique est qu'il faut disposer idéalement d'une forme marquée de chacun des composés d'intérêt, c'est-à-dire jusqu'à plusieurs centaines, ce qui n'est pas envisageable par addition individuelle de chaque standard.

Quantification par Isotopomer Dilution Mass Spectrometry (IDMS)

Il s'agit d'une approche développée par l'équipe de Joseph Heijnen (Delft University, Pays-Bas) au milieu des années 2000. Elle repose sur le principe de dilution isotopique, généralisé à l'ensemble d'un métabolome. Comme cela a été vu précédemment (section 4.1.2), le standard IDMS repose sur la production biologique d'un échantillon dont tous les métabolites sont marqués isotopiquement. Cela signifie que pour chacun des métabolites d'intérêt contenu dans un échantillon non marqué, le standard IDMS contient son équivalent marqué. En ajoutant le standard IDMS à l'échantillon, on obtient ainsi un standard interne pour chaque métabolite

présent dans l'échantillon. Cela permet a minima une quantification relative lorsque l'on compare plusieurs échantillons entre eux, à condition d'ajouter la même quantité de standard IDMS à chacun (Figure 1.17). Si la quantité du métabolite marqué dans le standard IDMS est connue, on peut alors appliquer une règle de trois pour calculer la quantité du composé dans l'échantillon. Une façon de quantifier le composé marqué dans le standard IDMS est d'ajouter une quantité connue du composé non marqué (« contre-quantification »).

L'utilisation du standard IDMS permet également de compenser les variations dues à la préparation de l'échantillon et aux effets de matrice. Les phases d'extraction, d'évaporation et de lyophilisation nécessaires à la préparation de l'échantillon avant l'analyse en MS peuvent favoriser la dégradation ou la conversion de certains métabolites. L'ajout du standard interne IDMS au moment de l'extraction assure que ces variations soient subies de la même manière pour les composés marqués et non marqués.

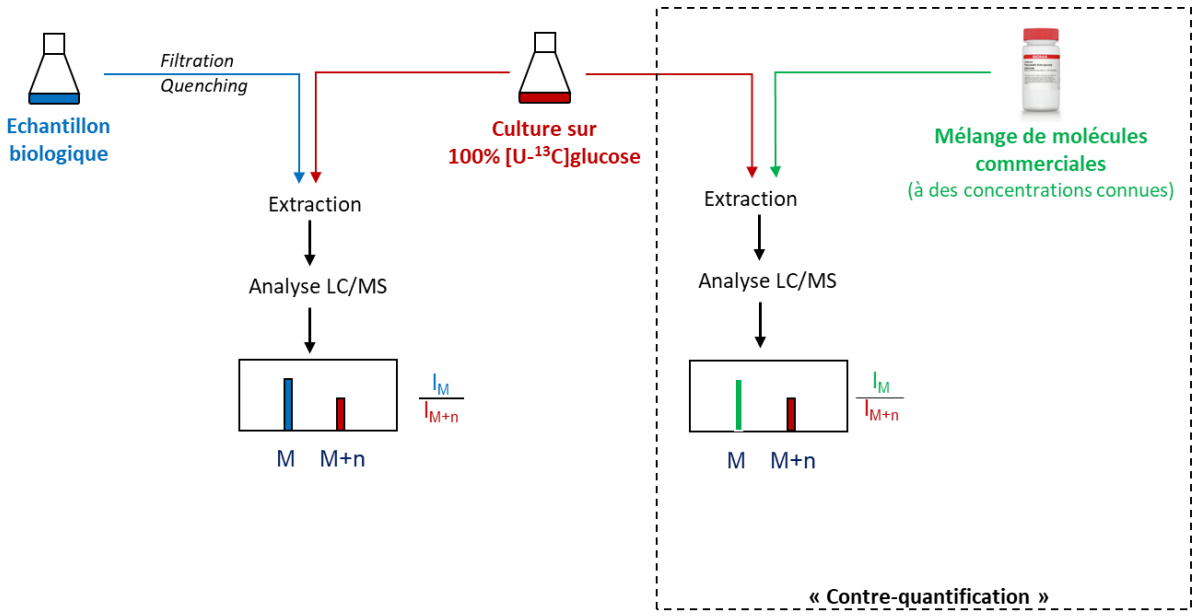


Figure 1.17 : Quantification des métabolites par la méthode IDMS (Isotopic Dilution Mass Spectrometry)

L'ajout de standard interne IDMS peut poser des difficultés lors de l'analyse MS, lorsqu'une molécule uniformément marquée (de masse M+n) possède la même masse qu'une autre molécule non marquée (M0). Ces deux molécules possèdent la même masse et génèrent donc une coélution sur le spectre de masse. Cet aspect peut interférer lors de la quantification absolue du métabolite d'intérêt. Ces différents inconvénients sont à considérer et doivent être évalués avant l'utilisation du standard IDMS.

Récemment, une approche basée la production d'un standard doublement marqué au $^{13}C^{15}N$ a été développée pour permettre de mesurer simultanément la concentration absolue des

métabolites et l'incorporation isotopique par spectrométrie de masse [Heuillet et al, 2020]. Cette stratégie de marquage a également l'avantage de réduire la quantité d'échantillons à collecter et analyser en doublant les informations obtenues via une seule expérience de marquage. Elle reste cependant limitée à la quantification de composés azotés.

5. Etude du métabolisme par traçage isotopique

Le deuxième axe d'utilisation des isotopes est consacré aux études de traçage isotopique, qui visent à identifier les réactions et les voies métaboliques actives dans un organisme. Le principe général de ces approches est schématisé sur la Figure 1.18. Elles reposent sur l'administration d'un substrat (ou nutriment) marqué au ^{13}C (ou autres isotopes tels que ^{15}N, ^{2}H…) dont on peut identifier le devenir métabolique grâce à l'incorporation de l'isotope dans les molécules filles [Christensen and Nielsen, 1999 ; Violante et al, 2019 ; Dong et al, 2022]. Le devenir de ce traceur isotopique est suivi en mesurant son incorporation au sein des métabolites de l'organisme, le plus couramment par spectrométrie de masse ou RMN (Section 3.4.1). L'analyse des profils d'incorporation du marquage ^{13}C des différents métabolites permet d'identifier les voies métaboliques et la topologie du réseau, mais également de mesurer les flux métaboliques (Section 5.2 et 7). Nous allons dans un premier temps détailler le protocole expérimental permettant de mettre en place ces approches (design expérimental, stratégies d'analyses, etc). Comme la thèse se concentre sur l'apport de la HRMS pour les études de traçage, les aspects spécifiques à cette technique seront développés mais comme nous le verrons tout au long du manuscrit, certains aspects sont génériques et peuvent intégrer des données issues d'autres techniques analytiques, telle que la RMN.

Dans la suite, le protocole sera développé pour le cas d'études de cellules *in vitro*. Le traçage peut aussi être appliqué à des tissus *ex vivo* ou à des organismes *in vivo*. Si le principe général des expériences reste le même, les conditions expérimentales, le choix et les modes d'adminitration du traceur, de collecte et de préparation des échantillons sont très différents.

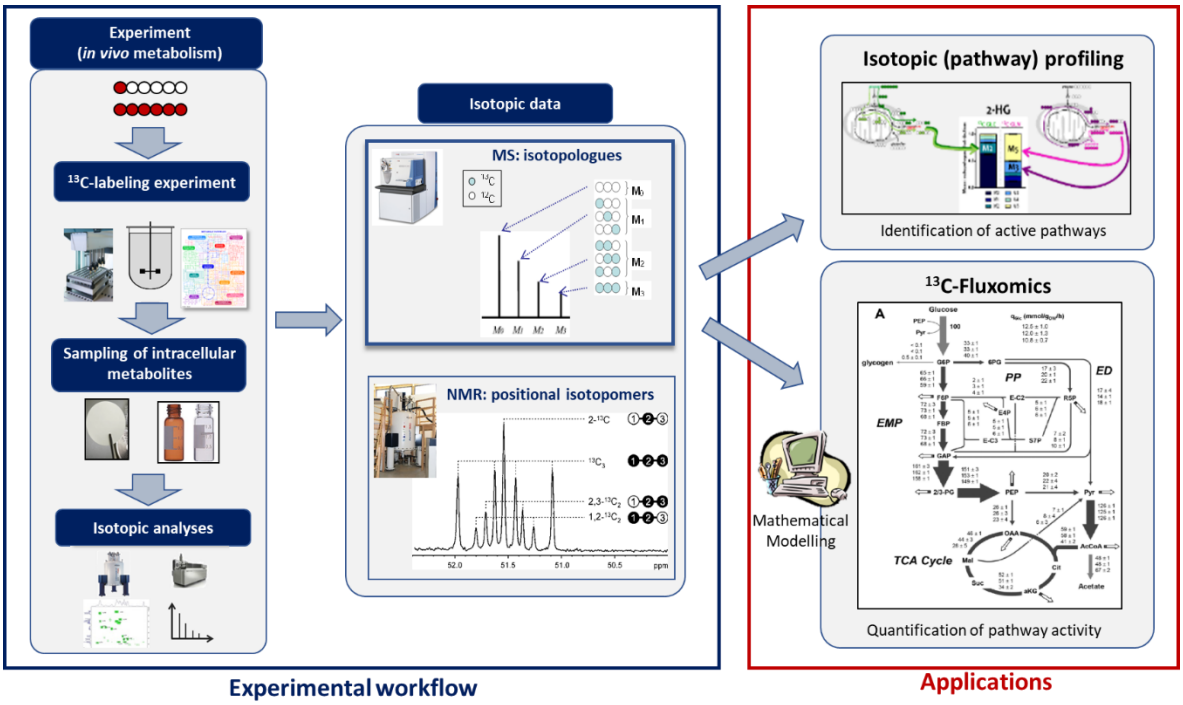


Figure 1.18 : Principe général des approches de traçage isotopique pour identifier et quantifier les activités des voies métaboliques dans un organisme.

5.1.Protocole expérimental

Le protocole expérimental des expériences de traçage isotopique par spectrométrie de masse suit un enchaînement d'étapes communes avec les approches de métabolomiques classiques (extraction, préparation d'échantillons, acquisition et traitement des données, analyses statistiques) (Figure 1.19). Les particularités des approches isotopiques sont la mise en place d'une expérience de marquage, au cours de laquelle les cellules sont cultivées sur un substrat spécifique marqué isotopiquement, et la génération de données isotopiques (CID, enrichissement isotopique). L'ensemble des étapes énoncées ci-dessus doit être adapté aux spécificités apportées par la dimension isotopique (section 3.5).

En métabolomique par spectrométrie de masse, on distingue deux grandes stratégies d'analyse : les approches ciblées, pour lesquelles on cible un ou plusieurs métabolite(s) d'intérêt bien défini(s), et les approches non-ciblées pour couvrir l'ensemble du métabolome détectable. Dans le protocole présenté ci-dessous, les trois étapes différenciant ces deux stratégies d'analyse sont le choix du mode d'acquisition par MS (mode full scan), la procédure de traitement des données et la génération des données isotopiques (non corrigées si les métabolites n'ont pas été identifiés). Dans les approches non-ciblées, le traitement, l'analyse et l'interprétation des ensembles de données complexes qui sont générés constituent encore des défis majeurs.

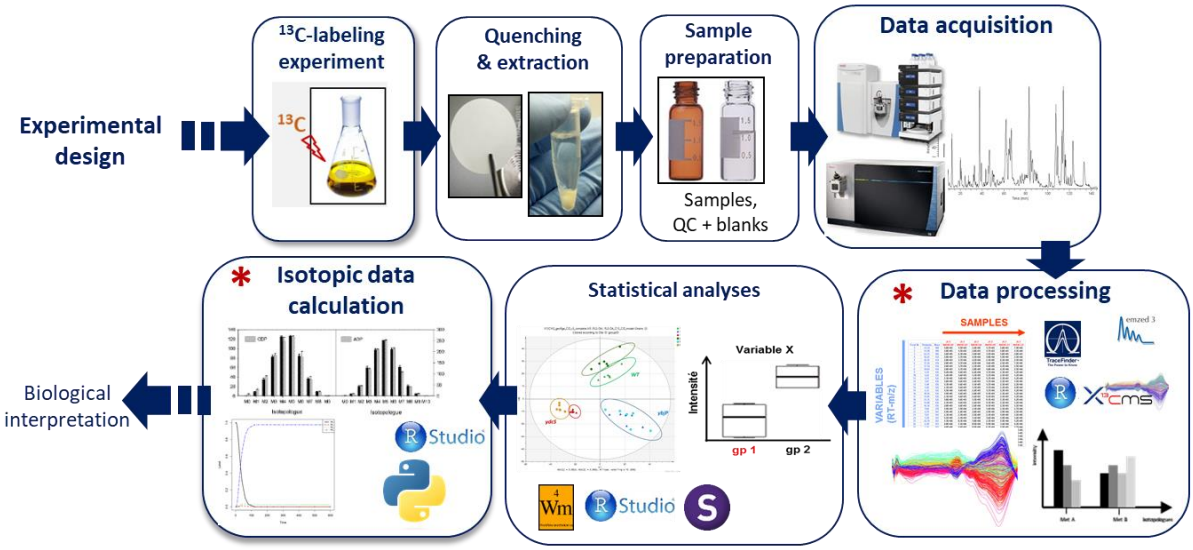


Figure 1.19 : Protocole des expériences de marquage isotopique C13 par spectrométrie de masse. Les astérisques (*) repèrent les étapes différant entre les approches isotopiques ciblées et non-ciblées.

5.1.1. Design expérimental

Le design expérimental est une étape clef pour maximiser et optimiser les informations isotopiques récupérées par les approches de traçage isotopique, tout en limitant les efforts et les coûts liés à la mise en place de l'expérience. Le design des expériences de marquage isotopique consiste à préparer chacune des étapes en amont de notre expérience, à savoir la culture, l'échantillonnage, le choix du mode d'analyse et le traitement des données. Ces différentes étapes vont dépendre principalement de la question biologique et de l'organisme étudié. En particulier, si le principe général des expériences reste le même, les conditions expérimentales, le choix et les modes d'adminitration du traceur, de collecte et de préparation des échantillons sont très différents suivant le modèle biologique étudié et le niveau d'intégration biologique considéré (celle, tissu, organisme). Le design d'études de traçage réalisées *in vivo* chez l'animal ou l'Homme diffère ainsi grandement de celui adapté à l'étude de cellules *in vitro*. Dans la suite, le protocole sera développé pour ce dernier cas, qui a été mis en œuvre dans la thèse.

5.1.1.1. Culture cellulaire

La mise en place de la culture cellulaire consiste à déterminer dans quelles conditions les cellules seront cultivées, notamment la composition du milieu de culture, le choix du traceur isotopique et l'état métabolique dans lequel les cellules vont se trouver au moment de l'échantillonnage.

- ***Composition du milieu***

Le milieu de culture va permettre à la cellule de se multiplier et de produire des composés. Chez *E. coli* par exemple, on va pour cela retrouver dans un milieu de cellule des composés inorganiques qui vont alimenter les cellules en azote, phosphore, souffre et en ions métalliques mais aussi des composés organiques, notamment une source de carbone, des vitamines ou des hormones. La composition du milieu depend de plusieurs paramètres tels que l'organisme étudié, le process de culture et peut également dépendre de la question biologique. Le milieu peut être simple et contenir une seule source de carbone – on parle alors de milieu minimum – ou complexe avec plusieurs sources de carbone – on parle dans ce cas de milieu riche. Dans le cadre d'expériences de marquage isotopique, il est nécessaire de connaitre précisément la composition du milieu, c'est à dire les composés qui le constitue ainsi que leur concentration. L'idéal est de cultiver les cellules dans un milieu minimum, ce qui simplifie grandement les expériences de marquage. Cet aspect est généralement applicable pour des organismes simples (ex : *E. coli*). Les organismes complexes (ex : cellules mammifères) consomment plusieurs sources de carbone et nécessiteront un milieu riche, ce qui complexifie l'expérience notamment car les cellules peuvent consommer plusieurs substrats en parallèle, avec des cinétiques de consommation différentes. Dans ce cas, plusieurs expériences de marquage peuvent être nécessaires avec chacune des sources de carbone consommées par la cellule (ex : glucose et glutamine pour les cellules mammifères).

- ***Choix du nutriment***

Le choix du nutriment marqué dépend principalement du type d'organisme étudié, de l'objet de l'étude et des voies métaboliques d'intérêt [Crown et al, 2012]. Il est important que l'assimilation du substrat marqué soit assurée par l'organisme et qu'il soit intégré dans le métabolisme aussi près que possible de la / des voie(s) étudiée(s). A titre d'exemple, le ^{13}C-glucose est très classiquement utilisé pour étudier la glycolyse et la voie des pentoses-phosphates. Le cycle de Krebs peut être étudié directement à partir de ^{13}C-pyruvate ou de ^{13}C-acétate, mais peut également être étudié à partir de sources plus éloignées telles que le ^{13}C-glucose, la ^{13}C-glutamine, un acide gras marqué ou encore à partir d'une combinaison de plusieurs d'entre eux. Le choix du nutriment marqué dépend également de la nature du système biologique étudié. Ainsi, pour l'étude de cellules mammifères, le glucose et la glutamine sont privilégiés car ce sont des nutriments majeurs pour ces cellules et ils donnent accès à l'essentiel des voies du métabolisme carboné central [Crown et al, 2012 ; Antoniewicz, 2013]. D'autres

substrats marqués comme le ^{13}C-lactate, ou le ^{13}C-pyruvate sont également largement utilisés dans ce cadre. D'un point de vue pratique, le prix et la disponibilité du traceur utilisé sont également pris en compte.

- ***Traceur isotopique / choix du marquage***

Le type de marquage du composé lui-même, c'est-à-dire le nombre et la position des atomes remplacés par l'isotope est également déterminant pour le nombre et la nature des flux accessibles et la précision de l'estimation des flux. A titre d'exemple, il a été montré que le [1,2-^{13}C]-glucose est un traceur privilégié pour une estimation précise de la partition des flux entre la glycolyse et la voie des pentoses phosphates [Metallo et al, 2009]. Différentes études ont été menées pour déterminer le type de marquage optimal [Boros et al, 2005 ; Munger et al, 2008], notamment dans les systèmes complexes tels que les cellules mammifères [Metallo, et al. 2009]. Des logiciels tels que IsoDesign [Millard et al, 2014] ont été spécifiquement développés pour optimiser le traceur en fonction de la question biologique à traiter, notamment dans le cadre d'analyses de flux métaboliques. La combinaison de plusieurs marquage peut également augmenter le nombre d'information obtenu et augmenter la précision sur la mesure des flux métaboliques. La publication de Crown et al. expose les différents marquages possibles du glucose et leur impact sur la mesure des flux chez *E.coli* pour l'ensemble des voies du métabolisme central [Crown et al, 2016].

- ***État métabolique***

Le dernier point à considérer est l'état métabolique dans lequel se trouvent les cellules au moment de l'échantillonnage. On parle d'état stationnaire métabolique lorsque la concentration des métabolites au cours du temps ne varie pas au sein de la cellule. A contrario, un état métabolique dynamique signifie que la concentration d'un métabolite donné varie au cours du temps. Il existe des conditions de culture pour lesquelles l'état stationnaire métabolique ne sera jamais atteint. Dans ce cas là, il est possible mais difficile de réaliser des expériences de fluxomique. Les cellules peuvent croître en suspension ou adhérer. Si l'on prend l'exemple de cellules en suspension, il existe trois principaux modes de cultures (Figure 1.20). Les cultures en batch consistent à alimenter les cellules avec une quantité définie de source de carbone au départ et les laisser se multiplier jusqu'à consommation complète de la source de carbone. Dans ce type d'expérience, l'état stationnaire métabolique est atteint uniquement pendant la phase exponentielle de croissance. Le fedbatch consiste à donner le même volume

de source de carbone au cours du temps alors que le nombre de cellules se multiplient au fur et à mesure. Dans ce type de culture, l'état stationnaire métabolique n'est jamais atteint. Enfin les cultures continues consistent à amener en continu une source de carbone et à l'enlever à la même vitesse du milieu et des cellules. Dans ce cas-là, l'état stationnaire métabolique est atteint dès que le système est stabilisé.

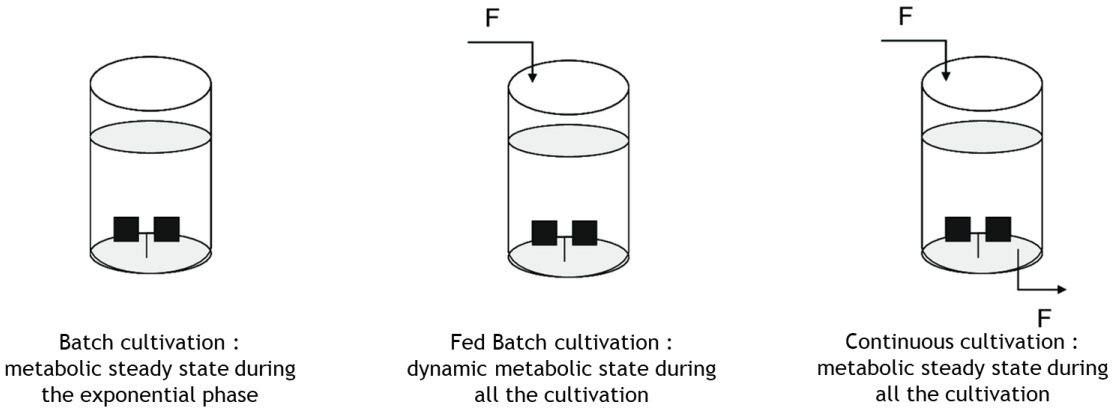


Figure 1.20 : Exemples de systèmes de culture pour les cellules en suspension.

5.1.1.2. Échantillonnage

Le choix de la méthode d'échantillonnage est déterminé en fonction de l'organisme étudié et des métabolites d'intérêts. Il est possible de réaliser des expériences de marquage sur différents types d'organismes : procaryotes ou eucaryotes. En fonction du type cellulaire utilisé, les méthodes d'échantillonnage vont varier et peuvent être plus ou moins complexes.

En fonction de la question biologique, la composition en solvant de la solution d'extraction, sa température et son pH doivent être adaptés. Le choix de la méthode d'extraction peut influer sur le type et la quantité de métabolites extraits. Les métabolites sont extraits des cellules par broyage dans un solvant. La nature de ce solvant va dépendre des propriétés physico-chimiques des composés d'intérêt. Par exemple un solvant polaire (MeOH, EtOH, ACN) sera utilisé pour extraire des composés polaires. Une analyse des différentes méthodes d'extraction a été réalisé chez la levure [Canelas et al, 2009] ou chez *E. coli* [Winder et al, 2008 ; Millard et al, 2014]. De manière générale, il est préférable de limiter le nombre d'étapes dans la procédure d'extraction pour limiter les biais. Pour les approches de profilage non-ciblées, une méthode d'extraction non séléctive est souhaitée pour détecter la plus large gamme de métabolites possibles.

L'échantillonnage des intermédiaires métaboliques est une étape critique afin d'obtenir une information isotopique fiable. Il a pour but d'isoler l'échantillon, comme par exemple pour

isoler les cellules du surnageant de culture. En fonction de la matrice de l'échantillon à collecter, différents protocoles peuvent être utilisés comme par exemple la fast filtration ou la centrifugation [Bolten et al, 2007 ; da Luz et al ; 2013 ; Millard et al, 2014]. Le métabolome étant instable et l'activité enzymatique continuant ex-vivo, il est nécessaire de bloquer le plus rapidement possible l'activité enzymatique pour étudier le métabolisme. Cette étape est nommée étape de quenching. Elle peut être réalisée par congélation rapide en utilisant notamment de l'azote liquide ou par trempe dans du méthanol (éthanol ou autre solvant) froid (-40°C) ou directement dans un solvant d'extraction. Le processus de trempe au méthanol peut cependant endommager la membrane cellulaire et provoquer une fuite importante de métabolites intracellulaires dans le milieu. La filtration rapide est une méthode d'échantillonnage qui consiste à collecter en quelques secondes les cellules par filtration sous vide. Cette méthode est appliquée avant l'extraction et permet de réduire la perte de métabolites [Bolten et al, 2007]. Il est très important que cette étape soit réalisée le plus rapidement possible. En effet, le turnover les métabolites intracellulaires est très rapide. Par exemple, chez *E. coli*, il est de quelques secondes pour la glycolyse et de 30 minutes pour le cycle de krebs. Une préparation rapide évitera également un stress cellulaire qui pourrait impacter le métabolisme. Comme toute expérimentation, la préparation doit être homogène pour tous les échantillons. Il est crucial de respecter les températures et maintenir les échantillons au froid tout au long de la préparation pour éviter la dégradation ou la reprise du métabolisme. L'analyse des surnageants de culture permet quant à elle d'identifier les metabolites extracellulaires produits ou consommées par les cellules, et d'en analyser le marquage.

L'état isotopique des cellules doit être également considéré. Lorsque le marquage d'un métabolite est stable au cours du temps on parle d'état stationaire isotopique. Lorsque l'on peut atteindre cet état, un seul point d'échantillonnage sera nécessaire. Si le marquage varie au cours du temps, on parlera d'état isotopique instationnaire, et il sera nécessaire de procéder à une cinétique d'échantillonnage afin de suivre l'évolution du marquage pendant l'expérimentation. Ces différents états isotopiques seront détaillés dans la section 7.2 consacrée aux expériences de fluxomique.

5.1.2. Stratégies d'analyses par LC-MS

Le choix du mode d'analyse ayant été précedemment décrit (section 3.5), nous allons nous concentrer sur les différentes stratégies d'analyse utilisées en LC/HRMS.

5.1.2.1. Analyses ciblées

L'analyse métabolomique ciblée est l'analyse quantitative d'un groupe particulier de métabolites, correspondant soit à une famille physico-chimique, soit à différents métabolites impliqués dans une voie métabolique particulière. Elle nécessite de connaître en amont les composés d'intérêts, définis sur la base d'une question biologique précise et d'une base de données disponible pour l'identification des composés ciblés. Les approches ciblées sont plus sensibles de par leur sélectivité car la connaissance des composés recherchés permet d'adapter les procédures expérimentales (extraction, méthode analytique) en fonction des propriétés des métabolites recherchés.

Les progrès des instruments LC/MS et des outils de traitement de données ont permis de grandes avancées dans l'analyse métabolomique ciblée. Elle permet désormais la mesure précise de dizaines ou de centaines de métabolites dans des échantillons biologiques complexes. Les spectromètres de masse hybrides basse résolution type QqQ et QTrap sont largement utilisés dans les approches ciblées. Ces deux instruments s'appuient sur le mode de détection MRM (Multiple Reaction Monitoring) qui est la méthode la plus sensible pour la quantification de métabolites dans un mélange complexe [Sherwood et al, 2009]. Ce mode est adapté à des spectromètres de masse à triple quadripôle (Figure 1.21). Dans ce mode, l'ion d'intérêt est d'abord sélectionné dans le premier analyseur Q1. Il est ensuite fragmenté dans le second analyseur Q2 pour produire une série d'ions fils. Un (ou plusieurs) de ces ions fils peut être sectionné dans le dernier analyseur Q3 à des fins de quantification. Ce mode d'acquisition présente une double sélectivité. En ignorant tous les autres ions qui entre dans le spectromètre de masse, l'expérience gagne en sensibilité. Il est particulièrement adapté pour la mesure de biomarqueurs, en santé notamment pour la compréhension des mécanismes ou des reprogrammations métaboliques liées à certaines pathologies. Les spectromètres de masse haute résolution sont également utilisés pour la quantification métabolique ciblée [Doerfler et al, 2014 ; Dunn et al, 2013 ; Zhou et al, 2016].

Dans le cadre d'analyses ciblées, les données sont prédéfinies (métabolites connus et transitions MRM). Le profilage isotopique ciblé s'avère plus complexe. Un fragment marqué possède un parent par isotopologue. Dans le cas où les fragments contiennent des atomes de carbone, il est donc nécessaire de multiplier le nombre de transitions pour obtenir l'ensemble du massif isotopique. Les transitions dépendent de la molécule et du (des) fragment(s) fils auxquels on a accès.

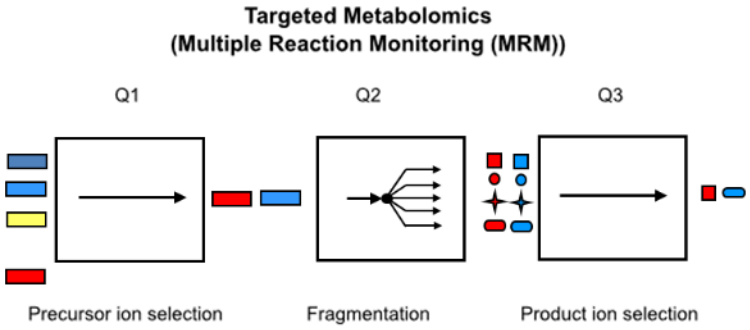


Figure 1.21 : Illustration du mode de fragmentation MRM. Dans le cadre d'analyses de profilage isotopique, le nombre de fragments avec des atomes de C augmente le nombre de transitions.

5.1.2.2. Analyses non-ciblées

Les analyses non-ciblées sont utilisées pour couvrir l'ensemble du métabolome. Contrairement aux analyses ciblées, ces méthodes enregistrent chaque molécule ionisable de l'échantillon et ont donc une capacité de profilage bien supérieure. Leur utilisation est adaptée à l'analyse des réseaux métaboliques à large échelle car elle permet de fournir une information plus exhaustive et sans *a priori* sur le métabolome d'un organisme.

L'utilisation de spectromètres de masse haute résolution est privilégiée dans le cadre d'analyses non ciblées. En effet, les mélanges complexes contiennent généralement des centaines de métabolites ayant des différences de masse étroites et une résolution d'au moins 0,1 mDa est nécessaire pour permettre une bonne séparation de tous les ions générés [Marshall et al., 2008]. La complexité du métabolome rend compliquée la mesure de tous les métabolites via une seule analyse, et il peut être nécessaire de combiner plusieurs méthodes d'extraction, plusieurs colonnes chromatographiques (ex : C18, HILIC) ou plusieurs modes d'ionisation (positive/négative) pour optimiser la récupération des métabolites.

Les approches non-ciblées sont globalement utilisées pour l'identification sans *a priori* de biomarqueurs liés à une perturbation métabolique ou pour identifier de nouveaux métabolites et de nouvelles voies métaboliques. Les principaux défis liés aux approches non-ciblées résident dans les protocoles de traitement de grandes quantités de données et d'identification des métabolites. Cette étape d'identification reste la principale difficulté, liée aux bases de données métabolomiques qui ne permettent pas d'identifier sans ambiguïté tous les signaux obtenus lors de l'analyse. Malgré les excellents progrès réalisés dans l'élargissement des bases de données (section 3.5), une partie importante des signaux détectés dans les études non ciblées ne peut toujours pas être identifiée et reste non annotée [Zamboni et al, 2015].

Contrairement à la métabolomique non-ciblée qui est aujourd'hui largement utilisée pour l'étude du métabolisme et pour laquelle des méthodes et outils sont bien référencés, le profilage isotopique non-ciblé reste un domaine relativement récent et sous-exploité [Dange et al, 2020]. Dans l'ensemble, l'analyse non-ciblée isotopique peut être complexe et le résultat peut être difficile à interpréter.

5.1.3. Traitement des données isotopiques

Les analyses de traçage isotopique par spectrométrie de masse génèrent un grand nombre de données (Figure 1.22). Elles nécessitent l'utilisation de protocoles et d'outils de traitements de données performants et adaptés pour extraire l'information pertinente des spectres de masse obtenus.

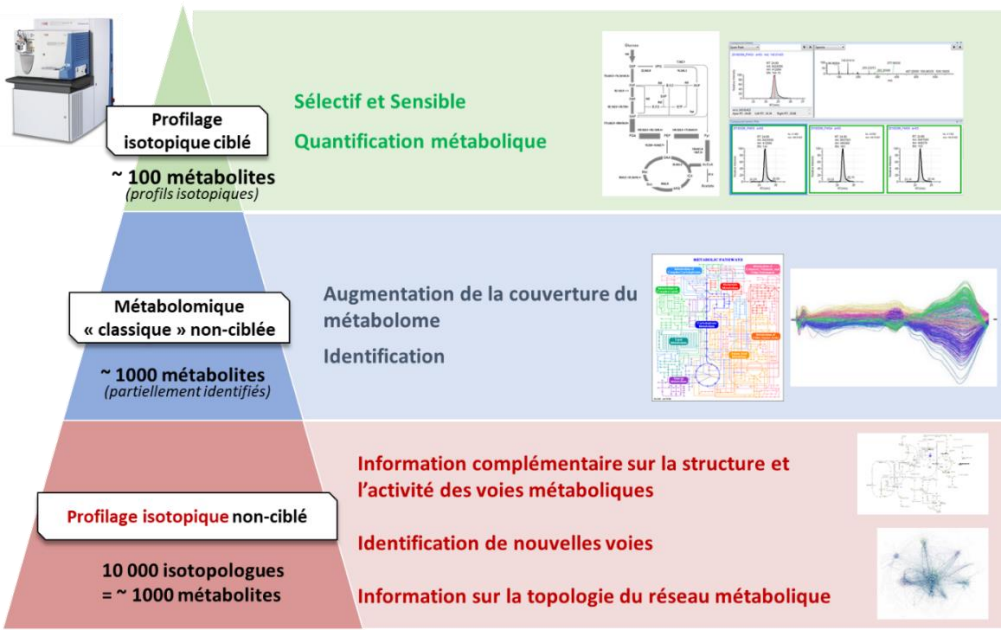


Figure 1.22 : Ordre de grandeur des données générées en LC/MS par des approches de profilage isotopique ciblées, de métabolomique non-ciblées, de profilage isotopique non-ciblées et informations apportées par ces différentes approches pour l'étude du métabolisme.

Le traitement des données HRMS issues d'expériences de marquage se décompose en plusieurs étapes pour extraire et filtrer les signaux de masse, générer les données isotopiques (mesure des CID et de l'enrichissement isotopique) et exploiter ces informations selon l'objet de l'étude : recherche de biomarqueurs, identification de nouvelles voies métaboliques ou calcul de flux. Les approches ciblées et non-ciblées diffèrent principalement par la nature et la quantité d'informations générées ainsi que dans le traitement de données à appliquer pour extraire l'information (Figure 1.23).

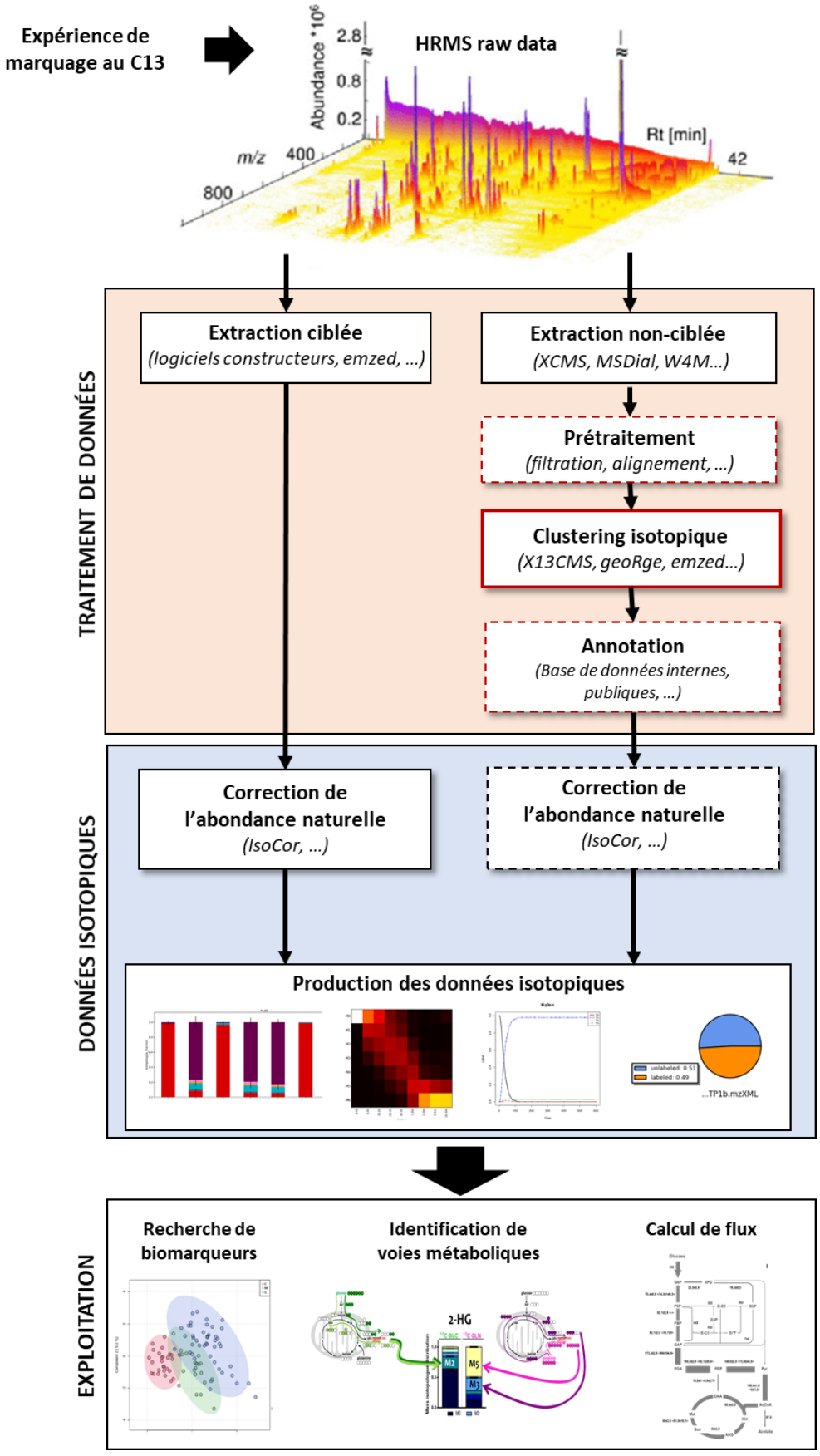


Figure 1.23 : Comparaison des étapes de traitement de données HRMS issues d'expériences de marquage via des analyses ciblées (à gauche) et non-ciblées (à droite). Les encadrés en pointillés représentent les étapes optionnelles et les encadrés rouges représentent les étapes spécifiques du traitement non-ciblé.

La première étape consiste à extraire les données des spectres de masse. On distingue l'extraction ciblée qui permet de récupérer directement le profil isotopique complet de métabolites d'intérêts prédéfinis, de l'extraction non-ciblée qui permet d'extraire l'ensemble des features (mz, RT, intensité) détectées dans les spectres de masse.

De nombreux logiciels constructeurs et *open source* ont été développés pour extraire ces données. Les premiers ont l'avantage d'être relativement simples d'utilisation mais restent limités aux formats de fichiers spécifiques du spectromètre de masse du constructeur (ThermoFischer : TraceFinder, Waters : MassLynx…) et encore peu d'entre eux permettent d'extraire des données acquises via des approches non-ciblées (Thermo : CompoundDiscoverer). Les logiciels *open source* quant à eux permettent de traiter les données après conversion des données brutes dans un format libre (.mzXML, .mzML, .NetCDF) [Kessner et al, 2008]. On distingue les logiciels permettant l'extraction spécifique (manuelle ou semi-automatique) de massifs isotopiques prédéfinis dans le cadre d'analyses ciblées (emzed [Kiefer et al, 2013], mzMatchIso [Chokkathukalam et al, 2013]) de ceux permettant l'extraction automatique de l'ensemble des features dans le cadre d'analyses non-ciblées (XCMS [Tautenhahn et al, 2012], MSDial [Tsugawa et al, 2015], …).

L'extraction des données brutes issues d'analyses isotopiques non-ciblées est particulièrement complexe. Outre la difficulté liée à la quantité de données générées par les approches non-ciblées, la complexité provient de la donnée isotopique en elle-même. L'impact de la présence des isotopes sur le spectre de masse est double et on pourra observer (i) une forte augmentation de la quantité de pics dans les spectres MS et (ii) une baisse globale de l'intensité de ces pics. En effet, les spectres MS recueillis sur du matériel marqué sont plus complexes à traiter que ceux issus du (même) matériel non marqué car tous les isotopologues de chaque métabolite sont susceptibles d'être générés et détectés. De plus, l'intensité totale du signal MS d'un analyte donné reste la même, que le composé soit marqué ou non. Pour un même analyte, dans l'échantillon marqué, l'intensité totale sera donc distribuée sur l'ensemble des isotopologues détectés (Figure 1.24A). Les spectres MS des composés marqués contiennent donc plus de signaux, chacun ayant une intensité plus faible que dans les spectres des composés non marqués correspondants. L'analyse étant non-ciblée, elle ne permet pas d'accéder directement au profil isotopique complet des métabolites, contrairement à l'analyse ciblée qui se base sur une liste de composés identifiés.

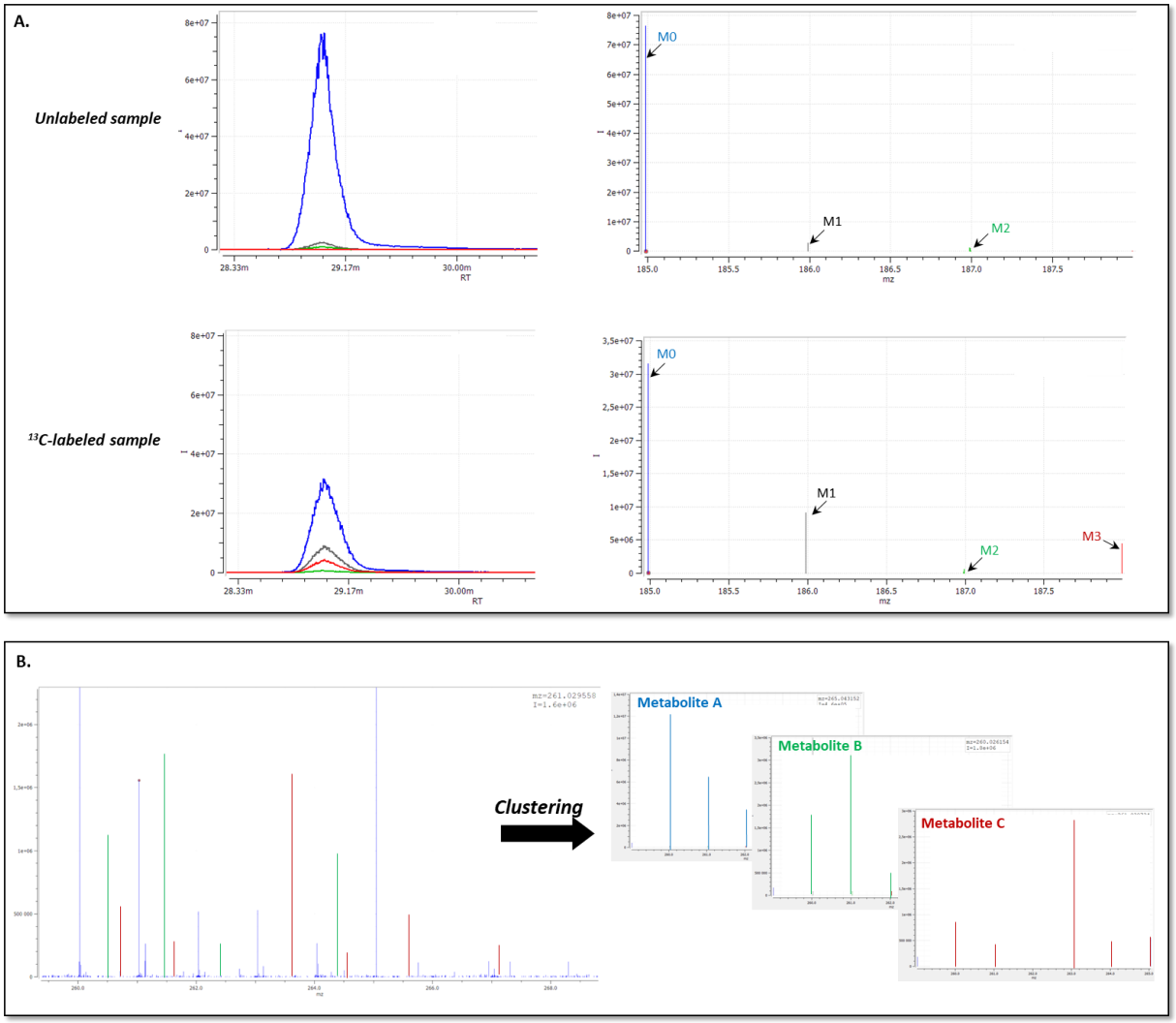


Figure 1.24 : A. Comparaison de spectres de masse d'un composé (2/3PG) non marqué et marqué au C13. B. Etape de groupement isotopique à partir d'un spectre de masse obtenu par des analyses non -ciblées sur du matériel marqué. Les données ont été extraites à l'aide du logiciel emzed [Kiefer et al, 2013].

La deuxième étape – optionnelle – du traitement non-ciblé consiste à filtrer les signaux polluants parmi l'ensemble des features extraites. Les approches non-ciblées sont utilisées dans l'objectif d'obtenir l'information la plus exhaustive possible et sont généralement associées à un nombre d'échantillons plus conséquent. Les jeux de données résultants contiennent généralement des milliers de pics de masse et peuvent être « pollués » par la présence d'artefacts, de contaminants, d'adduits ou de fragments. Pour pouvoir être exploitées correctement, ces données nécessitent généralement une étape supplémentaire de prétraitement permettant de passer des données instrumentales brutes à des données propres pour le traitement des données. Cela inclut des étapes de filtrations (S/N ratio, Bio/Blanc, …) pour nettoyer les données et éliminer les signaux non exploitables (<LOQ) et de réalignement du signal en cas de dérive analytique. Ces différentes étapes alourdissent considérablement le traitement des

données et peuvent être sources de perte d'informations. Elles nécessitent d'être bien dosées pour filtrer les signaux non exploitables et limiter l'information analytique redondante sans perte d'information métabolique. On retrouve des outils permettant la filtration et l'alignement automatique des signaux (mzMatch.R [Creek et al, 2012] ; mzMatchIso [Chokkathukalam et al, 2013].

La troisième étape, spécifique aux approches non-ciblées, consiste en un groupement automatique des signaux de masse extraits afin d'obtenir le profil isotopique complet des molécules détectées. Soit, pour un métabolite donné, son pic monoisotopique (M0) et les isotopologues associés (M1, M2…Mn) (Figure 1.24B). En effet, les features sont extraites de manière totalement indépendante les unes des autres, et la difficulté initiale réside dans le fait de pouvoir identifier et regrouper toutes celles qui appartiennent à un même massif isotopique, pour l'ensemble des analytes détectés. Bien qu'en évolution constante, le nombre de logiciels permettant de réaliser cette reconstruction de massifs isotopiques reste restreint et tous nécessitent encore des étapes importantes de curation manuelle des données. Le choix du logiciel dépend de plusieurs facteurs incluant l'objet de l'étude (identification, profilage, quantification), la stratégie de marquage appliquée (substrat uniformément marqué ou non) ou encore le choix de l'isotope utilisé. Cette étape est également très dépendante notamment du logiciel utilisé lors de la première étape pour extraire les features. Une analyse comparative des logiciels existants a été réalisée à partir de critères définis (Annexe 2). Une comparaison de certains de ces logiciels a été effectuée par Dange et al, 2020.

Dans le cadre de recherche de biomarqueurs, des analyses statistiques telles que des approches univariées ou multivariées peuvent être appliquées. Ces approches ne seront pas développées dans ce manuscrit mais elles peuvent être utilisées pour comparer les profils isotopiques de composés dans différentes conditions ou pour différentes souches ou lignées cellulaires.

Bien que les approches isotopiques non-ciblées se développent de plus en plus, des protocoles ou méthodologies permettant de garantir la qualité des données extraites font encore défaut. L'amélioration des stratégies de traitement non ciblé des données isotopiques a représenté un des deux objectifs de ma thèse et les résultats obtenus font l'objet du Chapitre 2 de ce manuscrit de thèse.

5.2. Applications des approches de traçage isotopique

5.2.1. Profilage isotopique

Comme cela a déjà été mentionné, la connaissance des concentrations en métabolites ne renseigne pas forcément sur la nature ou l'activité des voies métaboliques actives dans les conditions étudiées.

Le profilage isotopique est défini comme la mesure quantitative des différentes espèces isotopiques d'une même molécule. Les profils de marquage des différents métabolites obtenus à partir du traceur initial permettent l'identification des voies métaboliques en place dans le système étudié (c'est le principe de base du traçage isotopique). En effet le devenir des atomes de carbone est spécifique de chaque voie métabolique, en lien avec la nature et le mécanisme des réactions – et les transitions atomiques qui en découlent - qui s'y déroulent. Le traçage isotopique permet ainsi, à partir de l'analyse des profils isotopiques mesurés dans les différents métabolites détectés, de révéler l'activité des différentes voies métabolites et d'évaluer leur contribution à la production ou à la consommation de métabolites spécifiques. L'analyse du marquage ^{13}C des métabolites est ainsi précieuse pour identifier les voies et les réseaux métaboliques (nature des marquage observés) mais également pour apporter des informations sur leur activité (intensité des marquages). Des exemples d'exploitation de données isotopiques pour la caractérisation plus fine des réseaux métaboliques sont présentés ci-dessous.

5.2.1.1. Devenir des nutriments

Le traçage isotopique permet en premier lieu de connaître le devenir métabolique des nutriments utilisés par les cellules, à partir de l'incorporation du traceur dans toutes les molécules dérivées de celui-ci. A l'envers, cela permet de déterminer si le squelette carboné d'une molécule donnée est dérivé en tout ou partie d'un nutriment particulier, donc si ce nutrilent contribue ou non à la biosynthèse du composé. Pour cette analyse, l'utilisation d'un traceur isotopique totalement marqué (marquage uniforme) est recommandée. Un substrat marqué uniquement dans une position (marquage spécifique) peut perdre son carbone-13 lors d'une réaction de décarboxylation et la contribution du substrat aux voies métaboliques suivante ne peut pas être évaluée. Des stratégies combinant des expériences parallèles avec des substrats marqués spécifiquement pour différentes positions du nutriment peuvent cependant etre

envisagées pour déterminer le devenir de chaque atome de carbone de ce dernier [Crown et al, 2015].

La Figure 1.25A illustre le cas où deux sources de carbone contribuent à la formation d'un métabolite spécifique. La contribution relative de chacune de ces sources peut être déterminée par mesure directe de l'enrichissement isotopique obtenu après marquage de chacun des substrats. À l'état stationnaire isotopique, la somme des enrichissements obtenus à partir des deux substrats marqués doit être égale à 100% (Figure 1.25B). Si cette valeur n'est pas atteinte, cela peut signifier que d'autres sources contribuent à la synthèse du métabolite (Figure 1.25C), et donc que d'autres voies que celles déjà connues sont associées à sa synthèse. Cette analyse ne nécessite pas de connaître précisement le chemin emprunté par le substrat, mais permet de mesurer la fraction du métabolite formée à partir d'un nutriment donné à partir de la quantité de ^{13}C incorporé.

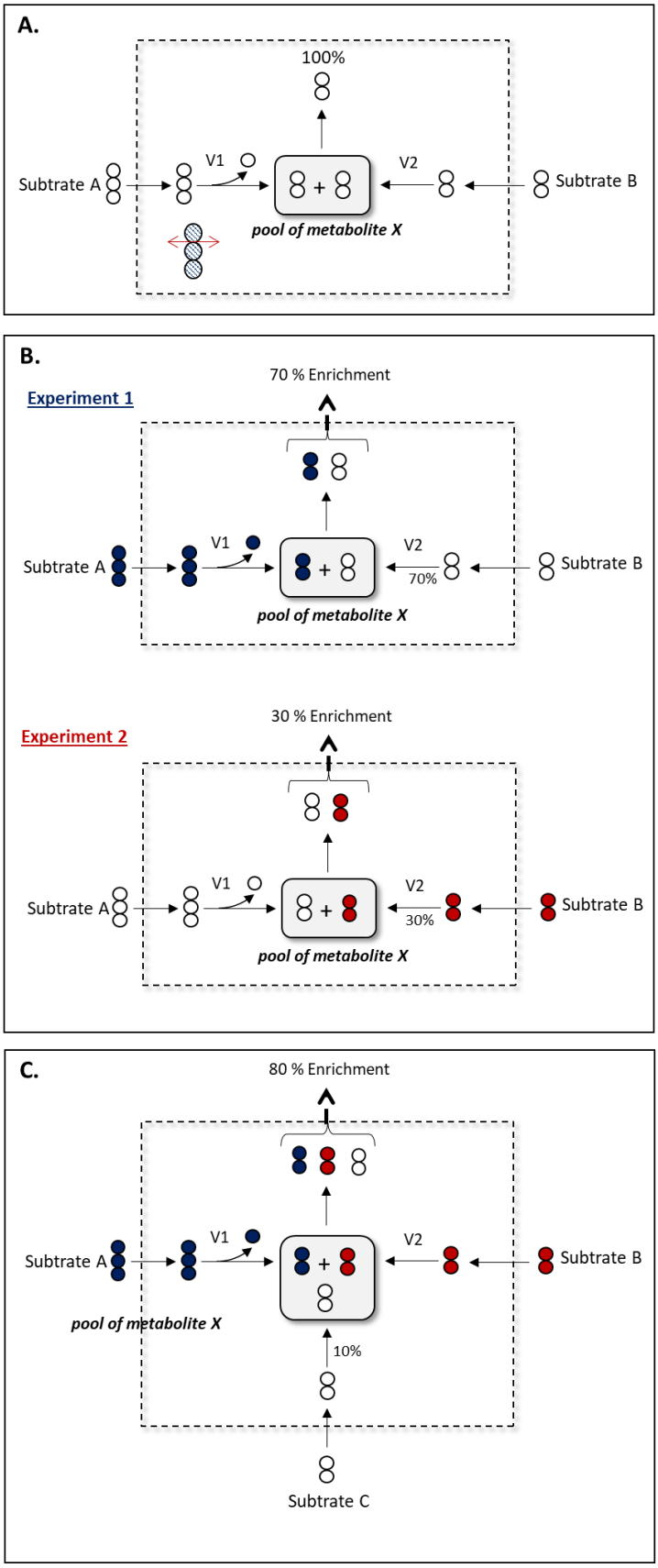


Figure 1.25 : Intérêt du traçage isotopique pour la mesure de la contribution de plusieurs sources de carbone à la formation d'un métabolite spécifique.

5.2.1.2. Élucidation de voies métaboliques

Le traçage isotopique est également utilisé pour identifier les voies métaboliques actives dans l'organisme étudié. Cela repose sur le fait que le devenir des atomes de carbone dans les voies métaboliques conduit à des profils de marquage caractéristiques dans les métabolites. A titre d'illustration, le devenir des atomes de carbone du glucose dans les 3 grandes voies glycolytiques est différent. En choisissant un marquage pertinent, et suivant la technique analytique utilisée, les trois voies peuvent être parfaitement discriminées à partir du profil isootpique du pyruvate ou de ses dérivés (Figure 1.26).

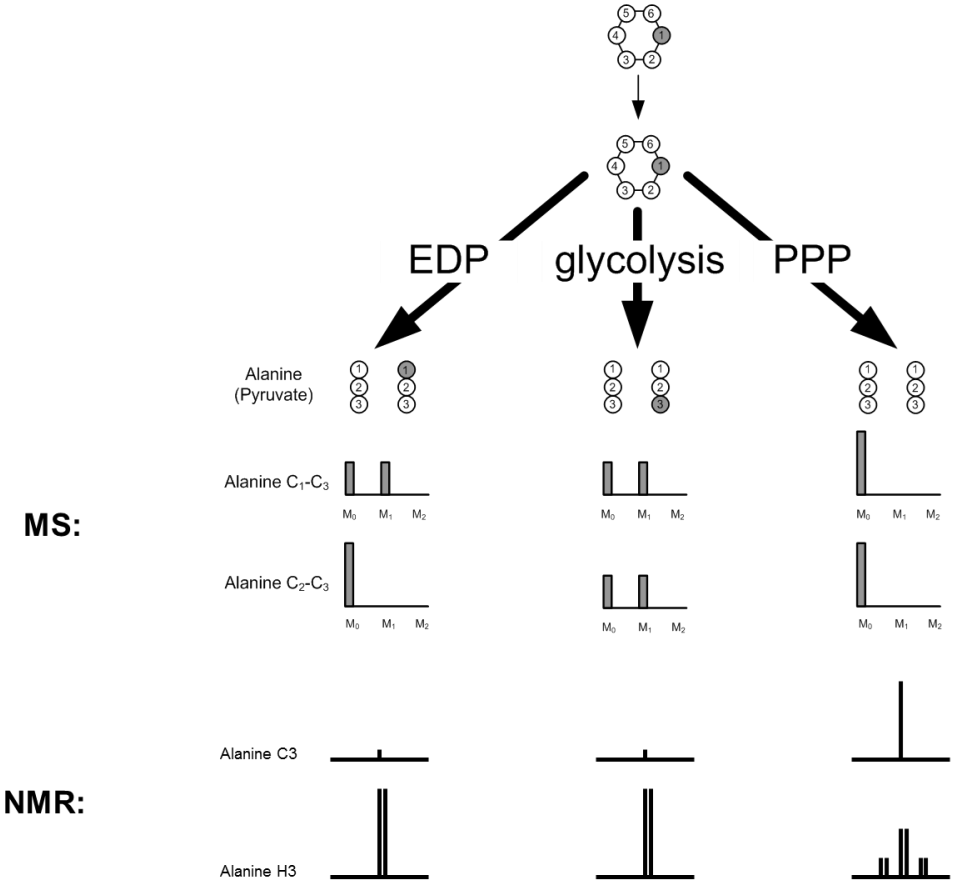


Figure 1.26 : Identification des 3 principales voies glycolytiques par traçage isotopique. Ces trois voies peuvent être discriminées en mesurant par MS ou RMN le profil isotopique du pyruvate ou de ses dérivés tels que l'alanine, après administration de [1-^{13}C]-glucose. EDP : voie d'Entner-Doudoroff, PPP : voie des pentoses-phosphates. (d'après Wittmann & Portais, 2013).

Comme on peut le voir sur la Figure 1.26, le devenir du C1 du glucose est différent dans chacune des voies glycolytiques. Il donne le C3 du pyruvate dans la glycolyse classique, le C1 dans la voie d'Entner-Doudoroff, et aucun par la voie de pentoses-phosphate. Le devenir spécifique des atomes de carbone au cours d'une réaction ou d'une voie métabolique est appelée transition atomique (atom mapping en anglais). Dans ce contexte, les données de traçage permettent non seulement de montrer l'activité de voies métaboliques connues, mais aussi, de

découvrir de nouvelles voies, lorsque les incorporations isotopiques observées ne peuvent pas être expliquées par des voies déjà connues. Comme exemple on peut citer la démonstration de voies métaboliques entièrement nouvelles telles que la voie de l'éthyl-malonyl-CoA dans certaines bactéries méthylotrophes [Peyraud et al, 2009], mais aussi l'identification de nouvelles combinaisons de réactions déjà bien connues telle qu'une nouvelle voie du catabolisme du glucose chez *E. coli* [Fisher and Sauer, 2003].

5.2.1.3. Partition de flux entre différentes voies métaboliques

Les données de marquage au ^{13}C permettent de mesurer la contribution de plusieurs voies métaboliques à la formation d'un même composé. Il faut pour cela que les substrats aient un marquage différent ou que les voies concernées aient une signature isotopique distincte (Figure 1.27). Il s'agit du principe de dilution isotopique qui est observé lorsque deux voies métaboliques convergent vers la formation du même métabolite : le marquage du métabolite formé par une voie est dilué par celui provenant de l'autre voie.

Le taux de dilution isotopique mesuré permet de calculer la proportion du métabolite formé par l'une et l'autre des deux voies, c'est-à-dire la distribution relative des flux de synthèse du composé entre les deux voies (Figure 1.27). Ce principe peut s'appliquer pour mesurer la contribution de deux voies ayant des substrats différents mais le même produit, mais aussi pour deux voies métaboliques ayant le même substrat et le même produit, c'est-à-dire deux voies dites parallèles [Christensen and Nielsen, 1999]. Cette analyse peut être particulièrement utile, en santé notamment, lorsque la production d'un métabolite est accrue dans un état pathologique. Sonder l'activité des voies participant à la production de ce composé permet de conduire à des stratégies thérapeutiques (comme une inhibition de voie).

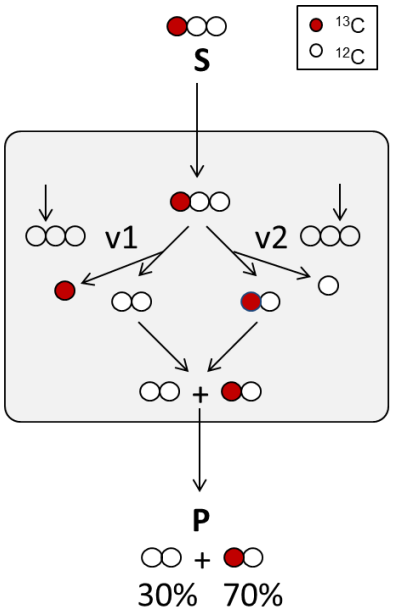


Figure 1.27 : Intérêt du principe de dilution isotopique pour mesurer des partitions de flux. Les fractions isotopiques des métabolites provenant de deux voies convergentes permettent d'estimer les flux relatifs de chacune des voies. Le marquage est perdu dans la voie 1 mais pas dans la voie 2. Les molécules du produit P formées par la voie 1 diluent le marquage des molécules de P formées par la voie 2. Le taux de dilution est directement proportionnel à la contribution respective des deux voies à la formation du produit. Dans le cas présenté, 30% des molécules de P sont formées par la voie 1 et 70% par la voie 2.

L'exemple ci-dessous présente la partition de flux entre deux voies parallèles : la voie des pentoses phosphates et la glycolyse qui produisent du pyruvate à partir du glucose. L'utilisation du glucose marqué en position 1 et 2 ([1,2-^{13}C]-glucose) permet de mesurer la contribution relative de chacune de ces voies dans la formation du pyruvate. Lorsque le [1,2-^{13}C-Glucose] est converti via la voie de la glycolyse, du pyruvate M+2 est formé. La voie des pentoses phosphates possède une branche oxydative et une branche non oxydative qui se connecte à la glycolyse à différents endroits [Brekke et al, 2012 ; Stincone et al, 2015]. Le C1 du glucose est éliminé sous forme de CO_2 dans la branche oxydative pour former du ribose-5-phophate M+1. Si celui-ci alimente la glycolyse via la branche non-oxydative, cela va produire du M+1 dans les métabolites glycolytiques (Figure 1.28A). La partition de flux entre la glycolyse et la voie des pentoses-phosphate peut être estimée sur la base des différents profils de marquage décrits ci-dessus. Les expériences de marquage fournissent un moyen simple de distinguer la partition entre les deux voies car le marquage du pyruvate produit est différent.

Ce type d'approche permet d'identifier des activités de voies initialement inattendues. Par exemple la détection de citrate M+5 dans les cellules mammifères alimentées en [U-^{13}C-Glutamine] a permis d'identifier la carboxylation réductrice de l'α-KG dans le cycle de Krebs contrairement à la voie métabolique standard du métabolisme de la glutamine qui produit du citrate M+4 via la voie oxydative de l'α-KG (Figure 1.28B) [Yoo et al, 2008 ; Metallo et al,

2011]. Le pourcentage de M+5 et de M+4 sur le profil isotopique du citrate indique que 50 % de la glutamine passe par la voie réductrice et 40 % par la voie oxydative. Cette voie métabolique a été prouvée par la perte du M+5 du citrate lors du knockout de l'Isocitrate Deshydrogénase (IDH) [Zhang et al, 2014].

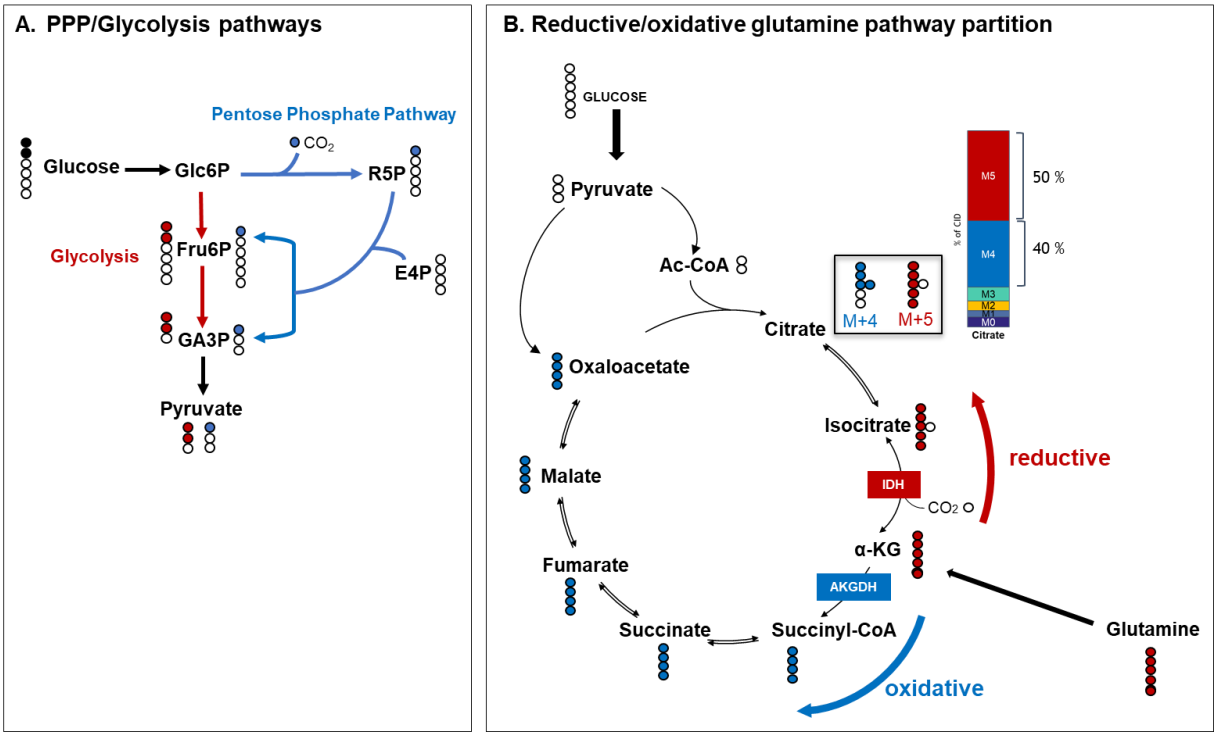


Figure 1.28 : Exemples d'applications de la partition de flux. A. Partition de flux entre la voie de la glycolyse et la voie oxydative des pentoses phosphates. B. Partition de la voie oxydative et réductive de la glutamine

5.2.1.4. Compartimentation

De nombreux métabolites et de nombreuses réactions sont présents dans plusieurs compartiments intracellulaires ce qui ajoute une couche de complexité à l'analyse du métabolisme. La plupart des techniques actuelles ne permettent pas de distinguer la provenance des métabolites concernés et fournissent un profil de marquage moyen d'un métabolite sur l'ensemble des compartiments qui le contient. Cependant selon le métabolite concerné, il est parfois possible de déduire les profils de marquage spécifiques à un compartiment à partir du marquage des métabolites qui sont produits exclusivement dans ce compartiment. Par exemple, le pyruvate se trouve à la fois dans le cytosol et dans les mitochondries et produit directement le lactate et l'alanine. L'enzyme responsable de la production du lactate à partir du pyruvate (lactate déshydrogénase) est une enzyme strictement cytosolique tandis que la production d'alanine est essentiellement liée au transporteur mitochondrial du pyruvate [Adeva et al, 2013 ; Vacanti et al,

2014 ; Buescher et al, 2015]. Ainsi, dans des conditions expérimentales où le milieu de culture n'apporte ni alanine ni lactate, le marquage du lactate reflète probablement celui du pyruvate cytosolique, tandis que le marquage de l'alanine reflète mieux le marquage du pyruvate mitochondrial.

5.2.2. Vers des analyses non-ciblées

Les études isotopiques du métabolisme se sont très longtemps appuyées sur des approches ciblées de mesure de l'incorporation isotopique dans des métabolites connus essentiellement associées au métabolisme central carboné [Zamboni et al, 2009]. Cependant, les approches ciblées ne permettent pas de couvrir l'ensemble du réseau métabolique, ce qui limite leur intérêt dans le cadre d'études systémiques. Elles limitent l'analyse à la partie du réseau accessible à partir de la méthode analytique utilisée. Même si les méthodes ciblées actuelles permettent de mesurer le marquage d'une centaine de métabolites, elles sont loin de couvrir la dimension des réseaux métaboliques cellulaires. Le développement récent des approches non ciblées permet de pallier en partie à ce problème, dans la mesure où elles donnent accès à un nombre bien plus important de métabolites, pouvant aller jusqu'à plusieurs centaines de composés dans une seule analyse. Si la couverture métabolique apportée par ces approches n'est pas encore totale, elles offrent cependant la possibilité d'explorer beaucoup plus largement les réseaux métaboliques. On assiste ainsi actuellement à un développement méthodologique important dans le domaine de l'analyse non ciblée dans le contexte d'expériences de marquage isotopique. Cela concerne aussi bien les aspects analytiques, pour augmenter la couverture du métabolome, que le traitement des données. En effet, comme énoncé précédemment, si l'extraction des données de marquage est mature dans les cas des approches ciblées, ce n'est pas le cas pour les approches non ciblées. Des outils logiciels permettant la détection, et l'extraction automatique des isotopologues sont apparus, mais restent encore à améliorer en termes de fonctionnalité, d'efficacité et d'intuitivité. De même, un aspect important pour la valeur biologique de ces données est la qualité des mesures isotopiques obtenues. En effet l'interprétation métabolique (biologique) des résultats dépend de manière très importante de la précision des mesures et pour cela il devient indispensable de développer des outils de validation adaptés aux méthodes analytiques utilisées pour l'analyse du marquage isotopique, qu'elles soient ciblées ou non. Enfin, l'interprétation des résultats en termes de flux métaboliques – ou de variation de flux métaboliques – nécessite une compréhension

approfondie du métabolisme cellulaire. Elle est actuellement compliquée par le nombre généralement élevé de composés, dont beaucoup restent non identifiés [Weindl et al, 2015].

5.2.3. Fluxomique

Les données isotopiques issues d'expériences de traçage peuvent également être utilisées dans le cadre d'approches de fluxomique. La fluxomique est la mesure de l'ensemble des flux au sein du système métabolique. Le calcul des flux métaboliques est possible grâce au développement de modèles mathématiques permettant de convertir les données de marquage en valeurs de flux métaboliques. Les différentes approches de fluxomique, et plus spécifiquement celles reposant sur le marquage isotopique, sont développées plus loin dans cette introduction. Au préalable, la notion de réseau métaboliques est introduite.

PARTIE II : Réseaux métaboliques : de la reconstruction à la modélisation pour le calcul de flux

6. Les réseaux métaboliques

Comme cela a été introduit au début du chapitre 1, l'analyse des réseaux métaboliques est une composante essentielle de la biologie des systèmes. Elle vise à à établir le réseau, à analyser sa topologie et ses propriétés fonctionnelles (notamment celles qui émergent de l'organisation en réseau), et à le modéliser soit pour expliquer le comportement métabolique observé (modèles explicatifs), soit pour prédire son comportement face à des modifications génétiques ou à de perturbations environnementales (modèles prédictifs).

Dans la suite de ce chapitre, les réseaux métaboliques vont être introduits, ainsi que les différentes approches de reconstruction et de modélisation permettant d'analyser leur topologie et leur fonctionnement.

6.1. Introduction aux réseaux métaboliques

L'unité de base du réseau métabolique est la réaction, qui correspond à la conversion d'un ou plusieurs métabolites d'entrée (substrats de la réaction) en un ou plusieurs métabolites de sortie (produits de la réaction). Ces réactions possèdent différentes propriétés. Entre autres elles peuvent être spontanées ou catalysées par des enzymes et peuvent être réversibles ou irréversibles dans les conditions étudiées. En théorie, toutes les réactions peuvent être réversibles, c'est-à-dire se dérouler dans un sens ou dans le sens inverse. Cependant la réaction inverse peut être limitée lorsqu'elle demande des conditions impossibles (concentration du produit supérieure à sa limite de solubilité), lorsqu'elle présente un coût énergétique très élevé ou encore lorsque le produit est immédiatement consommé.

Une série de réactions qui remplit une fonction de conversion déterminée, par exemple une séquence particulière de réactions au cours desquelles un substrat initial est transformé en un produit spécifique, est appelée une voie métabolique. Ces voies métaboliques permettent de comprendre comment un métabolite donné peut être transformé en un autre métabolite et décrivent les différentes étapes qui prennent part à ce processus. A titre d'exemple, la glycolyse, le cycle de Krebs et la voie des pentoses-phosphates sont des voies régulièrement analysées dans l'étude du métabolisme central et énergétique d'un organisme (Figure 1.29). Les voies

métaboliques sont reliées entre elles et peuvent partager des réactions et des métabolites dits métabolites intermédiaires. Elles forment un maillage de réactions interconnectées et interdépendantes qui constituent le réseau métabolique (Figure 1.29). Ce dernier représente la carte métabolique complète d'un organisme, cataloguant l'ensemble des potentialités de conversion de l'organisme étudié.

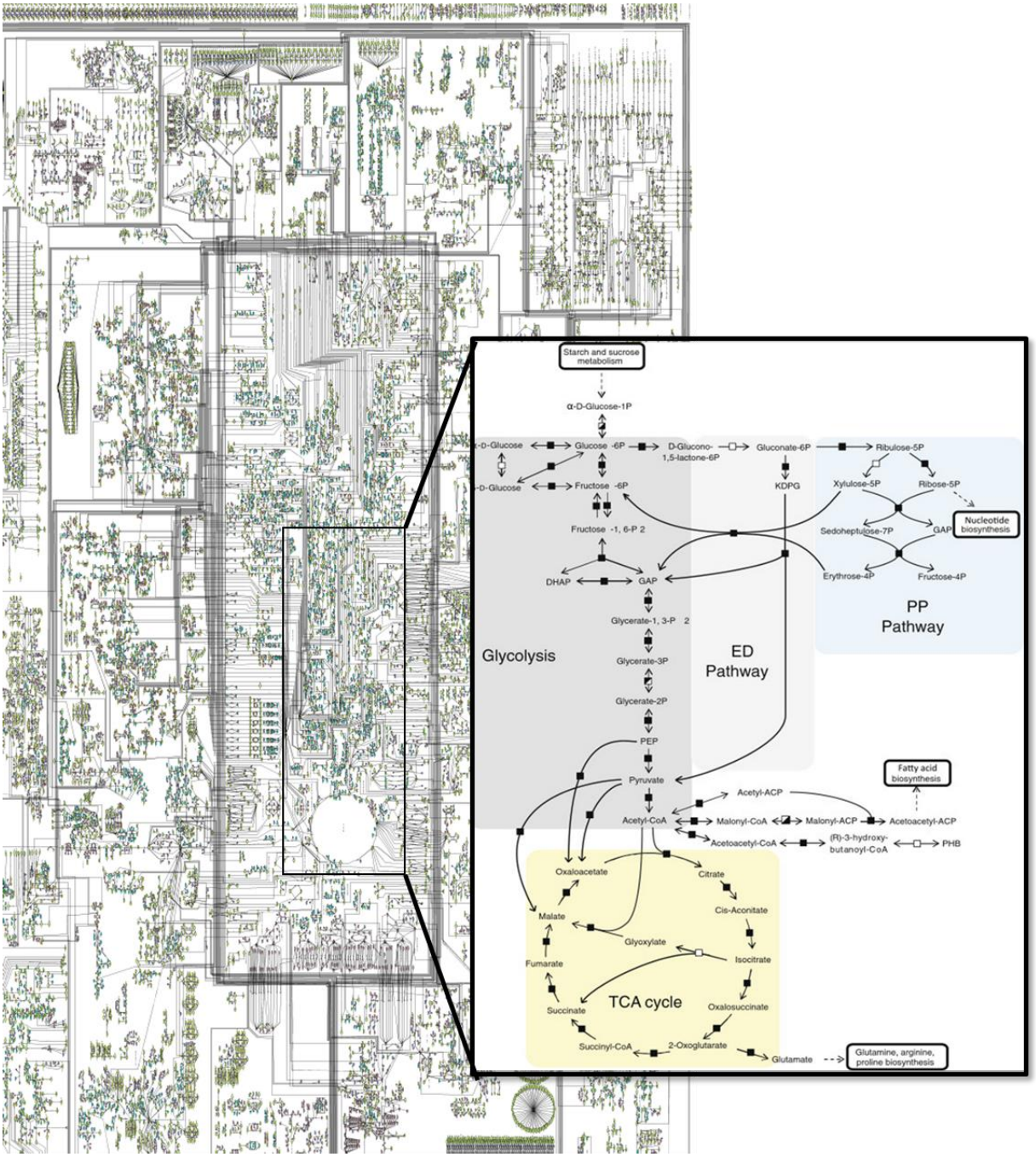


Figure 1.29 : Réseau métabolique et principales voies du métabolisme central et energétique. Réseau métabolique Recon3 chez l'humain, extrait de https://www.vmh.life/#reconmap et principales voies métaboliques : glycolyse, voies des pentoses phosphates et voie Entner-Doudoroff (Illustration adaptée de Tatsukami, 2013).

6.2. Reconstruction des réseaux métaboliques

La reconstruction de réseau métabolique est une étape clé de la biologie des systèmes et consiste à identifier les réactions pouvant se produire dans un organisme donné. L'approche la plus commune pour reconstruire un réseau métabolique est la reconstruction de réseau à l'échelle du génome. Cette reconstruction est basée en particulier sur l'annotation du génome et sur la description des activités enzymatiques de l'organisme considéré. Les réseaux métaboliques à l'échelle du génome (GSMN pour genome-scale metabolic network) représentent des bases de connaissances structurées sur le métabolisme qui rassemblent l'ensemble des réactions métaboliques qui peuvent avoir lieu dans un organisme particulier. Le développement de ces réseaux GSMN et l'émergence de nouvelles méthodes d'analyse *in silico* (analyse topologique, modélisation métabolique) ont élargi les capacités d'étude du métabolisme. Les réseaux GSMN de divers organismes sont également largement utilisés dans diverses applications industrielles et médicales [Durot et al, 2009 ; Kim et al, 2017 ; Gu et al, 2019].

6.2.1. Reconstruction d'un réseau métabolique à l'échelle du génome

La plupart des réseaux métaboliques sont reconstruits à partir de la séquence complète de leur génome. Thiele & Palsson présentent une méthode globale de reconstruction de réseau métabolique en essayant d'homogénéiser différents protocoles [Thiele and Palsson, 2010].

Le processus de reconstruction métabolique d'un organisme repose en premier lieu sur la reconstruction automatique d'une ébauche de réseau (Figure 1.30). Cette ébauche est basée sur l'identification et l'annotation fonctionnelle des gènes codants pour les enzymes métaboliques présentes dans le génome de l'organisme. L'ensemble de ces données permet de déterminer quelles enzymes peuvent être présentes dans l'organisme, et ainsi quelles réactions peuvent s'y dérouler. L'annotation de ces gènes est généralement obtenue par homologie de séquence avec celle de gènes déjà annotés dans un organisme similaire. La première ébauche de réseau est reconstruite par association de toutes les réactions catalysées par les enzymes ainsi identifiées. Des outils automatiques ont été développés pour faciliter ce processus de reconstruction en faisant correspondre les gènes du génome aux réactions associées aux enzymes identifiées (PathwayTools [Karp et al, 2021], GEMSystem [Arakawa et al, 2006]…). Cependant de nombreuses erreurs peuvent se produire lors de l'élaboration de l'ébauche du réseau métabolique. Il est souvent considéré comme (très) incomplet et présente des lacunes

liées à une annotation incomplète ou à des erreurs d'annotations [Monk et al, 2013]. D'autre part l'annotation par homologie de séquence reste basée sur de fortes hypothèses car on suppose que deux enzymes homologues issues de deux organismes différents catalysent les mêmes réactions, ce qui n'est pas forcément le cas. Les méthodes d'annotations ne prennent pas en compte les gènes peu ou pas homologues et qui n'ont pas été identifiés, ou les réactions spontanées qui ne sont associées à aucun gène. Ces erreurs aboutissent soit à l'inclusion dans le réseau GSMN de réactions qui n'existent pas dans l'organisme étudié ou, à l'inverse, l'absence de réactions présentes chez celui-ci. La propagation des annotations entre génomes peut conduire à la propagation des erreurs d'annotations, cumulées à chaque transfert.

L'ébauche du réseau métabolique doit ensuite être vérifiée et complétée. La seconde étape du processus de reconstruction consiste donc à affiner le modèle pour ajouter ou corriger les informations que les procédures automatiques ont manqué dans la reconstruction initiale du réseau. Cette étape de correction est principalement effectuée de manière manuelle sur la base des connaissances biochimiques acquises sur l'organisme ou sur des organismes proches via des bases de données génétiques ou biochimiques ou à partir de la littérature. Alors que l'étape de reconstruction automatique est rapide, le processus de correction manuelle demande beaucoup de travail et est parfois laborieux. Certaines méthodes automatiques visant à assister ce processus sont également utilisées pour affiner les réseaux en complétant les réactions manquantes (GapFind/GapFill [Satish Kumar et al, 2007], algorithme SMILEY [Reed et al, 2006]…)

Le réseau reconstruit à l'échelle du génome (GSMN) fournit une bonne représentation du métabolisme et rassemble les réactions connues pour avoir lieu dans un organisme. Ce réseau peut être converti en modèle mathématique (GSM - Genome Scale Model) et analysé via différentes approches (voir ci-dessous) pour analyser la topologie du réseau, répondre à diverses questions sur les capacités fonctionnelles des organismes et expliquer des observations expérimentales. Les incohérences déterminées entre les prédictions du modèle et les observations expérimentales vont conduire à un nouveau cycle de raffinement manuel. Ce processus itératif de raffinement se poursuit jusqu'à ce que le réseau atteigne la performance désirée déterminée par le constructeur du modèle.

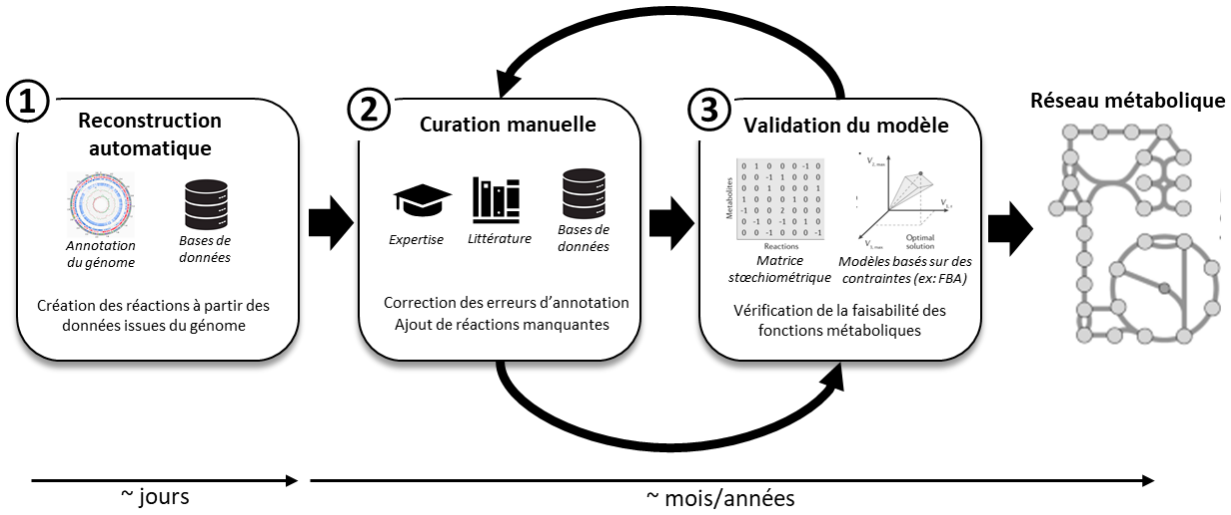


Figure 1.30 : Processus de reconstruction de réseau métabolique à l'échelle du génome

La reconstruction telle que décrite ci-dessus demande beaucoup de temps, de 6 mois à plusieurs années dans le cas de reconstruction de métabolismes complexes (ex : réseau métabolique de l'humain) [Duarte et al, 2007]. De plus, la valeur biologique apportée par ces réseaux est limitée par de multiples sources d'incertitudes. Les choix opérés à chaque étape de la reconstruction peuvent conduire à des réseaux reconstruits avec des structures différentes [Bernstein et al, 2021]. Cette approche reste limitée par notre connaissance imparfaite du métabolisme. Malgré une expérience et des connaissances croissantes, ainsi que l'amélioration des outils de reconstruction [Mendoza et al, 2019], nous ne sommes toujours pas en mesure, à ce jour, de reconstruire de manière entièrement automatique des réseaux métaboliques complets et de haute qualité.

La dimension d'un réseau peut varier selon l'organisme, de 500 à plus d'un millier de réactions [King et al, 2016]. Cette variété est liée aux différents degrés de complexité des organismes (organismes compartimentés, …). Des plateformes telles que BiGG [King et al, 2016], KEGG [Kanehisa et al, 2000], BioCyc [Karp et al, 2019] recensent les réseaux métaboliques de différents organismes. On y retrouve notamment le dernier réseau GSMN de l'Homme Recon3D composé de 5835 métabolites, 228 gènes et plus de 10600 réactions (Figure 1.29) [Brunk et al, 2018]. On peut également retrouver plusieurs réseaux par organisme, par exemple la base de données BiGG models propose 20 modèles pour *E. coli* de souches différentes [King et al, 2016; Ye et al, 2022]. Ces réseaux restent en constante évolution. A titre d'exemple, le réseau métabolique d'*E. coli* a vu son contenu évoluer entre le premier modèle iJE660 édité dans les années 2000 qui contenait 627 réactions et 438 métabolites [Edwards & Palsson, 2000] et le modèle iML1515 publiée en 2017 recensant 2719 réactions et 1192

métabolites [Monk et al, 2017]. Cette évolution est liée à une meilleure connaissance et à la disponibilité accrue des données biologiques au cours des dernières années.

L'utilisation d'un language standard de description permet de faciliter l'utilisation et le partage de ces réseaux. On retrouve différents formats pour décrire les réseaux métaboliques, notamment le format SBML (« System Biology Markup Language ») qui est le plus utilisé. Ce format liste l'ensemble des métabolites intervenant dans le réseau ainsi que la liste des réactions avec leur substrats et leurs produits [Hucka et al, 2003].

6.2.2. Stratégies expérimentales de reconstruction des réseaux métaboliques

Dans cette section sont abordées les approches expérimentales qui permettent de reconstruire des réseaux métaboliques. Ces approches peuvent être utilisées soit en complément de la reconstruction *in silico,* pour compléter ou corriger les réseaux GSMN, soit en tant que tel pour construire des réseaux métaboliques sur des bases purement expérimentales. En cohérence avec le niveau systémique de regard sur le métabolisme qui est envisagé dans cette introduction, ne sont abordées dans cette partie que les approches qui permettent d'accéder à une information globale sur le métabolisme. Les approches plus spécifiques (réductionnistes), bien que très utiles pour aborder certains aspects du métabolisme, ne seront pas abordées.

6.2.2.1. Apport des données omiques à la reconstruction des réseaux GSMN

L'émergence des méthodes « -omiques » a permis l'accès à des informations globales (à l'échelle du système) concernant l'expression des gènes, des protéines ou encore l'abondance des métabolites dans des conditions spécifiques. L'intégration de ces données dans le cadre de la reconstruction des réseaux métaboliques s'incrit dans plusieurs objectifs complémentaires.

Le premier objectif est celui de la validation ou de la correction expérimentale des réseaux reconstruits *in silico* à partir des données d'annotation du génome (réseaux GSMN). Les données expérimentales permettent en effet de montrer la réalité d'une annotation, ou de compléter les données d'annotation [Machado and Herrgard, 2014]. Par exemple, les données de métabolomique apportent des informations sur la nature des métabolites présents et sur leur concentrations relatives ou absolues. L'indication sur la nature des métabolites permet d'affiner et valider les réseaux métaboliques. La présence des substrats et produits d'une réaction annotée

permet de valider – ou d'appuyer – la présence réelle de la réaction. La métabolomique est également devenue un outil largement utilisé pour identifier des fonctions potentielles de gènes non annotés ou mal annotés. Il existe deux stratégies pour cela : celle basée sur des associations gène-métabolome et celles qui utilisent des approches de profilage métabolique [Ryu et al, 2015].

Le second objectif est de reconstruire le réseau métabolique contextuel. Il s'agit du réseau métabolique réellement en place dans le système biologique étudié (réseau cellule-spécifique, tissu-spécifique, etc) et dans les conditions physiologiques considérées. Comme cela est mentionné dans la section précédente, les réseaux GSMN sont principalement établis à partir de données d'annotation du génome et représentent l'ensemble des réactions qui peuvent être présentes dans l'organisme. Ils représentent le potentiel métabolique global de l'organisme. Cependant, si on prend l'exemple de l'Homme, toutes les cellules humaines possèdent le même génome (et donc le même potentiel métabolique global), mais l'expression de celui-ci diffère largement d'une cellule à une autre ou, pour la même cellule, d'une condition physiologique à une autre. L'équipement enzymatique en place dans une cellule donnée dépend donc à la fois du type cellulaire concerné mais aussi des conditions environementales (au sens large) dans lesquelles elle se trouve. Le même raisonnement s'applique à l'échelle tissulaire. En ce sens, le métabolisme est le produit de l'interaction entre le génome et l'environnement. Le réseau contextuel est donc un sous-ensemble du réseau global (un sous-réseau), qui représente le réseau exprimé dans le contexte étudié par rapport au réseau potentiel. Il peut être reconstruit à partir du réseau GSMN par suppression des réactions inactives, une opération qui peut être effectuée par différents algorithmes à partir par exemple des niveaux d'expression des gènes, de la présence de protéines ou de métabolites, de la disponibilité des données expérimentales et de la connaissance de la littérature. En restant dans le cas du réseau métabolique humain, l'exploitation des données « -omiques » a conduit à la construction de plusieurs modèles métaboliques spécifiques (Figure 1.31) [Ryu et al, 2015].

Enfin, les données omiques peuvent être aussi exploitées dans le cadre non plus de la reconstruction des réseaux, mais dans celui de la modélisation du métabolisme. Elles apportent des contraintes expérimentales qui permettent d'affiner les modèles et d'optimiser leur exploitation. Cela sera détaillé dans le chapitre 7.

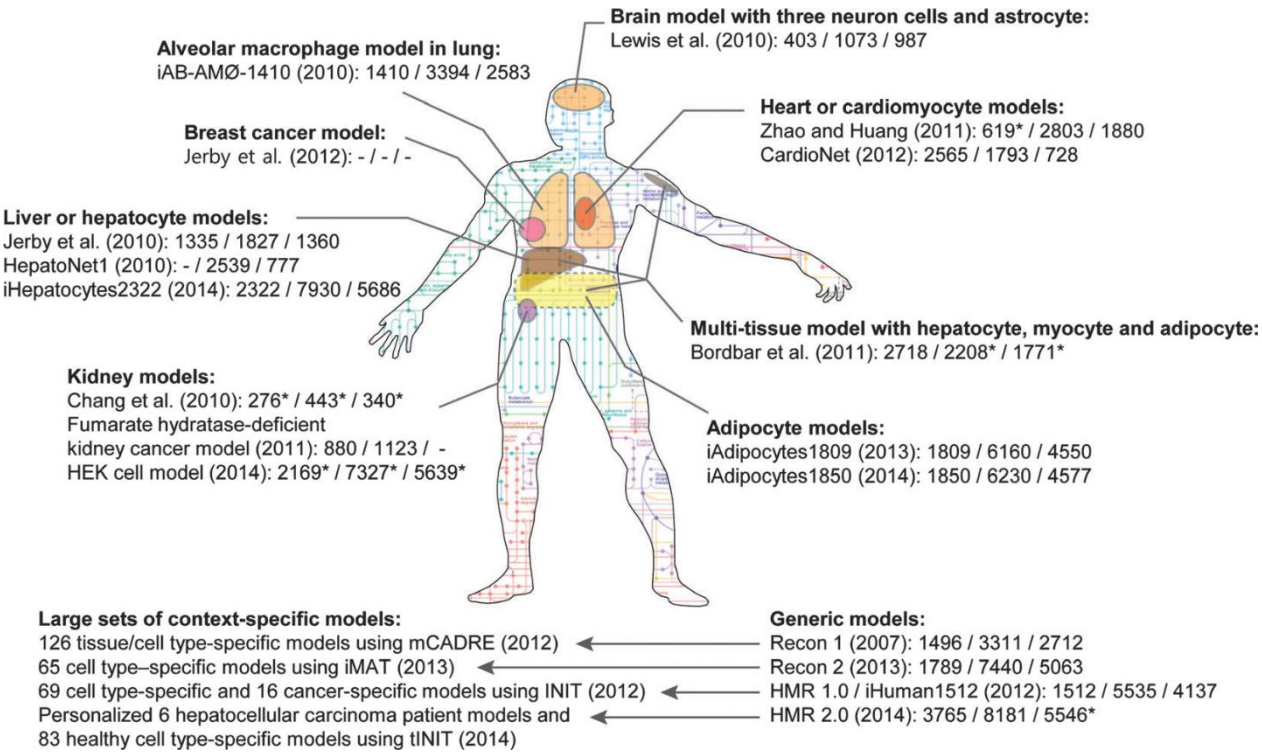


Figure 1.31 : État des réseaux métaboliques humains génériques et contextuels rapportés en 2015. Pour ces modèles, les nombres de gauche, du milieu et de droite indiquent le nombre de gènes, de réactions et de métabolites, respectivement. Les nombres marqués d'un astérisque ont été obtenus par traitement direct des fichiers SBML respectifs, car les statistiques du réseau n'étaient pas évidentes dans la littérature d'origine. Illustration tirée de Ryu et al, 2015.

6.2.2.2. Approches pilotées par les données (data-driven approaches)

Le réseau métabolique contextuel peut aussi être reconstruit directement à partir des données elles-mêmes (reconstruction pilotée par les données), sans appui sur le réseau GSMN. La puissance actuelle des techniques analytiques permet en effet la mise en place de méthodes de reconstruction métabolique alternatives permettant de reconstruire des réseaux métaboliques à partir de données expérimentales. Contrairement aux réseaux GSMN, ces approches de reconstruction pilotées par les données sont basées sur l'analyse directe du métabolome. L'objectif de ces approches est d'identifier le plus grand nombre possible de métabolites, qui constituent les éléments du réseau, et d'établir les liens de conversion (réactions) qui peuvent exister entre eux.

Les réseaux expérimentaux sont principalement dérivés de données métabolomiques obtenues via des techniques analytiques haute résolution. Bien que la spectroscopie RMN soit très utilisée en métabolomique, les réseaux expérimentaux présentés ci-dessous sont basés sur des données obtenues par spectrométrie de masse haute-résolution (HRMS) via des analyses

non-ciblées. L'utilisation de la spectrométrie de masse présente plusieurs avantages pour ce type d'approche. Tout d'abord leur haute sensibilité permet de détecter des composés présents en très faible quantité. De plus l'utilisation d'approches non-ciblées permettent de couvrir une grande partie du métabolome. Enfin, les méthodes de spectrométrie de masse en tandem sont un atout pour l'identification de métabolites inconnus.

Sur cette base d'analyses du métabolome par HRMS, il existe plusieurs approches de reconstruction de réseaux. On distingue principalement les réseaux basés sur les différences de masse, les réseau d'adduits, les réseaux de corrélation et les réseaux de similarités spectrales (Figure 1.32) [Amara et al, 2022]. Chacun de ces réseaux tend à établir un lien potentiel entre des métabolites pouvant être respectivement substrats et produits d'une réaction.

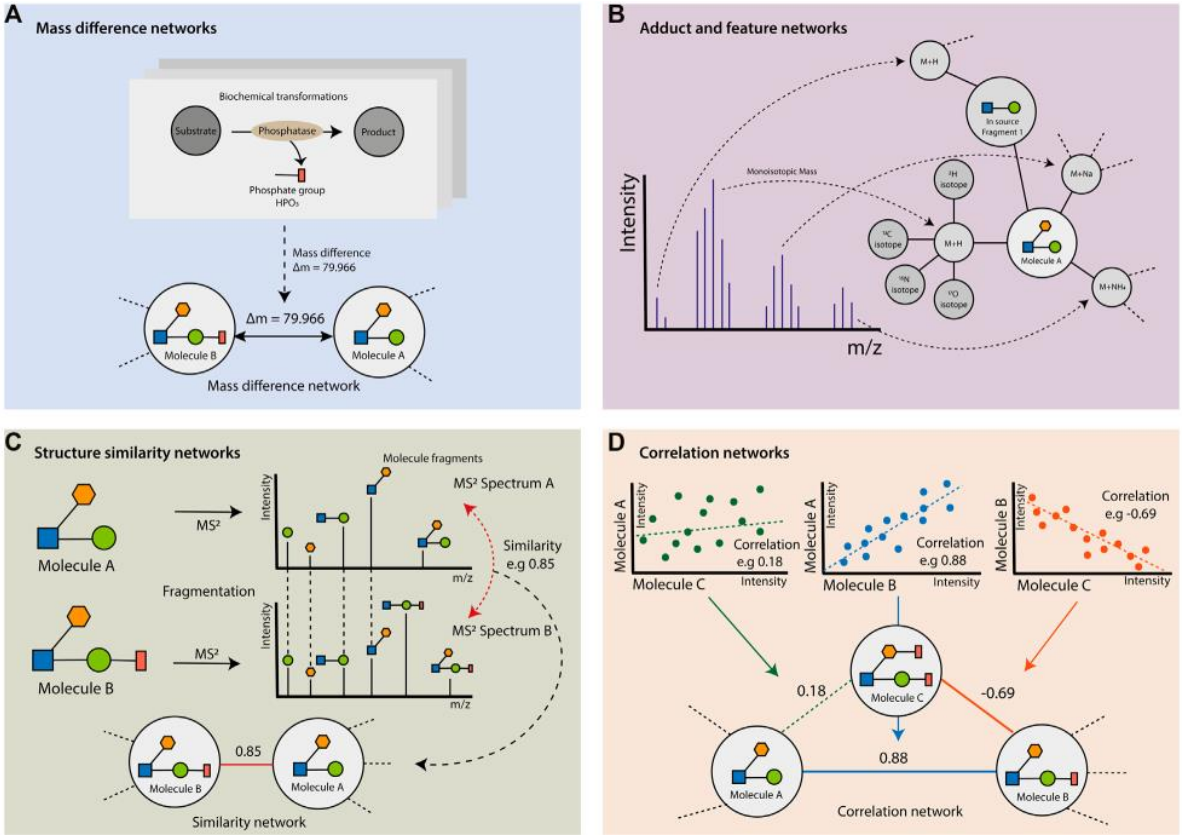


Figure 1.32 : Méthodes de reconstruction expérimentales. Illustration tirée de Amara et al, 2022.

6.2.2.2.1. Réseaux basés sur les différences de masses

La reconstruction de réseaux basée sur les différences de masse (MDiNs pour « mass difference networks ») consiste à extraire toutes les différences de masses existantes entre deux paires de pics de masses pour suggérer la présence d'une réaction biochimique (Figure 1.33). Les MDiNs sont généralement générés à partir de données de spectrométrie de masse haute-

résolution acquises via des analyses non-ciblées. Les MDiNs utilisent les valeurs de masse en tant que nœuds du réseau. Les différences de masse définies entre ces différentes valeurs sont quant à elles utilisées pour établir les arêtes du réseau. Cette approche permet d'obtenir un réseau au sein duquel tous les « potentiels » métabolites détectés au cours de l'analyse sont connectés les uns aux autres sur la base de leur différence de masse. Ce réseau peut être très difficile à exploiter car tous les métabolites détectés sont interconnectés et dans cette représentation, deux métabolites qui ne seraient pas liés biochimiquement seraient tout de même reliés. Une solution pour réduire les liens non pertinents entre les métabolites est de filtrer les liens ayant une faible corrélation. D'autres filtres, par exemple basés sur des variations prévisibles de temps de rétentions sont également exploitables pour améliorer la précision du réseau reconstruit [Amara et al, 2022]. Il existe plusieurs outils permettant de générer automatiquement des réseaux de différences de masse, tels que l'outil mzGoupAnalyzer [Doerfler et al, 2014], MetaNetter [Burgess et al, 2018] ou MetNet [Naake & Fernie, 2019].

Une autre approche consiste à exploiter les variations de masse, c'est-à-dire les pertes ou les gains de masse qui peuvent correspondre à des pertes/gains d'atomes, existantes entre deux paires de pics de masse et à les associer à une transformation biochimique existante. Cette approche a été initiée par les travaux de Breitling et al. Elle s'appuie sur le fait qu'un répertoire limité de 83 transformations chimiques, compilées à partir de manuels de biochimie, représente la majorité des réactions biochimiques opérant au sein des cellules [Breitling et al, 2006]. Cette liste de transformation est ensuite associée à la liste de variations de masse. Dans l'exemple présenté ci-dessous (Figure 1.33), une différence de masse de 14.01565 Da peut correspondre à une réaction de méthylation (variation de CH2 après substitution de l'atome d'hydrogène -H, par un groupe -CH3) entre le composé B et le composé C. Cette approche fournit un réseau ou sous-réseau contextuel. Cependant elle présente certaines limites. Notamment une différence de masse entre une paire de composés ne correspond pas systématiquement à une transformation biologique réelle, et en se limitant à ces données, il est difficile d'assurer la réalité de la réaction dans l'organisme.

L'avantage de ces deux approches est que, comparés au GSMN, les réseaux obtenus peuvent prendre en compte des réactions non enzymatiques spontanées. Les MDiN sont l'un des principaux types de réseaux expérimentaux utilisés pour l'analyse des données métabolomiques [Amara et al, 2022 ; Traquete et al, 2022]. Ils ont été utilisés dans des études sur des échantillons de levure [Liu et al., 2016], de tissu adipeux de souris et de plasma humain [Laber et al., 2021].

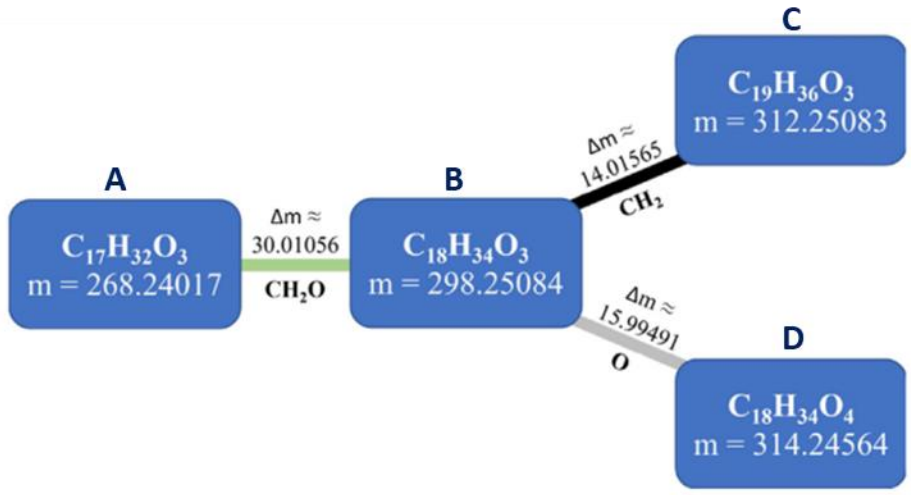


Figure 1.33 : Schématisation du concept des réseaux basés sur les différences de masse. Illustration tirée de Traquete et al, 2022

6.2.2.2.2. Les réseaux d'adduits

Les différences de masse ne sont pas seulement dues à des transformations biochimiques entre différents métabolites, mais peuvent également apparaître en raison d'effets physicochimiques différents lors de l'introduction des métabolites dans le spectromètre de masse. Ces différences de masse peuvent être représentées par des réseaux d'adduits. Les réseaux d'adduits mettent en relation des features correspondant à des adduits de métabolites, des isotopologues de ces adduits ou des ions fragments résultant de fragmentation en source. La fragmentation en source est un phénomène courant qui se produit lors de l'ionisation éléctrospray (ESI). Au cours de l'ionisation, les molécules acquièrent une énergie interne supplémentaire qui est libérée, ce qui entraîne la fragmentation de la molécule en source. Cette fragmentation génère des ions précurseurs supplémentaires qui peuvent conduire à des annotations faussement positives des ions moléculaires [Gathungu et al., 2018]. Les fragments de source détectés, les adduits et leurs isotopologues associés peuvent être représentés comme des nœuds d'un réseau, avec des liens les reliant à leurs métabolites associés (Figure 1.32B). Les réseaux d'adduits peuvent être utilisés pour améliorer la précision et la confiance dans l'annotation des métabolites dans les réseaux.

6.2.2.2.3. Réseaux de similarité spectrale

La reconstruction de réseaux de similarité spectrale effectue des liens substrats-produits sur la base de la similarité chimique des métabolites détectés [Amara et al, 2022]. Cette approche repose sur l'hypothèse que les métabolites connectés par des réactions biochimiques sont chimiquement similaires, c'est-à-dire qu'ils partagent des structures communes. Elle est habituellement associée à des analyses de spectrométrie de masse en tandem car ces composés partagent généralement des schémas de fragmentation communs. En métabolomique non-ciblée, elle est généralement effectuée via un mode d'acquisition dépendant des données (mode DDA - Data-Dependent Acquisition) qui fragmente les composés les plus abondants. Les différences et similarités spectrales peuvent être établies sur la base des masses des fragments ou des pertes de neutre entre les différentes masses. Pour établir des liens entre les composés, différents algorithmes ont été développés basée sur différentes métriques telles que le cosinus ou le cosinus modifié [Aguilar-Mogas et al., 2017]. La première application de réseau de similarité a été proposée par Watrous et al. basée sur un score cosinus modifié, qui prend en compte la différence de masse entre les masses des précurseurs. Ces différences de masse ont été appliquées aux fragments dans les spectres MS/MS, ce qui a conduit à une correspondance entre les pics des fragments [Watrous et al, 2012]. Les similarités entre deux spectres MS/MS peuvent également être calculées via l'index de Tanimoto [Bajusz et al, 2015].

Cette approche présente deux limites. D'une part, elle peut générer de faux positifs car de nombreux composés présentant une similitude structurelle ne sont pas forcément impliqués dans les mêmes réactions ou voies métaboliques. D'autre part, elle est dépendante des méthodes analytiques utilisées. En effet, en fonction des méthodes et instruments utilisés, tous les composés ne seront pas forcément fragmentés. De plus, elle nécessite d'avoir identifié en amont les métabolites concernés pour avoir accès à leur structure chimique.

6.2.2.2.4. Réseaux de corrélation

Les réseaux de corrélation sont reconstruits directement via le calcul des corrélations entre métabolites sur la base des abondances mesurées dans les ensembles de données métabolomiques. Les métabolites qui sont connectés par des réactions biochimiques présentent souvent une codépendance qui s'observe sur la base de leurs concentrations [Rosato et al, 2018]. Les corrélations entre ces métabolites sont calculées par comparaison par paire de l'intensité des pics de masses. Deux métabolites sont liés si la valeur de corrélation atteint un niveau

considéré comme significatif. Le plus souvent, le calcul des corrélations s'effectue par le coefficient de corrélation de Pearson. Un des premiers réseaux de corrélation a été établi par Ursem et al. à l'aide de données métabolique recueillies sur un ensemble varié de génotypes de tomates par GC/MS [Ursem et al, 2008]. Cependant les réseaux obtenus sont généralement denses et complexes car ils prennent en compte des corrélations indirectes. En effet deux métabolites qui présentent une corrélation forte avec un autre métabolite seront également reliés entre eux. Une autre méthode consiste à utiliser une corrélation partielle basée sur des modèles graphiques gaussiens qui génère des réseaux plus robustes et plus stables [Krumsiek et al, 2011 ; Benedetti et al, 2020].

6.2.2.3. Avantages et limites de ces approches

Ces différentes approches sont les premières méthodes qui proposent une reconstruction de réseau métabolique basée sur les données expérimentales. Un grand avantage d'une reconstruction de réseau métabolique basée sur la métabolomique est qu'elle détecte les réactions non enzymatiques et qu'elle ne dépend pas de la disponibilité de l'annotation du génome. De plus, ces approches permettent d'accéder directement à un sous-réseau spécifique au contexte pour une cellule/tissu donné.

Cependant, les approches décrites ici dépendent de la précision et de la sensibilité des techniques analytiques employées. Certaines de ces masses peuvent également provenir d'artefacts et, en l'absence de procédure de filtration des données, peut induire des informations non pertinentes dans le process de reconstruction du réseau. Elles peuvent également générer de nombreux faux positifs, notamment en ce qui concerne les liens entre les métabolites. Par exemple, l'absence de certains intermédiaires métaboliques peut conduire à l'omission de certaines voies. De plus, l'utilisation de techniques à haute résolution ne garantit pas l'identification précise d'un métabolite et une masse peut en définitive correspondre à plusieurs, voire un très grand nombre de métabolites [Kind & Fiehn, 2007]. Cet ensemble de variables peut ajouter de l'imprécision dans les réseaux métaboliques produits à partir des données expérimentales. La combinaison de données génomiques et métabolomiques pour un même organisme peut améliorer considérablement la qualité des réseaux métaboliques reconstruits. Les réseaux reconstruits à l'échelle du génome et les réseaux expérimentaux sont liés. Les premiers peuvent fournir une aide pour aider à interpréter et analyser les réseaux expérimentaux, tandis que les seconds peuvent être utilisés pour affiner le GSM en identifiant des métabolites potentiellement manquants dans le réseau.

6.2.3. Stratégie alternative de reconstruction métabolique ab initio

La reconstruction de réseaux métaboliques est une démarche puissante pour analyser le fonctionnement du métabolisme à l'échelle du réseau. Les approches de reconstruction présentées ci-dessus (GSMN et plus récemment les modèles de reconstructions expérimentales) ont montré leurs avantages pour l'analyse du métabolisme et l'étude de la relation topologie-fonction au sein du réseau.

Cependant, les réseaux obtenus restent relativement incomplets et présentent des lacunes : métabolites ou réactions manquantes, manque d'information sur la réalité de la réaction, manque d'identification des métabolites, compléments d'approches compliqués et chronophages pour accéder au réseau contextuel complet (c'est-à-dire au réseau actif dans un type cellulaire spécifique ou dans un contexte donné). De plus ces approches ne nous permettent pas d'obtenir le réseau métabolique d'organismes complexes, méconnus ou totalement inconnus pour lesquels l'accès à l'annotation du génome n'est pas garanti. Une stratégie alternative consiste à reconstruire un réseau métabolique *ab initio* à partir d'expériences de marquage isotopique non-ciblées pour donner accès directement au réseau actif dans une condition donnée.

Un des objectifs principaux de cette thèse est de développer une approche de reconstruction métabolique *ab initio*, basée sur des stratégies d'analyses non-ciblées de marquage isotopique dynamique par couplage chromatographie liquide – spectrométrie de masse haute-résolution (LC-HRMS). La mise en place de cette approche a nécessité le développement d'outils d'intégration de données et de modélisation métabolique novateurs. Son intérêt principal est qu'elle inclue une dimension isotopique qui est un avantage non négligeable pour s'assurer de la réalité des réactions biochimiques identifiées entre différents métabolites. Comme il a été démontré au début de ce chapitre, l'utilisation d'isotopes a également plusieurs bénéfices qui peuvent être exploités dans cette approche basée sur les données expérimentales, tels que l'utilisation de standards isotopiquement marqués pour s'assurer de la qualité des données ou en tant que support pour l'identification de métabolites.

Le principe de cette approche sera détaillé dans le chapitre 3 de cette thèse.

6.3. Visualisation des réseaux métaboliques

Les réseaux reconstruits via les différentes approches énoncées ci-dessus sont des structures complexes comprenant des milliers de réactions et de métabolites. La visualisation d'un réseau est une méthode essentielle pour pouvoir les exploiter, les analyser, générer des hypothèses testables sur des données biologiques et en extraire des connaissances sur leur fonctions biologiques. Le défi consiste à créer des représentations visuelles d'une partie ou de l'ensemble des réactions et des métabolites constituant les réseaux métaboliques. Les outils de visualisation des réseaux peuvent avoir plusieurs fonctions tels que représenter un réseau métabolique à l'échelle du génome, intégrer et cartographier des données omiques et analyser les réseaux via des modes de calculs automatiques intégrés dans l'outil. Il existe un très grand nombre de logiciels ou d'applications web permettant de représenter et analyser les réseaux métaboliques. Nous évoquons dans cette section quelques uns des outils existants et leur fonctionnalité.

Un des outils les plus utilisés pour réaliser ces fonctions est le logiciel open-source Cytoscape [Shannon et al, 2003]. Il permet de visualiser des réseaux complexes et d'intégrer n'importe quel type de données. Cytoscape s'est imposé dans la communauté scientifique car il fournit une API (interface de programmation d'applications) bien documentée, qu'il est libre et qu'il propose un nombre croissant d'applications qui visent à traiter les données. Cytoscape offre aussi la possibilité d'interaction directe avec des bases de données.

Escher [King et al, 2015] et PathwayTools [Karp et al, 2002] sont des applications populaires conçues pour construire et partager des visualisations de voies métaboliques sur plusieurs organismes modèles. MetExploreViz [Chazalviel et al, 2018] permet d'importer des données et de les cartographier, et a pour objectif la visualisation de données omiques sur les réactions et les métabolites. Plus récemment, Fluxer [Hari and Lobo, 2020] a été conçu pour calculer et visualiser les réseaux de flux métaboliques à l'échelle du génome.

Globalement, la visualisation des réseaux métaboliques accélère la compréhension et l'analyse scientifiques et permet d'effectuer de nombreuses opérations d'analyse basées sur les réseaux, telles que la comparaison des réseaux d'un ou de plusieurs organismes et l'interprétation de plusieurs ensembles de données omiques.

6.4. Analyse topologique des réseaux métaboliques

L'analyse de la topologie des réseaux métaboliques a pour objet l'analyse des connexions entre les différents éléments du réseau. L'organisation topologique d'un réseau métabolique peut être étudiée à différentes échelles : à l'échelle des métabolites, des réactions, des voies métaboliques ou du réseau métabolique dans son ensemble. Cette dernière échelle est la plus représentative du fonctionnement du métabolisme d'un organisme car elle représente l'ensemble du potentiel métabolique de l'organisme étudié. L'une des principales motivations pour étudier la topologie des réseaux est que les fonctions métaboliques, et leurs dysfonctionnements éventuels, sont déterminés en partie par la structure du réseau [Wunderlich & Mirny, 2006].

6.4.1. Les graphes métaboliques

Une première manière d'étudier la topologie des réseaux métaboliques (et d'en simplifier la représentation) consiste à les modéliser sous forme de graphes. Un graphe est un objet mathématique qui décrit les connexions entre différents éléments. Il peut être représenté comme un ensemble de nœuds connectés par des liens. Les nœuds représentent les éléments du système, et les liens représentent la relation entre les éléments considérés du réseau et peuvent être unidirectionnels ou bidirectionnels. La représentation du réseau métabolique sous forme de graphes nécessite d'identifier en amont les entités biologiques associées aux nœuds et le sens des connexions existants entre eux. Ce choix dépend de la question que l'on souhaite aborder à l'aide de ce modèle de graphe. On peut trouver des graphes de réactions ou d'enzymes mais le plus souvent, les nœuds sont utilisés pour représenter les métabolites et sont reliés entre eux par les réactions qui convertissent l'un en l'autre [Lacroix et al, 2008]. On distingue les graphes dits bipartes pour lesquels il existe plusieurs types de nœuds (métabolites, réactions) et les hypergraphes pour lesquels les liens relient l'ensemble des substrats et l'ensemble des produits (Figure 1.34).

La méthode des graphes est utilisée dans les bases de données telles que KEGG [Kanehisa et al, 2000] et EcoCyc [Karp et al, 2018]. Dans ces bases de données, les voies métaboliques sont représentées explicitement avec des nœuds symbolisant les métabolites, et les arêtes représentant les réactions. La méthode des graphes permet d'effectuer rapidement des analyses sur des réseaux de grande dimension.

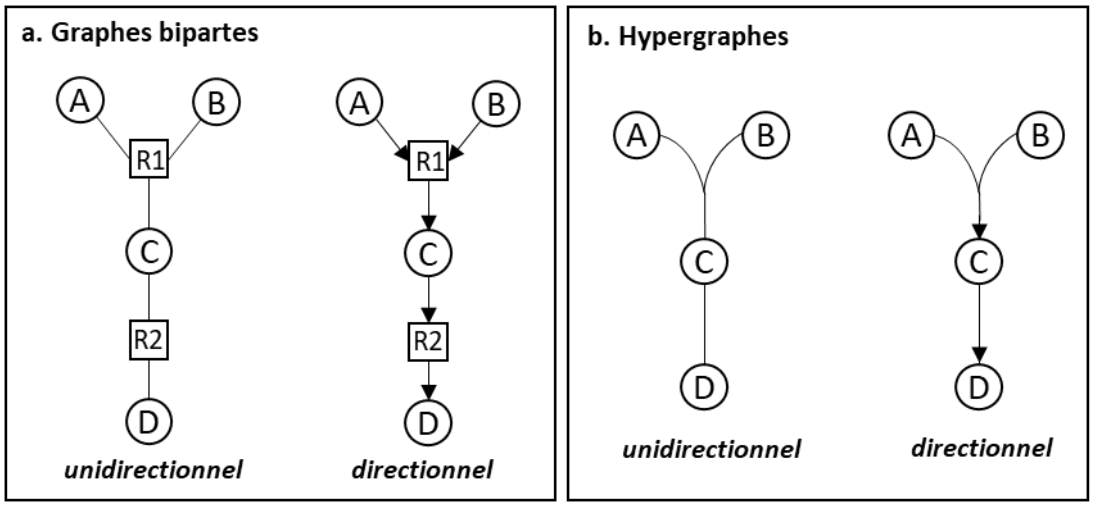


Figure 1.34 : Les principaux types de graphes représentant les réseaux métaboliques. a.Graphe biparte : les nœuds représentent des métabolites (cercle) et des réactions (carré). Un métabolite ne peut être relié qu'à des réactions et inversement. b.Hypergraphes : les nœuds représentent des métabolites et les réactions sont représentées par des arêtes. Une arête peut relier plus de deux nœuds. On distingue les hypergraphes unidirectionnels qui ne prennent pas en compte le sens des réactions et les graphes directionnels. Illustration adaptée de Lacroix et al, 2008

6.4.2. Modes élémentaires

Une autre approche d'analyse topologique des réseaux métaboliques a été développée par Schuster et al en 1999 [Schuster et al, 1999]. Cette approche vise à identifier toutes les capacités de conversion (tous les chemins métaboliques) stables qu'offre un réseau métabolique. Elle introduit la notion de « mode de flux » (flux mode) pour décrire tout processus métabolique qui peut être stable dans le temps (flux constant à travers le processus), c'est à dire fonctionner à l'état stationnaire. Un mode de flux est dit « élémentaire » s'il est non décomposable en modes de flux plus petits (Figure 1.35). Exprimé de manière moins formelle, cette approche permet d'identifier tous les chemins métaboliques stables pouvant exister au sein d'un réseau métabolique donné. Dit autrement, elle permet d'identifier toutes les capacités fonctionnelles de ce réseau, et fait ainsi le lien entre la topologie du réseau et ses fonctions. Il est ainsi possible d'identifier, par exemple, tous les chemins métaboliques d'un réseau permettant de former de la biomasse, ou encore tous les chemins ne conduisant qu'à la production d'ATP. Ces différents chemins peuvent ensuite être classés en fonction de critères d'intérêt. Par exemple, les chemins conduisant à la biomasse peuvent être classés en fonction de leur rendement énergétique, ou de leur rendement en carbone, etc. Cette approche, ainsi que d'autres approches complémentaires (minimal cut sets, extreme pathways, etc), est très utilisée à la fois pour identifier le potentiel métabolique complet d'un organisme (pouvant conduire à identifier de nouvelles voies métaboliques), pour en étudier la mise en oeuvre dans des

condtions physiologiques d'intérêt, mais aussi comme outil de prédiction pour optimiser ou modifier le métabolisme en vue d'applications biotechnologiques ou autres.

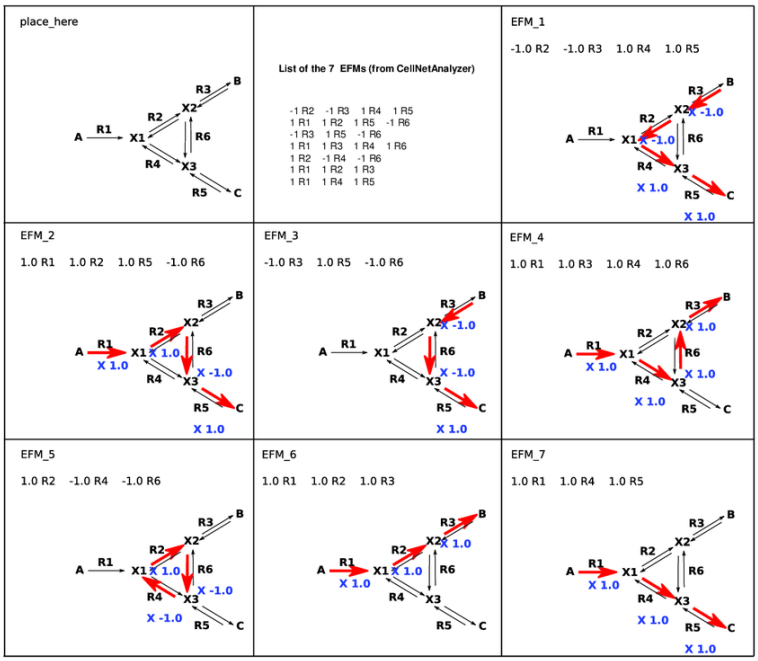


Figure 1.35 : Exemple de décomposition d'un réseau en modes élémentaires de flux. Du fait de sa topologie, le réseau de 6 réactions montré dans le carré supérieur gauche peut être décomposé en 7 modes élémentaires de flux représentés dans les différents carrés. Illustration tirée de Rose et al, 2018.

6.5.Modélisation du métabolisme

6.5.1.Du réseau métabolique au modèle

En raison de la dimension et de la complexité des systèmes métaboliques, l'analyse des réseaux métaboliques s'appuie sur la construction de modèles. La définition de modèle varie très largement suivant les domaines concernés ou des objectifs visés par la modélisation. Ici, de manière simple, un modèle métabolique est défini comme une représentation formelle (mathématique) d'un réseau métabolique. En ce sens, les graphes métaboliques ou les modes élémentaires développés dans la section 6.4 sont des modèles métaboliques qui ont pour objectif d'analyser la topologie des réseaux et le lien entre celle-ci et les propriétés fonctionnelles du réseau. Le modèle est un objet mathématique qu'il est possible de manipuler pour effectuer des simulations et des prédictions [Borodina and Nielsen, 2005]. Ils présentent une valeur pour tester la validité d'un réseau métabolique reconstruit et y apporter des corrections éventuelles. Ces modèles sont également très largement utilisés comme des outils pour analyser, tester et prédire le fonctionnement du métabolisme. Une distinction est souvent faite entre modèles explicatifs, visant à expliquer des comportements métaboliques observés, et modèles prédictifs

visant soit à prédire la réponse du système face à une perturbation (interne ou externe) soit à déterminer comment agir sur le réseau pour atteindre un objectif spécifique (améliorer un procédé biotechnologique, évaluer la valeur d'une cible thérapeutique, etc). Ils peuvent être développé à l'échelle du réseau complet de l'organisme, ou à l'échelle de sous-réseaux de façon à être focalisés sur les processus métaboliques d'intérêt pour la question biologique considérée [Machado et al, 2011]. S'il existe plusieurs possibilités de décrire le fonctionnement du métabolisme, la notion de flux métabolique est celle qui rend compte le plus précisément de l'activité du réseau, que ce soit en précisant quelles voies métaboliques sont réellement actives, mais aussi en indiquant leur niveau d'activité de manière quantitative.

Dans la suite du manuscrit, le terme de modèle métabolique sera utilisé uniquement pour parler de modèle visant à calculer ou prédire les flux métaboliques au sein du réseau. Avant de voir les principaux types de modélisation métabolique, la notion de flux est introduite.

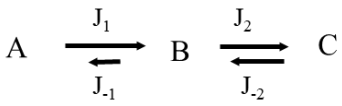
6.5.2. Notion de flux métabolique

Les flux métaboliques sont les vitesses réelles des réactions métaboliques, et la mesure des flux est donc la mesure de la dynamique des systèmes métaboliques. Ils sont exprimés en quantité de molécules transformées par unité de temps et normalisés au nombre de cellules ou à la quantité de protéines. La fluxomique est la mesure de l'ensemble des flux au sein du système métabolique. Les flux varient suivant les conditions physiologiques dans lesquelles se trouve l'organisme étudié et représentent le phénotype métabolique ultime puisqu'ils intégrent à la fois la topologie du réseau qui est opérationnel dans les conditions étudiées et son activité. Il n'existe cependant pas de méthode directe de mesure des flux métaboliques internes à une cellule, à un tissu ou à un organisme. Les méthodes actuelles s'appuient toutes sur des approches de modélisation intégrant des données expérimentales plus ou moins complexes.

En réalité, il faut distinguer la notion de flux de la notion de vitesse. Un flux est un débit, c'est-à-dire une quantité de matière (de molécules) convertie par unité de temps. C'est une notion associée à la notion de réaction, c'est-à-dire à une conversion chimique d'une espèce en une autre. La vitesse est quant à elle une notion de vélocité, associée à la catalyse enzymatique. Or, réactions et enzymes ne sont pas du tout équivalents. Une réaction peut être catalysée ou non, et si elle est catalysée, elle peut l'être par une ou par plusieurs enzyme(s). A l'inverse, une enzyme peut catalyser une ou plusieurs réactions différentes. Le terme de vitesse (notée V) sera donc utilisé pour décrire des vitesses enzymatiques et le terme de flux (noté J) sera utilisé pour

décrire des flux de conversion moléculaire, indépendamment d'une éventuelle catalyse ou du nombre de catalyseurs.

En pratique, on distingue différents flux. En prenant comme exemple les deux réactions ci-dessous, on peut distiguer deux flux par réaction : le flux aller et le flux retour (J_1 et J_{-1} pour la première réaction, J_2 et J_{-2} pour la seconde).



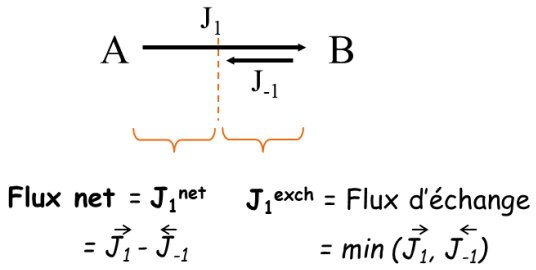
La variation de concentration de B ([B]) dans le temps est décrite par :

$$d[B]/dt = J_1 + J_{-2} - J_{-1} - J_2$$

Et, à l'état stationnaire :

$$d[B]/dt = 0$$

On définit également flux net et flux d'échange, comme suit :



Le flux net représente la différence entre le flux aller et le flux retour. Le flux d'échange représente le nombre de molécules qui sont échangées dans la réaction, il est lié au taux de réversibilité de la réaction. Pour une réaction proche de l'équilibre thermodynamique, on aura :

$$J^{net} << J^{xch}$$

Pour une raction éloignée de l'équilibre thermodynamique, on aura à l'inverse:

$$J^{net} >> J^{xch}$$

6.5.3. Méthodes de modélisation des flux métaboliques

On peut distinguer deux principales approches de modélisation des flux métaboliques: les modèles basés sur la conservation de la matière, incluant les modèles stoechiométriques, la modélisation sous contraintes et la fluxomique par marquage ^{13}C, et les modèles cinétiques (Figure 1.36). Ces méthodes sont utilisées pour modéliser l'activité des systèmes métaboliques soit pour caractériser leurs propriétés fonctionnelles, soit pour expliquer des observations (modèles explicatifs) ou pour générer des hypothèses (modèles prédictifs).

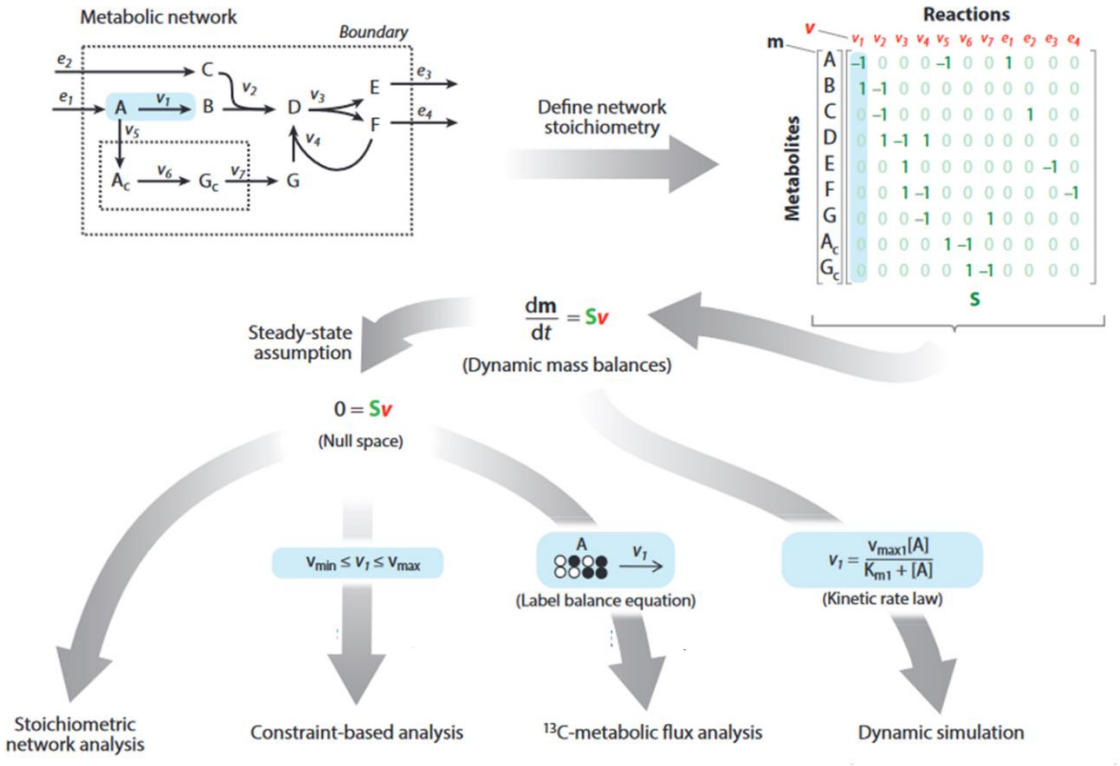


Figure 1.36 : Approches de modélisation métabolique. Illustration adaptée de Clark et al, 2020. Abbréviations : A : concentration du métabolite A ; Km : constante de liaison de Michaelis ; vmax1 : vitesse maximale de la réaction 1

6.5.4. Modélisation cinétique

Le principe des modèles cinétiques est de calculer les flux à partir des propriétés cinétiques des enzymes. Alors que les modèles stoechiométriques sont applicables aux conditions d'état stationnaires métaboliques, le modèles cinétiques peuvent être utilisés pour simuler la réponse dynamique des systèmes métaboliques à différentes perturbations.

Ces modèles sont décrits par un système d'équations différentielles ordinaires (ODE) qui décrivent l'évolution des concentrations et des vitesses de réactions dans le temps :

$$\frac{dM}{dt} = S.v(M(t), \boldsymbol{\theta}), \mathrm{M}(0) = M0$$

Où S est la matrice stœchiométrique et v est le vecteur de flux des réactions calculés selon leurs lois de vitesse à partir des concentrations des espèces M et des paramètres de la loi de vitesse θ.

La loi de vitesse de Michaelis-Menten [Johnson & Goody, 2011] permet d'établir une relation entre la concentration en substrat et les principales caractéristiques de l'enzyme, à savoir sa vitesse maximale v_{max} et son affinité avec le substrat K_M :

$$v_i = \frac{v_{max}[M]}{K_M + [M]}$$

Cette équation prévoit que la vitesse initiale de la réaction i tend vers 0 quand la concentration initiale en substrat [M] tend vers 0 ; et tend asymptotiquement vers une limite v_{max} quand la concentration en substrat tend vers l'infini. D'atres équations cinétiques plus complexes peuvent être utilisées.

Cette méthode de modélisation fournit une image très détaillée du métabolisme mais nécessite des connaissances précises sur les propriétés cinétiques des enzymes et les concentrations de métabolites. Ce type de modèle n'est donc pas adapté aux études à l'échelle d'un organisme à partir de réseaux métaboliques à l'échelle du génome pour lesquels la plupart des réactions n'ont pas été caractérisées *in vivo* [Vasilakou et al, 2016 ; Ramon et al, 2018].

6.5.5. Modélisation thermodynamique

Une autre approche de modélisation est la modélisation thermodynamique. Il s'agit d'une approche qui vient souvent en complément d'une approche de modélisation sous contraintes type FBA. Cette dernière peut conduire à la prédiction de solutions de flux thermodynamiquement infaisables. L'analyse thermodynamique des flux impose des contraintes supplémentaires aux modèles stoechiométriques pour garantir des flux thermodynamiquement valides et permet d''intégrer des données métabolomiques dans les modèles. En comparant les concentrations des substrats et produits d'une réaction à la constante d'équilibre de cette réaction, il est en effet possible d'indiquer la direction spontanée de la réaction. Cette approche est donc utilisée pour contraindre la directionnalité des réactions du modèle.

6.5.6. Modèles basés sur la conservation de la matière

Dans ces approches, on considère simplement que la matière est conservée à chaque nœud du réseau. La somme de toutes les réactions de formation d'un composé est nécessairement égale à la somme de l'accumulation du composé et de toutes les réactions qui l'utilise. On peut distinguer deux déclinaisons principales de ce principe : les modèles stoechiométriques (conservation des espèces chimiques) et ses dérivés, et les modèles isotopiques (conservation des espèces isotopiques).

6.5.6.1. Modèles stoechiométriques

Comme son nom l'indique, cette approche est basée sur la connaissance de la stœchiométrie du réseau métabolique. Le réseau peut en effet être décrit par une matrice stœchiométrique S(*m*,*r*) dont les éléments sont les coefficients stœchiométriques associés aux métabolites *m* dans les réactions *r* (Figure 1.36). Les coefficients négatifs sont associés à la consommation des métabolites tandis que les coefficients positifs sont associés aux métabolites produits. Un coefficient stœchiométrique de zéro est utilisé pour chaque métabolite qui ne participe pas à une réaction particulière. Sur la base du modèle stœchiométrique, le principe de conservation de la matière s'exprime par l'équation :

$$(1) \qquad dM/dt = S.\nu$$

où M représente le vecteur de concentration des métabolites et v le vecteur de flux à travers les réactions du réseau.

A l'état stationnaire métabolique, qui est défini comme un état pour lequel les concentrations des métabolites sont toutes constantes, l'accumulation des composés est nulle. Dans ce cas, l'équation (1) devient :

$$(2) \qquad dM/dt = S.\nu = 0$$

La condition d'état stationnaire amène une autre simplification. Toutes les réactions d'une voie linéaire ont le même flux net et peuvent être représentées par une réaction globale unique. Cela permet de réduire la dimension du système d'équations.

Sur cette base, plusieurs approches de modélisation stoechiométrique ont été développées, en fonction de la complexité du réseau métabolique étudié et des informations expérimentales exploitées.

- ***L'analyse des flux métaboliques (MFA)***

L'analyse des flux métaboliques (MFA) vise à calculer l'ensemble des flux dans le réseau métabolique d'un organisme [Wiechert et al, 2001]. En fonction de la dimension du réseau étudié, l'équation (2) peut admettre une infinité de solution. Le degré de liberté (F= *n*-*m*) permet de définir si un système est déterminé (s'il peut être résolu) ou non. Pour la plupart des systèmes métaboliques, la matrice stœchiométrique correspondante est sous-déterminée (F>0), c'est-à-dire que le nombre de métabolites *m* du réseau est inférieur au nombre de réactions *n* (*n*<*m*). Pour estimer les flux métaboliques, les contraintes stœchiométriques sont

complétées par des mesures de flux extracellulaires (r), telles que les vitesses de consommation des nutriments et de sécrétion des produits qui peuvent être mesurées expérimentalement.

La combinaison de ces données avec les contraintes stœchiométriques à l'état stationnaire (2) conduit au système d'équations suivant :

$$(3) \qquad SSR = \sum \frac{(r - r_m)^2}{\sigma_r^2}$$

Les flux sont estimés en minimisant la somme des carrés des résidus entre les flux extracellulaires mesurés (r_m) et les flux prédits (r) par le modèle. Les résidus sont pondérés par les erreurs de mesure (σ).

Le résultat d'une analyse MFA est une carte qui représente la distribution des flux sur les différentes voies du réseau métabolique (« carte des flux »). La MFA fournit une quantification détaillée de l'activité métabolique réelle et présente un intérêt particulier pour la génomique fonctionnelle et la biologie des systèmes afin d'établir la relation entre le génotype et le phénotype métabolique. Elle a été largement utilisée pour quantifier les flux dans des cellules cultivées en conditions de croissance industrielles pertinentes, telles que la limitation du nutriment ou la présence d'inhibiteurs de croissance [Antoniewicz, 2020]. Cependant elle présente trois limitations majeures. La première est liée à l'état stationnaire, car dans ces conditions on ne peut accéder qu'aux seuls flux nets. La réversibilité des réactions n'est donc pas accessible par cette approche. La seconde limitation est liée à la dimension des réseaux métaboliques et par conséquent au nombre d'équations à résoudre. Les réseaux GSMN contiennent en effet plusieurs centaines voire milliers de réactions alors qu'expérimentalement le nombre de flux extracellulaires mesurables est généralement d'une dizaine. Les systèmes d'équations sont donc très fortement sous-déterminés. En pratique, l'approche MFA est limitée à des réseaux de « petite dimension », la plupart des travaux dans ce domaine étant focalisés sur le métabolisme carboné central. La troisième limitation est d'ordre topologique. En effet, un certain nombre de motifs métaboliques ne sont pas accessibles via une analyse MFA tels que les réactions parallèles, les cycles ou la canalisation (Figure 1.37) [Wiechert et al, 2001].

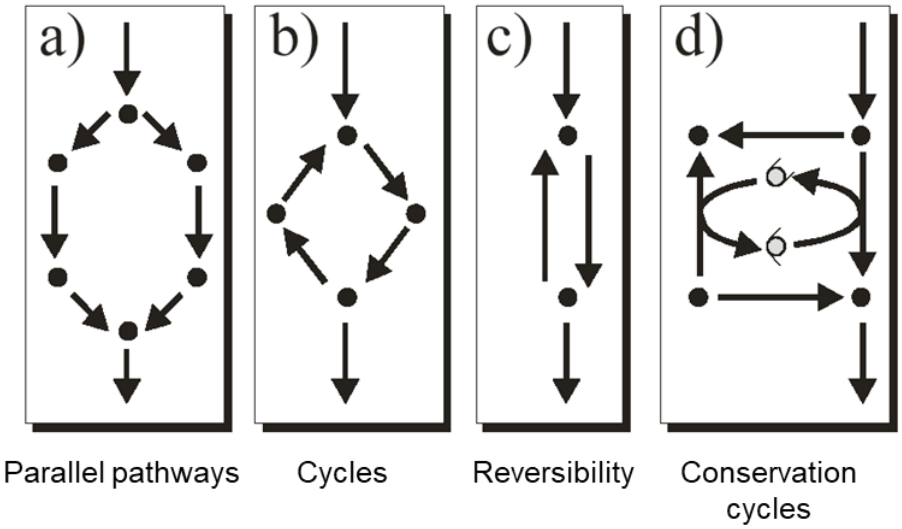


Figure 1.37 : Exemples de motifs de réseau métabolique non accessibles par une analyse MFA. Figure tirée de Wiechert et al, 2001.

- ***Modélisation sous contrainte : analyse des bilans de flux (FBA pour Flux Balance Analysis)***

La modélisation sous contrainte (CBM pour Constraint-based modeling) est par essence une extension de l'approche MFA, en cela qu'il s'agit d'une approche stoechiométrique à l'état stationnaire et qu'il s'agit donc de résoudre l'équation (1) :

$$(1) \qquad dM/dt = S.v = 0$$

La différence avec la MFA est que la modélisation sous contrainte vise à calculer les flux sur la totalité d'un réseau GSMN. La limitation initiale est la très forte sous-détermination du système d'équations obtenu pour un réseau GSMN. Par exemple le modèle *E. coli* iML1515 contient 2712 réactions pour 1877 métabolites [King et al, 2016], ce qui conduit à une matrice de 2712x1877 éléments. En supposant que l'on mesure une dizaine de flux expérimentaux, le degré de liberté du système d'équations est particulièrement élevé. Dans l'approche sous contrainte, cette sous-détermination est levée par l'addition de contraintes sur le réseau de façon à réduire l'espace des solutions possibles pour le système d'équations (Figure 1.38).

$$a_i \leq v_i \leq b_i$$

où ai et bi sont les limites inférieures et supérieures qui définissent les flux maximaux et minimaux admissibles de la réaction i.

Ces contraintes sont définies pour analyser les capacités métaboliques de l'organisme en fonction de la question biologique posée et peuvent être de diverses natures (direction des réactions, limites de valeur de flux, composition de biomasse, …). L'intégration de données

omiques permet notamment d'apporter un grand nombre de contraintes en raison de leur dimension globale, grace au développement d'algorithmes dédiés. Par exemple, l'algorithme GIMME (Gene Inactivity Moderated by Metabolism and Expression) [Machado et al, 2014] utilise des données d'expression des gènes et une ou plusieurs fonctions objectives pour déterminer une distribution de flux qui optimise l'objectif donné et minimise ensuite l'utilisation de réactions inactives.

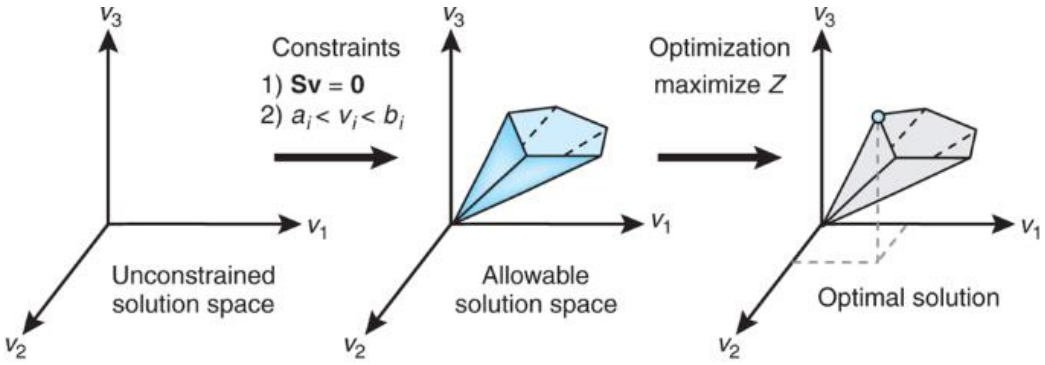


Figure 1.38 : Illustration de la modélisation basée sur les contraintes

.

Toutefois, si elle permet de restreindre l'espace de solutions, cette addition de contraintes n'est pas suffisante pour résoudre le système. La stratégie consiste ensuite à définir un objectif biologiquement pertinent et trouver dans l'espace de solutions restreint une distribution de flux optimale pour cet objectif (Figure 1.38). Mathématiquement, l'objectif est représenté par une "fonction objective" qui peut être simple (un critère d'optimisation) ou complexe (plusieurs critères).

La FBA peut être par exemple être utilisée pour simuler la croissance d'un organisme sur différents milieux ou optimiser la production d'un composé particulier. L'application la plus utilisée est l'optimisation de la production de biomasse (BM) [Orth et al, 2010]. Elle consiste à résoudre le système d'équations linéaires suivant :

$$\text{Maximize } \nu BM$$

$$S.\nu = 0$$

$$ai \leq \nu i \leq bi$$

La FBA est avant tout un outil de prédiction qui est utilisé dans un large éventail d'applications dans le domaine de la biologie des systèmes, l'ingénierie métabolique ou de la santé. Par exemple pour la prédiction de taux de croissance chez *E. coli* [Edwards and Palsson, 2000], pour la prédiction de cibles thérapeutiques [Kim et al, 2012] ou encore pour la prédiction de cibles de manipulation génétique [Park et al, 2012]. La FBA peut également permettre de

compléter les lacunes des réseaux métaboliques à l'échelle du génome en comparant des données produites *in silico* à des résultats expérimentaux [Orth et al, 2010]. Un autre intérêt de la FBA est de pouvoir intégrer différentes données omiques pour affiner le modèle contextuel (ajustement de la matrice stœchiométrique S) ou de fixer les limites de flux [Becker & Palsson, 2008; Colijn et al, 2009; Wang et al, 2012]. On retrouve beaucoup d'applications qui utilisent des données transcriptomiques pour améliorer le modèle [Tian and Reed, 2018 ; Granata et al, 2019].

La FBA repose sur une hypothèse forte de fonctionnement optimal du métabolisme, ce qui n'est pas forcément le cas en conditions physiologiques. Par ailleurs plusieurs – voire un très grand nombre – de solutions peuvent répondre à la fonction objective. Elle peut fournir un espace de solutions plus qu'une solution unique [Antoniewicz et al, 2021].

Il existe des extensions de la FBA [Lewis et al. 2012], incluant notamment :

- L'analyse de la variabilité des flux (FVA pour *Flux Variability Analysis*), qui ne vise pas à identifier un jeu de flux optimal, mais à déterminer la plage de valeurs de flux possibles pour chaque réaction, toujours en lien avec une fonction objective spécifiée. Cela permet d'identifier des réactions métaboliques déterminantes par rapport à l'objectif biologique considéré.
- La FBA régulée (rFBA pour *Regulatory Flux Balance Analysis*), dans laquelle des contraintes de régulation de l'expression des enzymes sont introduites, en lien avec des données d'expression du génome.

6.5.6.2. Modèles isotopiques

Les modèles isotopiques sont basés sur la combinaison d'approches expérimentales de traçage isotopique et d'équations décrivant le devenir de l'isotope dans le métabolisme. Ces modèles sont construits de façon analogue aux modèles stoichiométriques. Ils reposent sur la conservation des espèces isotopiques à travers les réactions de formation et d'utilisation d'un métabolite donné. Ils peuvent être développés de manière indépendante de toute autre donnée (e.g. approche MetaFor, [Szyperski et al 1999]) donnant accès à des distributions relatives de flux. Ils peuvent aussi être combinés à la modélisation stoechiometrique et dans ce cas-là ils permettent non seulement d'obtenir des valeurs absolues de flux mais aussi d'améliorer la précision sur leur mesure. Ces approches peuvent se décliner à l'état stationnaire, mais aussi en conditions (semi-) dynamiques. Comme elles sont au cœur du travail de thèse, elles seront présentées plus en détail dans la partie 7 de cette introduction.

7. Fluxomique par marquage isotopique

L'analyse des flux métaboliques par marquage isotopique ^{13}C (^{13}C-MFA) permet la quantification des flux à partir de données issues d'expériences de marquage isotopique. Le principe général est présenté sur la figure 1.39. Il inclut une phase expérimentale permettant de fournir des données d'incorporation isotopique dans les métabolites et une phase de calcul qui permet d'extraire des valeurs de flux à partir des données expérimentales de marquage. La phase expérimentale est une expérience de traçage isotopique telle que décrite au début de ce chapitre, pour laquelle toutes les considérations liées au choix du traceur vues précédemment restent valides. La phase de calcul repose sur des modèles isotopiques, c'est à dire des modèles contenant un système d'équations décrivant le devenir de l'isotope au sein du réseau. Les modèles les plus complets associent modélisation stœchiométrique et modélisation isotopique du réseau métabolique étudié.

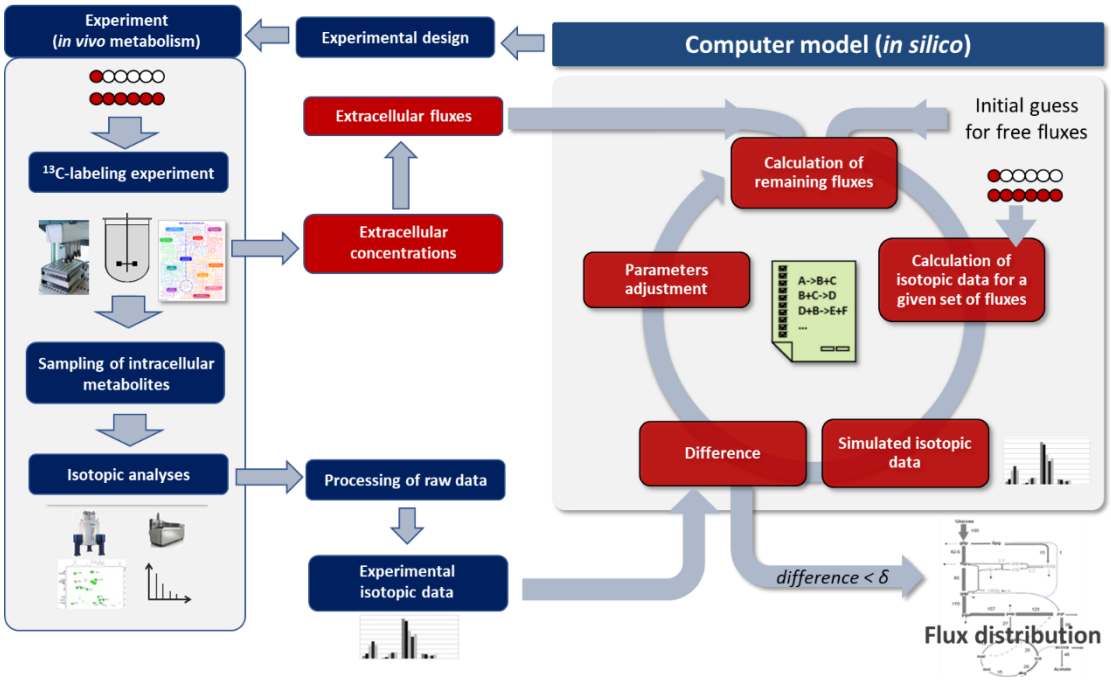


Figure 1.39 : Protocole des approches de fluxomique. Les encadrés en bleu correspondent aux étapes communes de la fluxomique avec les approches de profilage isotopique.

Pour permettre la quantification des flux intracellulaires, la ^{13}C-MFA s'appuie sur les mesures des flux extracellulaires, ainsi que sur les mesures de marquage isotopique issues d'une ou de plusieurs expériences de traçage isotopique. Ces mesures sont utilisées avec un modèle de réseau métabolique pour trouver la solution de flux la plus probable, c'est-à-dire celle dans laquelle le modèle correspond le mieux aux schémas de marquage des métabolites mesurés (Figure 1.38).

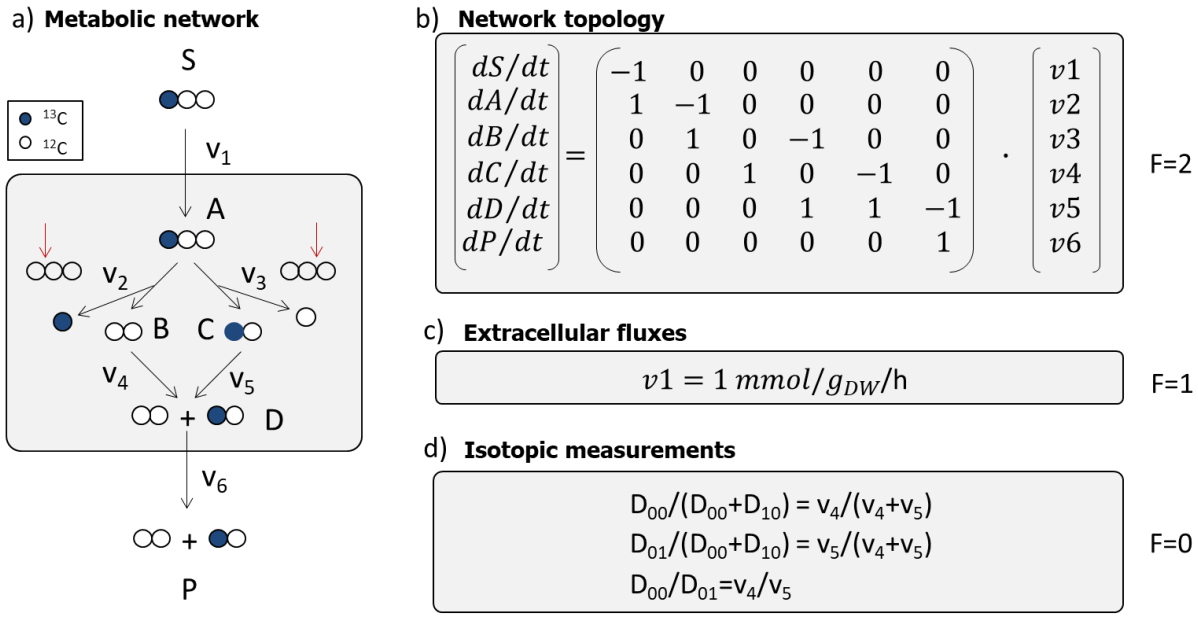


Figure 1.40 : Principe général de l'analyse ^{13}C-MFA. a) Représentation d'un réseau métabolique simple qui consomme le substrat S pour former le produit P en 4 étapes. b) Matrice stœchiométrique S calculée sur la base de la topologie du réseau. À l'état stationnaire métabolique, S.v=dc/dt=0, où v est le vecteur de flux. c) Les flux extracellulaires peuvent être déterminés à partir de la mesure des concentrations des métabolites extracellulaires au cours du temps. d) Équations supplémentaires obtenues à partir du bilan de masse des isotopes pour fixer les degrés de liberté F restants.

- ***Modèle métabolique***

Le calcul des flux intracellulaires est réalisé à partir d'un modèle de réseau métabolique qui doit inclure les réactions et leur transitions respectives d'atomes de carbone. Le modèle doit pouvoir expliquer les origines possibles de tous les carbones. L'exemple ci-dessus (Figure 1.40a) présente un réseau métabolique simple contenant deux voies parallèles v2 et v3 entrainant respectivement la perte du C1 et du C3 du métabolite A. La première étape de l'analyse ^{13}C-MFA consiste à modéliser le réseau sous la forme de la matrice stœchiométrique calculée sur la base de la topologie du réseau à l'état d'équilibre métabolique (Figure 1.40b).

- ***Flux extracellulaires***

Les flux extracellulaires correspondent au flux d'échange de matière entre les cellules et leur environnement. Ils doivent être mesurés pour fournir des contraintes supplémentaires permettant la détermination des flux intracellulaires. Ces flux extracellulaires peuvent être définis à partir de mesures des concentrations au cours du temps de la biomasse et des métabolites extracellulaires. Ils peuvent être déterminés à l'aide de modèles mathématiques tels que décrit par l'outil Physiofit, qui simule les flux à partir des données observées [Peiro et al, 2019]. Le modèle général, qui tient compte de la dégradation non enzymatique des substrats ou

des produits, relie les changements de concentrations aux flux en utilisant le système suivant d'équations différentielles ordinaires :

$$\frac{dX}{dt} = \mu . X(t) \quad (1)$$

$$\frac{dM_i}{dt} = q_{M_i} . X(t) \quad (2)$$

Où X est la concentration de la biomasse (g_{DW}/L), μ est le facteur de croissance (h^{-1}) et M_i est la concentration du métabolite extracellulaire i (mmol/L) avec un flux d'échange q_{M_i}(mmol. $g_{DW}^{-1}.h^{-1}$). L'intégration des équations (5) et (6) fournit les fonctions analytiques suivantes :

$$X(t) = X_0 . e^{\mu . t} \quad (3)$$

$$M_i(t) = M_i^0 + q_{M_i} . \frac{X_0}{\mu} . (e^{\mu . t} - 1) \quad (4)$$

Étant donné que les concentrations de biomasse sont généralement mesurées en tant que densités optiques (valeurs OD_{600}) à l'aide d'un spectrophotomètre, un facteur de conversion OD_{600}/concentration de biomasse approprié doit également être déterminé afin de convertir les valeurs OD_{600} en grammes de poids sec par litre (g_{DW}/L) pour les concentrations de biomasse [Long and Antoniewicz, 2019].

- ***Données isotopiques***

Les données isotopiques sont mesurées à l'aide d'une ou plusieurs techniques analytiques telles que les méthodes décrites précédemment (section 3.5). A partir des données brutes obtenues, les informations isotopiques doivent être extraites pour déterminer la répartition des atomes du traceur ^{13}C dans les métabolites. Par exemple sur la Figure 1.40d, D01/(D00+D10) représente la proportion relative du métabolite D avec 1 atome de ^{13}C par rapport au pool total de D. Ces équations isotopiques fournissent des contraintes supplémentaires pour le calcul des flux intracellulaires.

7.1.Principe du calcul de flux

Le calcul des flux est basé sur la simulation des profils isotopiques des métabolites intracellulaires. Pour effectuer cette simulation, on considère un jeu initial de valeurs de flux. Sur la base de ces valeurs initiales de flux, de la composition du substrat d'entrée et de la mesure des flux extracellulaires, les données isotopiques peuvent être simulées grace aux équations stoechiométriques et isotopiques (Figure 1.38). Les données isotopiques simulées (x) avec ces données sont ensuite comparées avec les données isotopiques expérimentales (x_m) afin de

calculer leur écart, qui est généralement représenté mathématiquement par une somme des carrés où chaque valeur est pondérée par l'écart type (σ) de la mesure correspondante, appelée résidu

$$(5) \qquad SSR = \sum \frac{(x-x_m)2}{\sigma_x^2} + \sum \frac{(r-r_m)2}{\sigma_r^2}$$

Le modèle ajuste ensuite itérativement les valeurs de flux afin de minimiser le résidu jusqu'à ce qu'un ajustement optimal soit obtenu. A chaque itération, le marquage isotopique des métabolites intracellulaires est simulé pour un ensemble de flux en résolvant un grand nombre d'équations. En raison de la taille et de la non-linéarité du système d'équation reliant les flux métaboliques aux données isotopiques, cette procédure d'ajustement est un processus lent et complexe. Récemment, de nouvelles méthodes mathématiques ont été développées pour réduire la dimension du problème et rendre la simulation des données plus efficace [Antoniewicz et al, 2007]. Plusieurs logiciels ont été développés pour calculer les flux intracellulaires tels que INCA [Young et al, 2014], 13CFlux2 [Weitlzel et al, 2013] ou influx_si [Sokol et al, 2012].

Afin d'éviter des interprétations biologiques erronées, une analyse statistique est appliquée à la fin de la procédure d'ajustement des paramètres. Elle permet de vérifier la qualité de l'ajustement (c'est-à-dire vérifier si les données expérimentales corrèlent avec les données simulées) et déterminer des intervalles de confiance sur les flux. Dans cet objectif, plusieurs méthodes ont été développées mais les plus rigoureuses sont la simulation de Monte-Carlo [Quek et al, 2009] ou la mesure du Chi2 (χ2) [Antoniewicz et al, 2006]. Le test χ2 est appliqué à la somme minimale des carrées obtenue pour valider la qualité de l'ajustement. Pour être valide, la plage de valeur SSR doit être comprise entre :

$$\frac{x^2\alpha}{2(n-p)} \quad \text{et} \quad \frac{x^2(1-\alpha)}{2(n-p)}$$

Où n représente le nombre de mesures ajustées, p le nombre de paramètres estimés et α = 0.05 pour un intervalle de confiance de 95%.

Si la valeur mesurée est trop élevée, cela peut révéler plusieurs erreurs : (i) le modèle métabolique utilisé pour mesurer les flux est incomplet ; (ii) la réversibilité est non comprise dans la modélisation ou (iii) des erreurs de mesures des données isotopiques [Antoniewicz, 2018]. Les flux intracellulaires sont dits structurellement identifiables s'il existe un ensemble unique de valeurs de flux produisant les données mesurées [Wiechert et al, 2001].

Le résultat d'une analyse ^{13}C-MFA est une carte de flux qui montre la distribution des flux dans le réseau métabolique (Figure 1.41).

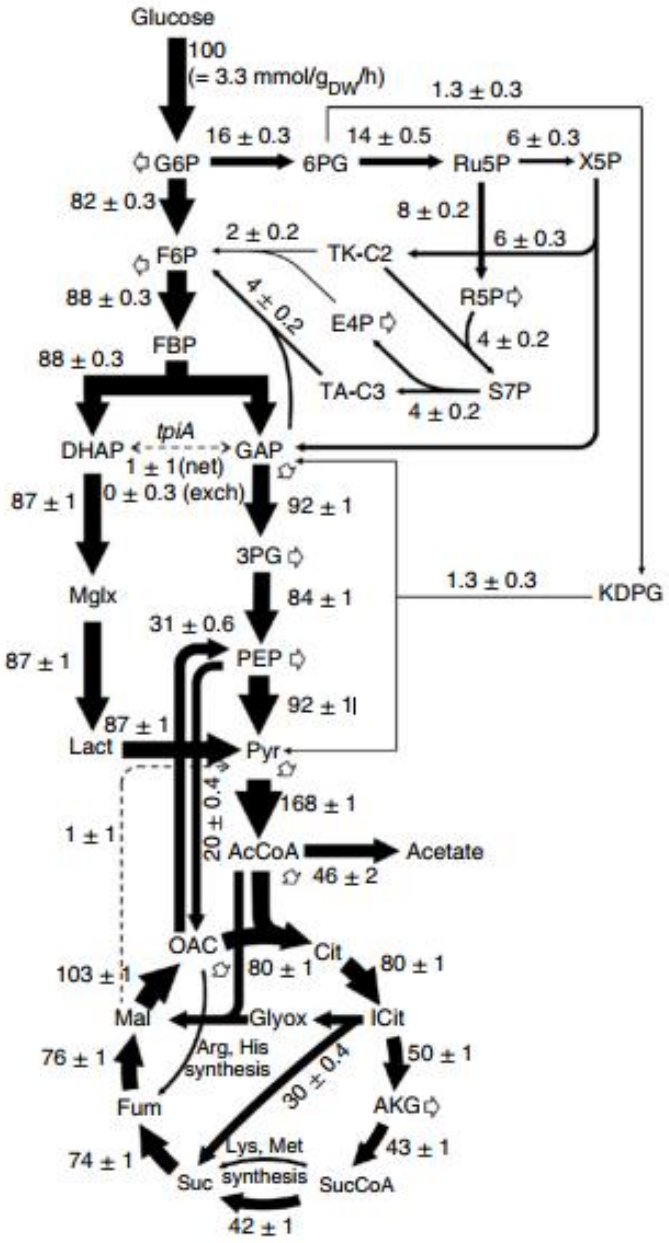


Figure 1.41 : Carte de flux. Exemple d'une carte de flux sur une souche E. coli à partir de mesures GC-MS. Les flux (fit ± sd) ont été normalisés à un taux d'absorption du glucose de 100. Illustration tirée de Long and Antoniewicz, 2019.

7.2.Stratégies d'analyses de flux

Le niveau d'information obtenu lors d'une analyse de flux dépend en premier de l'expérience de marquage qui est réalisée. Il existe différentes approches de marquage isotopique au ^{13}C pour analyser les flux métaboliques : l'approche stationnaire, l'approche instationnaire et l'approche dynamique (Figure 1.42). Le choix de l'une ou l'autre de ces approches pour une étude spécifique va être déterminé en fonction de la question biologique, de l'organisme étudié et des conditions biologiques et expérimentales.

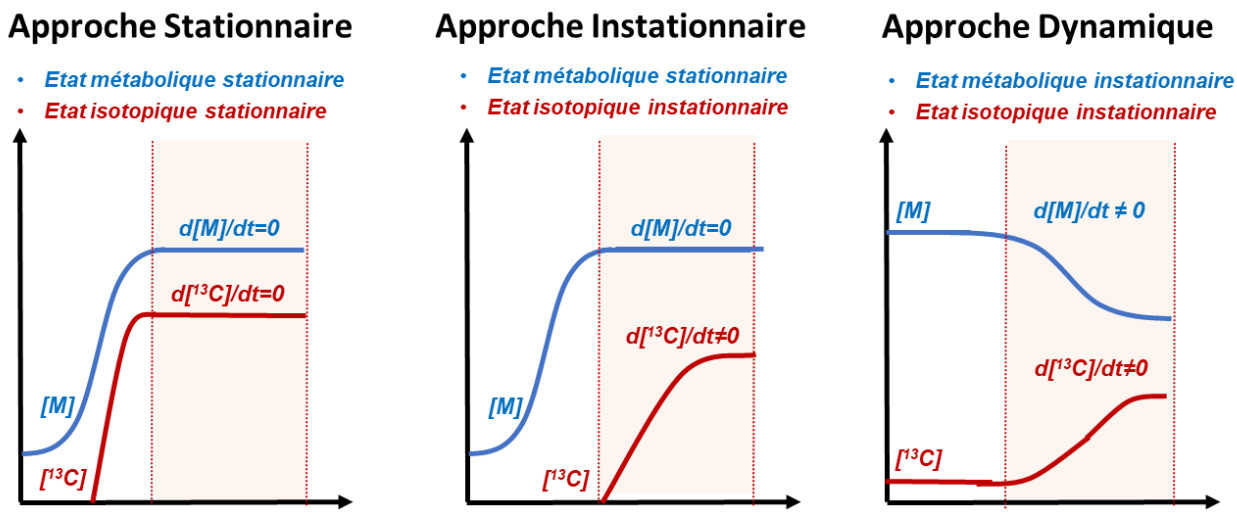


Figure 1.42 : Comparaison des différentes approches d'analyse des flux par marquage isotopique. [M] : concentration métabolique, [13C] : enrichissement isotopique, F : Flux métabolique. Les zones en rose correspondent aux états physiologiques et de marquage utilisés dans les expériences et les simulations.

7.2.1. L'approche stationnaire

L'approche stationnaire correspond à l'exploitation de données de marquage collectées lorsque la concentration des métabolites intracellulaires est stable dans le temps (état stationnaire métabolique) et que leur composition isotopique l'est également (état stationnaire isotopique). Cela conduit à des simplifications mathématiques (les équations sont toutes linéaires) et pratiques (absence de contrainte temporelle pour la collecte d'échantillons, analyse de composés fortement concentrés, etc).

Le temps nécessaire au système biologique pour atteindre ces états stationnaires métabolique et isotopique dépend de plusieurs facteurs tels que l'activité des cellules, les pools de métabolites mesurés, les voies métaboliques concernées, le substrat utilisé comme traceur et la composition du milieu. Par exemple lors d'une expérience de marquage au ^{13}C-glucose chez *E. coli*, l'état stationnaire isotopique est atteint en quelques secondes pour la voie de la glycolyse, en plusieurs minutes pour le cycle de Krebs et en plusieurs heures pour les nucléotides.

L'approche stationnaire est couramment utilisée par sa simplicité de mise en place expérimentale. Puisque le marquage isotopique est stable dans le temps, la collecte d'un seul échantillon est suffisante pour mesurer les profils isotopiques des métabolites. Il est également possible d'exploiter le marquage des produits terminaux du métabolisme plutôt que les intermédiaires. En effet, en conditions stationnaires le marquage des produits métaboliques reflète directement celui de leurs précurseurs. Ils s'accumulent cependant de manière beaucoup

plus importante que ces derniers, ce qui est avantageux en termes de sensibilité et de précision des mesures des marquage isotopique. Elle permet également de mesurer la simplification des équations mathématiques linéaires.

L'analyse ^{13}C-MFA a été largement appliquée pour mieux comprendre le métabolisme d'organismes tels que *E. coli* ou Saccharomyces cerevisiae. Elle a notamment permis de découvrir de nouvelles voies métaboliques ou des fonctions enzymatiques qui n'avaient pas été prédites par les annotations du génome. Parmi les récentes découvertes chez *E. coli* on peut citer la découverte d'une nouvelle réaction dans la voie des pentoses phosphates [Nakahigashi et al., 2009] ou la découverte d'un flux inverse du pyruvate vers le phosphoenolpyruvate [Long et al, 2017].

Cette approche est valide pour étudier des systèmes biologiques qui se maintiennent – ou peuvent être maintenues – en condition d'état stationnaire métabolique suffisamment longue pour atteindre l'état stationnaire isotopique. Cependant, il existe de nombreux cas de figures pour lesquels cet état stationnaire peut être difficile à atteindre, notamment les cellules mammifères qui sont des systèmes plus complexes et peuvent mettre plusieurs jours pour l'atteindre [Murphy et al, 2013]. Elle ne reflète également pas la réalité physiologique, la situation d'état stationnaire étant rare dans la nature.

7.2.2. L'approche instationnaire isotopique

L'approche isotopique instationnaire correspond à des expériences réalisées en condition d'état stationnaire métabolique pour lesquelles on mesure la cinétique d'incorporation isotopique dans les métabolites. La principale différence entre avec l'approche précédente est que dans ces conditions, les équations différentielles ordinaires (ODE) décrivant la conservation de l'isotope dans le réseau ne se linéarisent pas. Il est donc nécessaire d'intégrer la dimension temporelle et de collecter expérimentalement les cinétiques de propagation de l'isotope au sein du réseau (x(t)). Les techniques de régression non-linéraires des moindres carrés sont similaires à celles utilisées à l'état stationnaire isotopique. Cependant, en plus de l'estimation des flux métaboliques, les tailles des pools de métabolites sont ajustées (C) pour tenir compte des données transitoire de marquage observées [Wiechert et al, 2005] :

$$(6) \qquad SSR = \sum \frac{(x(t)-x_m(t))2}{\sigma_x^2} + \sum \frac{(r-r_m)2}{\sigma_r^2} + \sum \frac{(C-C_m)^2}{\sigma_C^2}$$

Sur le plan expérimental, le traceur isotopique n'est introduit qu'une fois l'état stationnaire métabolique atteint, puis des échantillons sont collectés après différents temps pour mesurer l'incorporation du traceur isotopique au cours du temps. Même si les réseaux métaboliques peuvent être de taille similaire, les temps de simulation pour l'approche instationnaire sont généralement beaucoup plus longs en raison du nombre accru de points de mesure qui doivent être simulés.

L'approche instationnaire permet d'accéder à de nouvelles informations sur la topologie des réseaux métaboliques et sur leurs activités. Elle offre la possibilité de quantifier le recyclage des pools de réserve ou encore de mesurer de manière indirecte des pools métaboliques non mesurables par des méthodes analytiques [Nöh and Wiechert, 2011]. Les approches instationnaires permettent également de mesurer les flux métaboliques chez des organismes autotrophiques ou des organismes à croissance lente ou nulle chez lesquelles la réalisation d'expériences isotopiques stationnaires n'est pas informative (ex : organismes méthylotrophes) [Anh et Antoniewitz, 2013].

7.2.3. L'approche dynamique

Il s'agit de l'approche la plus complexe puisque le système n'est ni à l'état stationnaire métabolique ni à l'état stationnaire isotopique. Sur le plan mathématique, il n'y a aucune simplification des équations stœchiométriques ni des équations isotopiques. Cela veut dire que sur le plan expérimental, il est nécessaire de mesurer à la fois la cinétique d'évolution de la concentration (par métabolomique quantitative) et de la cinétique de marquage de chacun des métabolites du réseau. Bien qu'il s'agisse de l'approche permettant potentiellement l'apport d'information le plus riche sur l'activité du réseau métabolique, elle n'est que peu exploitée directement en raison de la complexité expérimentale et de l'absence d'outils permettant de résoudre la distribution des flux sur la base seule des expériences de marquage. L'approche dynamique peut cependant être mise en place en couplant la ^{13}C-fluxomique à des modèles cinétiques du réseau métaboliques (couplage de modèles). Dans ce cas-là, le modèle est paramétré grâce aux caractéristiques cinétiques des enzymes, puis les données de métabolomique quantitative et de marquage isotopique sont simulées pour calculer les valeurs de flux. Cette approche a par exemple été utilisée pour comprendre la réorganisation du métabolisme du glucose des adipocytes en réponse à l'insuline [Quek et al, 2020].

7.2.4. Approches parallèles : COMPLETE-MFA

L'analyse des flux métaboliques a connu récemment une avancée importante : l'utilisation d'expériences de marquage en parallèle combinées à une intégration rigoureuse des données afin d'estimer avec une grande précision les flux métaboliques dans des systèmes complexes. Cette stratégie d'analyse est appelée COMPLETE-MFA (« complementary parallel labeling experiments technique for metabolic flux analysis ») [Leighty and Antoniewicz, 2013].

La sélection optimale du traceur est désormais bien reconnue comme une étape importante de l'analyse des flux [Antoniewicz, 2013]. Il est toutefois devenu de plus en plus clair qu'il n'existe pas de traceur unique optimal qui permette de déterminer tous les flux d'un modèle métabolique avec une grande précision [Leighty and Antoniewicz, 2013]. Les expériences de marquage en parallèle offrent une solution à ce problème et nécessitent peu d'augmentation de l'effort expérimental, analytique et informatique. Dans ce type d'approche, plusieurs traceurs isotopiques sont utilisés en parallèle. Les données isotopiques générées via les différentes expériences, sont analysées en ajustant simultanément les mesures à un modèle de flux unique. Cette approche permet d'adapter les traceurs individuels à des parties spécifiques du modèle de réseau, de sorte que tous les flux d'intérêt peuvent être estimés avec une grande précision.

Les stratégies de marquage peuvent être différentes selon le modèle biologique étudié. Par exemple des traceurs isotopiques possédant un marquage positionnel différent seront utilisés pour les organismes simples cultivés sur une source de carbone unique. Pour les organismes complexes (cellules mammifères), deux substrats différents pourront être utilisés, tels que le glucose et la glutamine (Figure 1.43). Cette approche a été utilisée via l'utilisation de deux traceurs [1,2-^{13}C]glucose et de la [U-^{13}C]glutamine pour déterminer les flux métaboliques dans les cellules CHO [Ahn and Antoniewicz, 2013]. Un autre avantage de cette approche est que le temps de marquage peut être réduit de manière significative en introduisant plusieurs points d'entrée des isotopes dans le métabolisme.

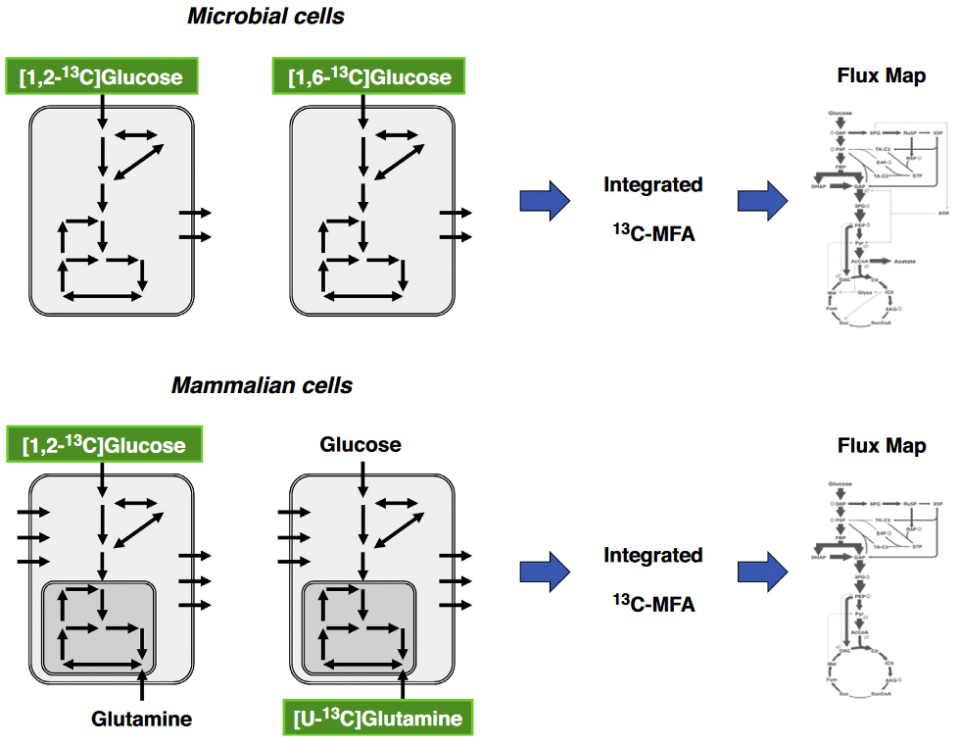


Figure 1.43 : Stratégies de marquage en parallèle selon l'organisme considéré pour une analyse COMPLETE-MFA. Illustration tirée de Antoniewicz, 2015.

7.2.5. Décomposition des flux : approche scalable

L'approche ScalaFlux ((« Scalable Metabolic Flux Analysis ») est une approche alternative qui consiste à mesurer les flux intracellulaires à l'échelle de sous-réseaux métaboliques que l'on peut ensuite combiner pour reconstruire le réseau global [Millard et al, 2020]. Dans ce cadre, les flux sont quantifiés en simulant la propagation du marquage directement à partir des précurseurs métaboliques de ces sous-réseaux. L'intérêt de cette approche est qu'elle permet de s'affranchir de toute mesure ou information supplémentaire telle que les flux extracellulaires, le marquage du nutriment ou la connaissance des réactions et transitions carbonées en amont du sous-réseau, ce qui réduit les exigences en matière d'expérimentation et de calcul. Elle exploite les concepts de l'approche instationnaire isotopique en simulant la propagation du marquage dans les métabolites intracellulaires au cours du temps et respecte ainsi la topologie du réseau (cycle, voies convergentes ou tout autre motif composant les réseaux métaboliques). Cette approche est générique et peut être utilisé pour des mesures de flux à haut débit dans pratiquement tous les systèmes métaboliques, à partir d'expériences de traçage isotopique (au ^{15}N, ^{13}C ou autre traceur isotopique), et de n'importe quel type de mesures isotopiques (MS, MS/MS, RMN…).

Cette approche récente a notamment été utilisée pour l'étude du métabolisme des plantes en estimant les flux métaboliques à partir de marquage au ^{15}N dans les feuilles de *Brassica napus* [Dellero et al, 2020]. Elle sera également utilisée dans le cadre de l'approche de reconstruction métabolique *ab initio* développée dans le chapitre 3 de ce manuscrit de thèse.

7.3.Avantages et limites

Le principal avantage de l'analyse ^{13}C-MFA est que les mesures de marquage isotopiques fournissent un grand nombre de contraintes pour l'estimation des flux. Par exemple 50 à 100 mesures de marquage isotopique vont permettent d'estimer 10 à 20 flux [Antoniewicz, 2020]. La précision des résultats de l'analyse des flux dépend à la fois de la précision de l'ensemble des mesures et de la précision du modèle utilisé pour interpréter les données [Wiechert et al., 2001]. Pour assurer la fiabilité de l'interprétation biologique, une attention particulière doit être portée à la qualité des données isotopiques mesurées [Butin et al, 2022].

L'analyse ^{13}C-MFA reste cependant un outil coûteux, complexe et à faible débit. Généralement, peu de conditions sont étudiées et dans la plupart des cas sans réplicats biologiques. La demande croissante de phénotypage complet en biologie des systèmes entraîne la nécessité d'approches à plus haut débit permettant l'étude de vastes ensembles d'organismes, de souches ou de conditions physiologiques [Bergès et al, 2021].

7.4.Couplage de modèles

Chaque type de modèlisation présente ses avantages et ses défauts. Une stratégie en développement consiste à coupler les modèles, c'est-à-dire à intégrer dans une représentation unique du réseau les différentes approches de modélisation. Il s'agit en général de modèles de type FBA intégrant soit des données thermodynamiques, soit des équations isotopiques, soit encore des équations cinétiques, ou encore des combinaions de ces trois types d'information.

8. Objectifs de recherche

L'un des principaux défis de la biologie des systèmes est d'identifier le lien entre structure et capacité fonctionnelle dans des organismes vivants. L'étude du métabolisme à l'échelle du système est un domaine complexe qui combine des approches expérimentales et outils informatiques puissants pour atteindre l'objectif global d'expliquer et de prédire les comportements cellulaires complexes des systèmes biologiques. Dans ce contexte, l'analyse du métabolisme d'un organisme à l'échelle du réseau métabolique est une approche puissante. Ce travail de thèse s'inscrit dans la biologie des systèmes pour répondre à un besoin d'approches méthodologiques et d'outils bio-informatiques robustes et puissants dans ce domaine.

L'objectif principal de ce travail de thèse est de mettre en place une approche de fluxomique originale basée sur les données pour la reconstruction *ab initio* de réseaux métaboliques. Cette approche permettra d'accéder au réseau métabolique spécifique d'un organisme dans un contexte donné.

Les résultats de ce travail sont présentés en trois grandes parties :

1- Un des premiers objectifs de cette thèse a consisté à développer une une approche de profilage isotopique non-ciblée par spectrométrie de masse haute résolution. Cela inclut l'acquisition des données de marquage par LC-HRMS mais aussi et principalement la mise en place des outils de traitement de données associées. Ce travail a fait l'objet d'une publication proposant une méthodologie permettant de s'assurer de la fiabilité des données isotopiques obtenues par ces approches.

2- Le deuxième objectif consiste au développement d'une approche de fluxomique dite « data-driven ». Elle est basée sur des stratégies d'analyses non-ciblées du marquage isotopique en dynamique par couplage LC-HRMS. Dans ce cadre, des outils informatiques ont été développés pour valider l'approche dans un premier temps à l'échelle de sous-réseaux métaboliques.

3- Pour finir, une approche de fluxomique à haut-débit a été développée. Elle permet l'optimisation et l'automatisation de l'ensemble du processus de fluxomique à hau-débit et l'intégration d'outils pertinentes. Elle a fait l'objet d'une publication pour étudier le fluxotype de 180 souches d'*E. coli*. Le fluxotype réfère à la distribution particulière de flux métaboliques mesurés pour une souche donnée dans des conditions physiologiques données.

Références

Adebiyi, A. O., Jazmin, L. J., & Young, J. D. (2015). 13 C flux analysis of cyanobacterial metabolism. *Photosynthesis Research*, *126*(1), 19–32. https://doi.org/10.1007/s11120-014-0045-1

Adeva, M., González-Lucán, M., Seco, M., & Donapetry, C. (2013). Enzymes involved in l-lactate metabolism in humans. *Mitochondrion*, *13*(6), 615–629. https://doi.org/10.1016/j.mito.2013.08.011

Aguilar-Mogas, A., Sales-Pardo, M., Navarro, M., Guimerà, R., & Yanes, O. (2017). iMet: A Network-Based Computational Tool To Assist in the Annotation of Metabolites from Tandem Mass Spectra. *Analytical Chemistry*, *89*(6), 3474–3482. https://doi.org/10.1021/acs.analchem.6b04512

Ahn, W. S., & Antoniewicz, M. R. (2013). Parallel labeling experiments with [1,2-(13)C]glucose and [U-(13)C]glutamine provide new insights into CHO cell metabolism. *Metabolic Engineering*, *15*, 34–47. https://doi.org/10.1016/j.ymben.2012.10.001

Amara, A., Frainay, C., Jourdan, F., Naake, T., Neumann, S., Novoa-Del-Toro, E. M., et al. (2022). Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Frontiers in Molecular Biosciences*, *9*, 841373. https://doi.org/10.3389/fmolb.2022.841373

Antoniewicz, M. R. (2013). 13C metabolic flux analysis: optimal design of isotopic labeling experiments. *Current Opinion in Biotechnology*, *24*(6), 1116–1121. https://doi.org/10.1016/j.copbio.2013.02.003

Antoniewicz, M. R. (2015). Parallel labeling experiments for pathway elucidation and 13C metabolic flux analysis. *Current Opinion in Biotechnology*, *36*, 91–97. https://doi.org/10.1016/j.copbio.2015.08.014

Antoniewicz, M. R. (2018). A guide to 13C metabolic flux analysis for the cancer biologist. *Experimental & Molecular Medicine*, *50*(4), 1–13. https://doi.org/10.1038/s12276-018-0060-y

Antoniewicz, M. R. (2020). A guide to deciphering microbial interactions and metabolic fluxes in microbiome communities. *Current Opinion in Biotechnology*, *64*, 230–237. https://doi.org/10.1016/j.copbio.2020.07.001

Antoniewicz, M. R. (2021). A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications. *Metabolic Engineering*, *63*, 2–12. https://doi.org/10.1016/j.ymben.2020.11.002

Antoniewicz, M. R., Kelleher, J. K., & Stephanopoulos, G. (2006). Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metabolic Engineering*, *8*(4), 324–337. https://doi.org/10.1016/j.ymben.2006.01.004

Antoniewicz, M. R., Kelleher, J. K., & Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic Engineering*, *9*(1), 68–86. https://doi.org/10.1016/j.ymben.2006.09.001

Arakawa, K., Yamada, Y., Shinoda, K., Nakayama, Y., & Tomita, M. (2006). GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics*, *7*(1), 168. https://doi.org/10.1186/1471-2105-7-168

Ayala-Cabrera, J. F., Montero, L., Meckelmann, S. W., Uteschil, F., & Schmitz, O. J. (2023). Review on atmospheric pressure ionization sources for gas chromatography-mass spectrometry. Part II:

Current applications. *Analytica Chimica Acta*, *1238*, 340379. https://doi.org/10.1016/j.aca.2022.340379

Babele, P. K., & Young, J. D. (2020). Applications of stable isotope-based metabolomics and fluxomics toward synthetic biology of cyanobacteria. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, *12*(3), e1472. https://doi.org/10.1002/wsbm.1472

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, *7*, 20. https://doi.org/10.1186/s13321-015-0069-3

Becker, S. A., & Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, *4*(5), e1000082. https://doi.org/10.1371/journal.pcbi.1000082

Benedetti, E., Pučić-Baković, M., Keser, T., Gerstner, N., Büyüközkan, M., Štambuk, T., et al. (2020). A strategy to incorporate prior knowledge into correlation network cutoff selection. *Nature Communications*, *11*(1), 5153. https://doi.org/10.1038/s41467-020-18675-3

Bergès, C., Cahoreau, E., Millard, P., Enjalbert, B., Dinclaux, M., Heuillet, M., et al. (2021). Exploring the Glucose Fluxotype of the E. coli y-ome Using High-Resolution Fluxomics. *Metabolites*, *11*(5), 271. https://doi.org/10.3390/metabo11050271

Bernstein, D. B., Sulheim, S., Almaas, E., & Segrè, D. (2021). Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology*, *22*, 64. https://doi.org/10.1186/s13059-021-02289-z

Bingol, K., & Brüschweiler, R. (2015). NMR/MS Translator for the Enhanced Simultaneous Analysis of Metabolomics Mixtures by NMR Spectroscopy and Mass Spectrometry: Application to Human Urine. *Journal of Proteome Research*, *14*(6), 2642–2648. https://doi.org/10.1021/acs.jproteome.5b00184

Bolten, C. J., Kiefer, P., Letisse, F., Portais, J.-C., & Wittmann, C. (2007). Sampling for metabolome analysis of microorganisms. *Analytical Chemistry*, *79*(10), 3843–3849. https://doi.org/10.1021/ac0623888

Bongaerts, M., Bonte, R., Demirdas, S., Jacobs, E. H., Oussoren, E., van der Ploeg, A. T., et al. (2020). Using Out-of-Batch Reference Populations to Improve Untargeted Metabolomics for Screening Inborn Errors of Metabolism. *Metabolites*, *11*(1), 8. https://doi.org/10.3390/metabo11010008

Borodina, I., & Nielsen, J. (2005). From genomes to in silico cells via metabolic networks. *Current Opinion in Biotechnology*, *16*(3), 350–355. https://doi.org/10.1016/j.copbio.2005.04.008

Boros, L. G., Lerner, M. R., Morgan, D. L., Taylor, S. L., Smith, B. J., Postier, R. G., & Brackett, D. J. (2005). [1,2-13C2]-D-glucose profiles of the serum, liver, pancreas, and DMBA-induced pancreatic tumors of rats. *Pancreas*, *31*(4), 337–343. https://doi.org/10.1097/01.mpa.0000186524.53253.fb

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., & Barrett, M. P. (2006). Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics: Official Journal of the Metabolomic Society*, *2*(3), 155–164. https://doi.org/10.1007/s11306-006-0029-z

Brekke, E. M. F., Walls, A. B., Schousboe, A., Waagepetersen, H. S., & Sonnewald, U. (2012). Quantitative importance of the pentose phosphate pathway determined by incorporation of 13C from [2-13C]- and [3-13C]glucose into TCA cycle intermediates and neurotransmitter amino acids

in functionally intact neurons. *Journal of Cerebral Blood Flow & Metabolism*, *32*(9), 1788–1799. https://doi.org/10.1038/jcbfm.2012.85

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, *36*(3), 272–281. https://doi.org/10.1038/nbt.4072

Buescher, J. M., Antoniewicz, M. R., Boros, L. G., Burgess, S. C., Brunengraber, H., Clish, C. B., et al. (2015). A roadmap for interpreting 13C metabolite labeling patterns from cells. *Current Opinion in Biotechnology*, *34*, 189–201. https://doi.org/10.1016/j.copbio.2015.02.003

Bueschl, C., Kluger, B., Neumann, N. K. N., Doppler, M., Maschietto, V., Thallinger, G. G., et al. (2017). MetExtract II: A Software Suite for Stable Isotope-Assisted Untargeted Metabolomics. *Analytical Chemistry*, *89*(17), 9518–9526. https://doi.org/10.1021/acs.analchem.7b02518

Burgess, K. E. V., Borutzki, Y., Rankin, N., Daly, R., & Jourdan, F. (2017). MetaNetter 2: A Cytoscape plugin for ab initio network analysis and metabolite feature classification. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, *1071*, 68–74. https://doi.org/10.1016/j.jchromb.2017.08.015

Butin, N., Bergès, C., Portais, J.-C., & Bellvert, F. (2022). An optimization method for untargeted MS-based isotopic tracing investigations of metabolism. *Metabolomics: Official Journal of the Metabolomic Society*, *18*(7), 41. https://doi.org/10.1007/s11306-022-01897-5

Canelas, A. B., ten Pierick, A., Ras, C., Seifar, R. M., van Dam, J. C., van Gulik, W. M., & Heijnen, J. J. (2009). Quantitative Evaluation of Intracellular Metabolite Extraction Techniques for Yeast Metabolomics. *Analytical Chemistry*, *81*(17), 7379–7389. https://doi.org/10.1021/ac900999t

Chazalviel, M., Frainay, C., Poupin, N., Vinson, F., Merlet, B., Gloaguen, Y., et al. (2018). MetExploreViz: web component for interactive metabolic network visualization. *Bioinformatics (Oxford, England)*, *34*(2), 312–313. https://doi.org/10.1093/bioinformatics/btx588

Chokkathukalam, A., Jankevics, A., Creek, D. J., Achcar, F., Barrett, M. P., & Breitling, R. (2013). mzMatch-ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics (Oxford, England)*, *29*(2), 281–283. https://doi.org/10.1093/bioinformatics/bts674

Chokkathukalam, A., Kim, D.-H., Barrett, M. P., Breitling, R., & Creek, D. J. (2014). Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis*, *6*(4), 511–524. https://doi.org/10.4155/bio.13.348

Christensen, B., & Nielsen, J. (2000). Metabolic network analysis. A powerful tool in metabolic engineering. *Advances in Biochemical Engineering/Biotechnology*, *66*, 209–231.

Clark, T. J., Guo, L., Morgan, J., & Schwender, J. (2020). Modeling Plant Metabolism: From Network Reconstruction to Mechanistic Models. *Annual Review of Plant Biology*, *71*, 303–326. https://doi.org/10.1146/annurev-arplant-050718-100221

Clendinen, C. S., Stupp, G. S., Ajredini, R., Lee-McMullen, B., Beecher, C., & Edison, A. S. (2015). An overview of methods using (13)C for improved compound identification in metabolomics and natural products. *Frontiers in Plant Science*, *6*, 611. https://doi.org/10.3389/fpls.2015.00611

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., et al. (2009). Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS computational biology*, *5*(8), e1000489. https://doi.org/10.1371/journal.pcbi.1000489

Cottret, L., & Jourdan, F. (2010). Graph methods for the investigation of metabolic networks in parasitology. *Parasitology*, *137*(9), 1393–1407. https://doi.org/10.1017/S0031182010000363

Creek, D. J., Chokkathukalam, A., Jankevics, A., Burgess, K. E. V., Breitling, R., & Barrett, M. P. (2012). Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. *Analytical Chemistry*, *84*(20), 8442–8447. https://doi.org/10.1021/ac3018795

Crown, S. B., Ahn, W. S., & Antoniewicz, M. R. (2012). Rational design of 13C-labeling experiments for metabolic flux analysis in mammalian cells. *BMC Systems Biology*, *6*(1), 43. https://doi.org/10.1186/1752-0509-6-43

Crown, S. B., Long, C. P., & Antoniewicz, M. R. (2015). Integrated 13C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. *Metabolic engineering*, *28*, 151–158. https://doi.org/10.1016/j.ymben.2015.01.001

Crown, S. B., Long, C. P., & Antoniewicz, M. R. (2016). Optimal tracers for parallel labeling experiments and 13C metabolic flux analysis: A new precision and synergy scoring system. *Metabolic Engineering*, *38*, 10–18. https://doi.org/10.1016/j.ymben.2016.06.001

da Luz, J. A., Hans, E., & Zeng, A.-P. (2014). Automated fast filtration and on-filter quenching improve the intracellular metabolite analysis of microorganisms. *Engineering in Life Sciences*, *14*(2), 135–142. https://doi.org/10.1002/elsc.201300099

Dange, M. C., Mishra, V., Mukherjee, B., Jaiswal, D., Merchant, M. S., Prasannan, C. B., & Wangikar, P. P. (2020). Evaluation of freely available software tools for untargeted quantification of 13C isotopic enrichment in cellular metabolome from HR-LC/MS data. *Metabolic Engineering Communications*, *10*, e00120. https://doi.org/10.1016/j.mec.2019.e00120

Dellero, Y., Heuillet, M., Marnet, N., Bellvert, F., Millard, P., & Bouchereau, A. (2020). Sink/Source Balance of Leaves Influences Amino Acid Pools and Their Associated Metabolic Fluxes in Winter Oilseed Rape (Brassica napus L.). *Metabolites*, *10*(4), 150. https://doi.org/10.3390/metabo10040150

Doerfler, H., Sun, X., Wang, L., Engelmeier, D., Lyon, D., & Weckwerth, W. (2014). mzGroupAnalyzer--predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. *PloS One*, *9*(5), e96188. https://doi.org/10.1371/journal.pone.0096188

Dong, W., Rawat, E. S., Stephanopoulos, G., & Abu-Remaileh, M. (2022). Isotope tracing in health and disease. *Current Opinion in Biotechnology*, *76*, 102739. https://doi.org/10.1016/j.copbio.2022.102739

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., et al. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(6), 1777–1782. https://doi.org/10.1073/pnas.0610772104

Durot, M., Bourguignon, P.-Y., & Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *Fems Microbiology Reviews*, *33*(1), 164–190. https://doi.org/10.1111/j.1574-6976.2008.00146.x

Edwards, J. S., & Palsson, B. O. (2000). The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5528–5533. https://doi.org/10.1073/pnas.97.10.5528

Edwards, Jeremy S, & Palsson, B. O. (2000). Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. *BMC Bioinformatics*, *1*, 1. https://doi.org/10.1186/1471-2105-1-1

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., et al. (2019). NMR Spectroscopy for Metabolomics Research. *Metabolites*, *9*(7), 123. https://doi.org/10.3390/metabo9070123

Enomoto, H., Sato, K., Miyamoto, K., Ohtsuka, A., & Yamane, H. (2018). Distribution Analysis of Anthocyanins, Sugars, and Organic Acids in Strawberry Fruits Using Matrix-Assisted Laser Desorption/Ionization-Imaging Mass Spectrometry. *Journal of Agricultural and Food Chemistry*, *66*(19), 4958–4965. https://doi.org/10.1021/acs.jafc.8b00853

Fan, T. W.-M., Lorkiewicz, P. K., Sellers, K., Moseley, H. N. B., Higashi, R. M., & Lane, A. N. (2012). Stable isotope-resolved metabolomics and applications for drug development. *Pharmacology & Therapeutics*, *133*(3), 366–391. https://doi.org/10.1016/j.pharmthera.2011.12.007

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science (New York, N.Y.)*, *246*(4926), 64–71. https://doi.org/10.1126/science.2675315

Fischer, E., & Sauer, U. (2003). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, *270*(5), 880–891. https://doi.org/10.1046/j.1432-1033.2003.03448.x

Gathungu, R. M., Larrea, P., Sniatynski, M. J., Marur, V. R., Bowden, J. A., Koelmel, J. P., et al. (2018). Optimization of Electrospray Ionization Source Parameters for Lipidomics To Reduce Misannotation of In-Source Fragments as Precursor Ions. *Analytical Chemistry*, *90*(22), 13523–13532. https://doi.org/10.1021/acs.analchem.8b03436

Giraudeau, P. (2020). NMR-based metabolomics and fluxomics: developments and future prospects. *Analyst*, *145*(7), 2457–2472. https://doi.org/10.1039/D0AN00142B

Gowda, G. A. N., & Djukovic, D. (2014). Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges. *Methods in molecular biology (Clifton, N.J.)*, *1198*, 3–12. https://doi.org/10.1007/978-1-4939-1258-2_1

Granata, I., Troiano, E., Sangiovanni, M., & Guarracino, M. R. (2019). Integration of transcriptomic data in a genome-scale metabolic model to investigate the link between obesity and breast cancer. *BMC Bioinformatics*, *20*(4), 162. https://doi.org/10.1186/s12859-019-2685-9

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*(1), 121. https://doi.org/10.1186/s13059-019-1730-3

Haggarty, J., & Burgess, K. E. (2017). Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Current Opinion in Biotechnology*, *43*, 77–85. https://doi.org/10.1016/j.copbio.2016.09.006

Hari, A., & Lobo, D. (2020). Fluxer: a web application to compute, analyze and visualize genome-scale metabolic flux networks. *Nucleic Acids Research*, *48*(W1), W427–W435. https://doi.org/10.1093/nar/gkaa409

Harrieder, E.-M., Kretschmer, F., Böcker, S., & Witting, M. (2022). Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics. *Journal of Chromatography B*, *1188*, 123069. https://doi.org/10.1016/j.jchromb.2021.123069

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., et al. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, *44*(D1), D1214-9. https://doi.org/10.1093/nar/gkv1031

Hegeman, A. D., Schulte, C. F., Cui, Q., Lewis, I. A., Huttlin, E. L., Eghbalnia, H., et al. (2007). Stable isotope assisted assignment of elemental compositions for metabolomics. *Analytical Chemistry*, *79*(18), 6912–6921. https://doi.org/10.1021/ac070346t

Heuillet, M., Bellvert, F., Cahoreau, E., Letisse, F., Millard, P., & Portais, J.-C. (2018). Methodology for the Validation of Isotopic Analyses by Mass Spectrometry in Stable-Isotope Labeling Experiments. *Analytical Chemistry*, *90*(3), 1852–1860. https://doi.org/10.1021/acs.analchem.7b03886

Heuillet, M., Millard, P., Cissé, M. Y., Linares, L. K., Létisse, F., Manié, S., et al. (2020). Simultaneous Measurement of Metabolite Concentration and Isotope Incorporation by Mass Spectrometry. *Analytical Chemistry*, *92*(8), 5890–5896. https://doi.org/10.1021/acs.analchem.9b05709

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, *19*(4), 524–531. https://doi.org/10.1093/bioinformatics/btg015

Jamshidi, N., & Palsson, B. Ø. (2006). Systems biology of SNPs. *Molecular Systems Biology*, *2*, 38. https://doi.org/10.1038/msb4100077

Jang, C., Chen, L., & Rabinowitz, J. D. (2018). Metabolomics and Isotope Tracing. *Cell*, *173*(4), 822–837. https://doi.org/10.1016/j.cell.2018.03.055

Johnson, K. A., & Goody, R. S. (2011). The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. *Biochemistry*, *50*(39), 8264–8269. https://doi.org/10.1021/bi201284u

Kaklamanos, G., Aprea, E., & Theodoridis, G. (2020). 11 - Mass spectrometry: principles and instrumentation. In Y. Pico (Ed.), *Chemical Analysis of Food (Second Edition)* (pp. 525–552). Academic Press. https://doi.org/10.1016/B978-0-12-813266-1.00011-5

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Karp, P. D., Midford, P. E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., et al. (2019). Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, *22*(1), 109–126. https://doi.org/10.1093/bib/bbz104

Karp, P. D., Midford, P. E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., et al. (2021). Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, *22*(1), 109–126. https://doi.org/10.1093/bib/bbz104

Karp, P. D., Ong, W. K., Paley, S., Billington, R., Caspi, R., Fulcher, C., et al. (2018). The EcoCyc Database. *EcoSal Plus*, *8*(1), 10.1128/ecosalplus.ESP-0006–2018. https://doi.org/10.1128/ecosalplus.ESP-0006-2018

Karp, P. D., Paley, S., & Romero, P. (2002). The Pathway Tools software. *Bioinformatics (Oxford, England)*, *18 Suppl 1*, S225-232. https://doi.org/10.1093/bioinformatics/18.suppl_1.s225

Karp, P. D., Riley, M., Paley, S. M., & Pellegrini-Toole, A. (2002). The MetaCyc Database. *Nucleic Acids Research*, *30*(1), 59–61.

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics (Oxford, England)*, *24*(21), 2534–2536. https://doi.org/10.1093/bioinformatics/btn323

Kiefer, P., Schmitt, U., & Vorholt, J. A. (2013). eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. *Bioinformatics (Oxford, England)*, *29*(7), 963–964. https://doi.org/10.1093/bioinformatics/btt080

Kim, I.-Y., Suh, S.-H., Lee, I.-K., & Wolfe, R. R. (2016). Applications of stable, nonradioactive isotope tracers in in vivo human metabolic research. *Experimental & Molecular Medicine*, *48*(1), e203. https://doi.org/10.1038/emm.2015.97

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2023). PubChem 2023 update. *Nucleic Acids Research*, *51*(D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J., & Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology*, *23*(4), 617–623. https://doi.org/10.1016/j.copbio.2011.10.007

Kim, W. J., Kim, H. U., & Lee, S. Y. (2017). Current state and applications of microbial genome-scale metabolic models. *Current Opinion in Systems Biology*, *2*, 10–18. https://doi.org/10.1016/j.coisb.2017.03.001

Kind, T., & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, *8*, 105. https://doi.org/10.1186/1471-2105-8-105

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., & Palsson, B. O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLOS Computational Biology*, *11*(8), e1004321. https://doi.org/10.1371/journal.pcbi.1004321

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., et al. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515–D522. https://doi.org/10.1093/nar/gkv1049

Klein, S., & Heinzle, E. (2012). Isotope labeling experiments in metabolomics and fluxomics. *WIREs Systems Biology and Medicine*, *4*(3), 261–272. https://doi.org/10.1002/wsbm.1167

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, *5*(1), 21. https://doi.org/10.1186/1752-0509-5-21

Kubryk, P., Kölschbach, J. S., Marozava, S., Lueders, T., Meckenstock, R. U., Niessner, R., & Ivleva, N. P. (2015). Exploring the Potential of Stable Isotope (Resonance) Raman Microspectroscopy and Surface-Enhanced Raman Scattering for the Analysis of Microorganisms at Single Cell Level. *Analytical Chemistry*, *87*(13), 6622–6630. https://doi.org/10.1021/acs.analchem.5b00673

Laber, S., Forcisi, S., Bentley, L., Petzold, J., Moritz, F., Smirnov, K. S., et al. (2021). Linking the FTO obesity rs1421085 variant circuitry to cellular, metabolic, and organismal phenotypes in vivo. *Science Advances*, *7*(30), eabg0108. https://doi.org/10.1126/sciadv.abg0108

Lacroix, V., Cottret, L., Thébault, P., & Sagot, M.-F. (2008). An introduction to metabolic networks and their structural analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, *5*(4), 594–617. https://doi.org/10.1109/TCBB.2008.79

Lane, A. N., & Fan, T. W.-M. (2017). NMR-Based Stable Isotope Resolved Metabolomics in Systems Biochemistry. *Archives of biochemistry and biophysics*, *628*, 123–131. https://doi.org/10.1016/j.abb.2017.02.009

Lee, D.-K., Na, E., Park, S., Park, J., Lim, J., & Kwon, S. (2018). In Vitro Tracking of Intracellular Metabolism-Derived Cancer Volatiles via Isotope Labeling. *ACS Central Science*, *4*. https://doi.org/10.1021/acscentsci.8b00296

Lee, W.-N. P., Wahjudi, P. N., Xu, J., & Go, V. L. (2010). Tracer-based Metabolomics: Concepts and Practices. *Clinical biochemistry*, *43*(16–17), 1269–1277. https://doi.org/10.1016/j.clinbiochem.2010.07.027

Leighty, R. W., & Antoniewicz, M. R. (2012). Parallel labeling experiments with [U-13C]glucose validate E. coli metabolic network model for 13C metabolic flux analysis. *Metabolic Engineering*, *14*(5), 533–541. https://doi.org/10.1016/j.ymben.2012.06.003

Leighty, R. W., & Antoniewicz, M. R. (2013). COMPLETE-MFA: Complementary parallel labeling experiments technique for metabolic flux analysis. *Metabolic Engineering*, *20*, 49–55. https://doi.org/10.1016/j.ymben.2013.08.006

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews. Microbiology*, *10*(4), 291–305. https://doi.org/10.1038/nrmicro2737

Liu, Y., Forcisi, S., Harir, M., Deleris-Bou, M., Krieger-Weber, S., Lucio, M., et al. (2016). New molecular evidence of wine yeast-bacteria interaction unraveled by non-targeted exometabolomic profiling. *Metabolomics*, *12*(4), 69. https://doi.org/10.1007/s11306-016-1001-1

Long, C. P., & Antoniewicz, M. R. (2019). High-resolution 13C metabolic flux analysis. *Nature Protocols*, *14*(10), 2856–2877. https://doi.org/10.1038/s41596-019-0204-0

Long, C. P., Au, J., Sandoval, N. R., Gebreselassie, N. A., & Antoniewicz, M. R. (2017). Enzyme I facilitates reverse flux from pyruvate to phosphoenolpyruvate in Escherichia coli. *Nature Communications*, *8*, 14316. https://doi.org/10.1038/ncomms14316

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., & Rocha, I. (2011). Modeling formalisms in Systems Biology. *AMB Express*, *1*, 45. https://doi.org/10.1186/2191-0855-1-45

Machado, D., & Herrgård, M. (2014). Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLOS Computational Biology*, *10*(4), e1003580. https://doi.org/10.1371/journal.pcbi.1003580

Markley, J. L., Brüschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., & Wishart, D. S. (2017). The future of NMR-based metabolomics. *Current Opinion in Biotechnology*, *43*, 34–40. https://doi.org/10.1016/j.copbio.2016.08.001

Marshall, A. G., & Hendrickson, C. L. (2008). High-resolution mass spectrometers. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)*, *1*, 579–599. https://doi.org/10.1146/annurev.anchem.1.031207.112945

Martín-Gómez, P., Barbeta, A., Voltas, J., Peñuelas, J., Dennis, K., Palacio, S., et al. (2015). Isotope-ratio infrared spectroscopy: a reliable tool for the investigation of plant-water sources? *New Phytologist*, *207*(3), 914–927. https://doi.org/10.1111/nph.13376

Mashego, M. R., Wu, L., Van Dam, J. C., Ras, C., Vinke, J. L., Van Winden, W. A., et al. (2004). MIRACLE: mass isotopomer ratio analysis of U-13C-labeled extracts. A new method for accurate

quantification of changes in concentrations of intracellular metabolites. *Biotechnology and Bioengineering*, *85*(6), 620–628. https://doi.org/10.1002/bit.10907

Mendoza, S. N., Olivier, B. G., Molenaar, D., & Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, *20*(1), 158. https://doi.org/10.1186/s13059-019-1769-1

Metallo, C. M., Gameiro, P. A., Bell, E. L., Mattaini, K. R., Yang, J., Hiller, K., et al. (2011). Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature*, *481*(7381), 380–384. https://doi.org/10.1038/nature10602

Metallo, C. M., Walther, J. L., & Stephanopoulos, G. (2009). Evaluation of 13C isotopic tracers for metabolic flux analysis in mammalian cells. *Journal of Biotechnology*, *144*(3), 167–174. https://doi.org/10.1016/j.jbiotec.2009.07.010

Midani, F. S., Wynn, M. L., & Schnell, S. (2017). The importance of accurately correcting for the natural abundance of stable isotopes. *Analytical Biochemistry*, *520*, 27–43. https://doi.org/10.1016/j.ab.2016.12.011

Millard, P., Letisse, F., Sokol, S., & Portais, J.-C. (2012). IsoCor: correcting MS data in isotope labeling experiments. *Bioinformatics (Oxford, England)*, *28*(9), 1294–1296. https://doi.org/10.1093/bioinformatics/bts127

Millard, P., Massou, S., Portais, J.-C., & Létisse, F. (2014). Isotopic studies of metabolic systems by mass spectrometry: using Pascal's triangle to produce biological standards with fully controlled labeling patterns. *Analytical Chemistry*, *86*(20), 10288–10295. https://doi.org/10.1021/ac502490g

Millard, P., Massou, S., Wittmann, C., Portais, J.-C., & Létisse, F. (2014). Sampling of intracellular metabolites for stationary and non-stationary 13C metabolic flux analysis in Escherichia coli. *Analytical Biochemistry*, *465*, 38–49. https://doi.org/10.1016/j.ab.2014.07.026

Millard, P., Schmitt, U., Kiefer, P., Vorholt, J. A., Heux, S., & Portais, J.-C. (2020). ScalaFlux: A scalable approach to quantify fluxes in metabolic subnetworks. *PLoS computational biology*, *16*(4), e1007799. https://doi.org/10.1371/journal.pcbi.1007799

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., et al. (2013). Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20338–20343. https://doi.org/10.1073/pnas.1307797110

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., et al. (2017). iML1515, a knowledgebase that computes Escherichia coli traits. *Nature biotechnology*, *35*(10), 904–908. https://doi.org/10.1038/nbt.3956

Munger, J., Bennett, B. D., Parikh, A., Feng, X.-J., McArdle, J., Rabitz, H. A., et al. (2008). Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy. *Nature Biotechnology*, *26*(10), 1179–1186. https://doi.org/10.1038/nbt.1500

Murphy, T. A., Dang, C. V., & Young, J. D. (2013). Isotopically nonstationary 13C flux analysis of Myc-induced metabolic reprogramming in B-cells. *Metabolic Engineering*, *15*, 206–217. https://doi.org/10.1016/j.ymben.2012.07.008

Naake, T., & Fernie, A. R. (2019). MetNet: Metabolite Network Prediction from High-Resolution Mass Spectrometry Data in R Aiding Metabolite Annotation. *Analytical Chemistry*, *91*(3), 1768–1772. https://doi.org/10.1021/acs.analchem.8b04096

Nakahigashi, K., Toya, Y., Ishii, N., Soga, T., Hasegawa, M., Watanabe, H., et al. (2009). Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Molecular Systems Biology*, *5*, 306. https://doi.org/10.1038/msb.2009.65

Neumann, N. K. N., Lehner, S. M., Kluger, B., Bueschl, C., Sedelmaier, K., Lemmens, M., et al. (2014). Automated LC-HRMS(/MS) approach for the annotation of fragment ions derived from stable isotope labeling-assisted untargeted metabolomics. *Analytical Chemistry*, *86*(15), 7320–7327. https://doi.org/10.1021/ac501358z

Nicholson, J. K., & Wilson, I. D. (1989). High resolution proton magnetic resonance spectroscopy of biological fluids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *21*(4), 449–501. https://doi.org/10.1016/0079-6565(89)80008-1

Nöh, K., & Wiechert, W. (2011). The benefits of being transient: isotope-based metabolic flux analysis at the short time scale. *Applied Microbiology and Biotechnology*, *91*(5), 1247–1265. https://doi.org/10.1007/s00253-011-3390-4

Noothalapati, H., & Shigeto, S. (2014). Exploring metabolic pathways in vivo by a combined approach of mixed stable isotope-labeled Raman microspectroscopy and multivariate curve resolution analysis. *Analytical Chemistry*, *86*(15), 7828–7834. https://doi.org/10.1021/ac501735c

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Park, Y.-C., Jun, S. Y., & Seo, J.-H. (2012). Construction and characterization of recombinant Bacillus subtilis JY123 able to transport xylose efficiently. *Journal of Biotechnology*, *161*(4), 402–406. https://doi.org/10.1016/j.jbiotec.2012.07.192

Peiro, C., Millard, P., de Simone, A., Cahoreau, E., Peyriga, L., Enjalbert, B., & Heux, S. (2019). Chemical and Metabolic Controls on Dihydroxyacetone Metabolism Lead to Suboptimal Growth of Escherichia coli. *Applied and Environmental Microbiology*, *85*(15), e00768-19. https://doi.org/10.1128/AEM.00768-19

Peyraud, R., Kiefer, P., Christen, P., Massou, S., Portais, J.-C., & Vorholt, J. A. (2009). Demonstration of the ethylmalonyl-CoA pathway by using 13C metabolomics. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(12), 4846–4851. https://doi.org/10.1073/pnas.0810932106

Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., et al. (2020). Implementation of liquid chromatography–high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: A tutorial. *Analytica Chimica Acta*, *1105*, 28–44. https://doi.org/10.1016/j.aca.2019.12.062

Quek, L.-E., Krycer, J. R., Ohno, S., Yugi, K., Fazakerley, D. J., Scalzo, R., et al. (2020). Dynamic 13C Flux Analysis Captures the Reorganization of Adipocyte Glucose Metabolism in Response to Insulin. *iScience*, *23*(2), 100855. https://doi.org/10.1016/j.isci.2020.100855

Quek, L.-E., Wittmann, C., Nielsen, L. K., & Krömer, J. O. (2009). OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis. *Microbial Cell Factories*, *8*(1), 25. https://doi.org/10.1186/1475-2859-8-25

Ramon, C., Gollub, M. G., & Stelling, J. (2018). Integrating -omics data into genome-scale metabolic network models: principles and challenges. *Essays in Biochemistry*, *62*(4), 563–574. https://doi.org/10.1042/EBC20180011

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., et al. (2006). Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(46), 17480–17484. https://doi.org/10.1073/pnas.0603364103

Rosato, A., Tenori, L., Cascante, M., De Atauri Carulla, P. R., Martins dos Santos, V. A. P., & Saccenti, E. (2018). From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics*, *14*(4), 37. https://doi.org/10.1007/s11306-018-1335-y

Rose, T. D., & Mazat, J.-P. (2018). FluxVisualizer, a Software to Visualize Fluxes through Metabolic Networks. *Processes*, *6*(5), 39. https://doi.org/10.3390/pr6050039

Ryu, J. Y., Kim, H. U., & Lee, S. Y. (2015). Reconstruction of genome-scale human metabolic models using omics data. *Integrative Biology: Quantitative Biosciences from Nano to Macro*, *7*(8), 859–868. https://doi.org/10.1039/c5ib00002e

Satish Kumar, V., Dasika, M. S., & Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC bioinformatics*, *8*, 212. https://doi.org/10.1186/1471-2105-8-212

Schoenheimer, R., & Rittenberg, D. (1935). Deuterium as an Indicator in the Study of Intermediary Metabolism. *Science*, *82*(2120), 156–157. https://doi.org/10.1126/science.82.2120.156

Schuster, S., Dandekar, T., & Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, *17*(2), 53–60. https://doi.org/10.1016/S0167-7799(98)01290-6

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sherwood, C. A., Eastham, A., Lee, L. W., Risler, J., Mirzaei, H., Falkner, J. A., & Martin, D. B. (2009). Rapid Optimization of MRM-MS Instrument Parameters by Subtle Alteration of Precursor and Product m/z Targets. *Journal of proteome research*, *8*(7), 3746–3751. https://doi.org/10.1021/pr801122b

Sokol, S., Millard, P., & Portais, J.-C. (2012). influx_s: increasing numerical stability and precision for metabolic flux analysis in isotope labelling experiments. *Bioinformatics*, *28*(5), 687–693. https://doi.org/10.1093/bioinformatics/btr716

Stincone, A., Prigione, A., Cramer, T., Wamelink, M. M. C., Campbell, K., Cheung, E., et al. (2015). The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. *Biological Reviews of the Cambridge Philosophical Society*, *90*(3), 927–963. https://doi.org/10.1111/brv.12140

Sutehall, S., Muniz-Pardos, B., Smajgl, D., Mandic, M., Jeglinski, C., Bosch, A., et al. (2021). The validity and reliability of a novel isotope ratio infrared spectrometer to quantify 13C enrichment of expired breath samples in exercise. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *130*(5), 1421–1426. https://doi.org/10.1152/japplphysiol.00805.2020

Szyperski, T., Glaser, R. W., Hochuli, M., Fiaux, J., Sauer, U., Bailey, J. E., & Wüthrich, K. (1999). Bioreaction network topology and metabolic flux ratio analysis by biosynthetic fractional 13C labeling and two-dimensional NMR spectroscopy. *Metabolic Engineering*, *1*(3), 189–197. https://doi.org/10.1006/mben.1999.0116

Takáts, Z., Wiseman, J. M., & Cooks, R. G. (2005). Ambient mass spectrometry using desorption electrospray ionization (DESI): instrumentation, mechanisms and applications in forensics,

chemistry, and biology. *Journal of mass spectrometry: JMS*, *40*(10), 1261–1275. https://doi.org/10.1002/jms.922

Tatsukami, Y., Nambu, M., Morisaka, H., Kuroda, K., & Ueda, M. (2013). Disclosure of the differences of Mesorhizobium loti under the free-living and symbiotic conditions by comparative proteome analysis without bacteroid isolation. *BMC microbiology*, *13*, 180. https://doi.org/10.1186/1471-2180-13-180

Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, *84*(11), 5035–5039. https://doi.org/10.1021/ac300698c

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Tian, M., & Reed, J. L. (2018). Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. *Bioinformatics*, *34*(22), 3882–3888. https://doi.org/10.1093/bioinformatics/bty445

Toechterle, P., Dublyansky, Y., Stöbener, N., Mandi, M., & Spötl, C. (2017). High Resolution Isotopic Monitoring of Cave Air CO2. *Rapid communications in mass spectrometry : RCM*, *31*. https://doi.org/10.1002/rcm.7859

Traquete, F., Luz, J., Cordeiro, C., Sousa Silva, M., & Ferreira, A. E. N. (2022). Graph Properties of Mass-Difference Networks for Profiling and Discrimination in Untargeted Metabolomics. *Frontiers in Molecular Biosciences*, *9*, 917911. https://doi.org/10.3389/fmolb.2022.917911

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., et al. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, *12*(6), 523–526. https://doi.org/10.1038/nmeth.3393

Ursem, R., Tikunov, Y., Bovy, A., van Berloo, R., & van Eeuwijk, F. (2008). A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica*, *161*(1), 181–193. https://doi.org/10.1007/s10681-008-9672-y

Vacanti, N. M., Divakaruni, A. S., Green, C. R., Parker, S. J., Henry, R. R., Ciaraldi, T. P., et al. (2014). Regulation of substrate utilization by the mitochondrial pyruvate carrier. *Molecular Cell*, *56*(3), 425–435. https://doi.org/10.1016/j.molcel.2014.09.024

Vasilakou, E., Machado, D., Theorell, A., Rocha, I., Nöh, K., Oldiges, M., & Wahl, S. A. (2016). Current state and challenges for dynamic metabolic modeling. *Current Opinion in Microbiology*, *33*, 97–104. https://doi.org/10.1016/j.mib.2016.07.008

Violante, S., Berisa, M., Thomas, T. H., & Cross, J. R. (2019). Stable Isotope Tracers for Metabolic Pathway Analysis. *Methods in Molecular Biology (Clifton, N.J.)*, *1978*, 269–283. https://doi.org/10.1007/978-1-4939-9236-2_17

Wang, C., Wang, H., Li, Y., & Liu, B. (2012). Identification of a fructose-1,6-bisphosphate aldolase gene and association of the single nucleotide polymorphisms with growth traits in the clam Meretrix meretrix. *Molecular Biology Reports*, *39*(4), 5017–5024. https://doi.org/10.1007/s11033-011-1298-9

Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., et al. (2012). Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(26), E1743-1752. https://doi.org/10.1073/pnas.1203689109

Weindl, D., Wegner, A., & Hiller, K. (2015). Metabolome-Wide Analysis of Stable Isotope Labeling-Is It Worth the Effort? *Frontiers in Physiology*, *6*, 344. https://doi.org/10.3389/fphys.2015.00344

Weindl, D., Wegner, A., Jäger, C., & Hiller, K. (2015). Isotopologue ratio normalization for non-targeted metabolomics. *Journal of Chromatography A*, *1389*, 112–119. https://doi.org/10.1016/j.chroma.2015.02.025

Weitzel, M., Nöh, K., Dalman, T., Niedenführ, S., Stute, B., & Wiechert, W. (2013). 13CFLUX2—high-performance software suite for 13C-metabolic flux analysis. *Bioinformatics*, *29*(1), 143–145. https://doi.org/10.1093/bioinformatics/bts646

Wiechert, W. (2001). 13C metabolic flux analysis. *Metabolic Engineering*, *3*(3), 195–206. https://doi.org/10.1006/mben.2001.0187

Wiechert, W., Möllney, M., Petersen, S., & de Graaf, A. A. (2001). A universal framework for 13C metabolic flux analysis. *Metabolic Engineering*, *3*(3), 265–283. https://doi.org/10.1006/mben.2001.0188

Wiechert, Wolfgang, & Nöh, K. (2005). From stationary to instationary metabolic flux analysis. *Advances in Biochemical Engineering/Biotechnology*, *92*, 145–172. https://doi.org/10.1007/b98921

Wilkinson, D. J. (2018). Historical and contemporary stable isotope tracer approaches to studying mammalian protein metabolism. *Mass Spectrometry Reviews*, *37*(1), 57–80. https://doi.org/10.1002/mas.21507

Winder, C. L., Dunn, W. B., Schuler, S., Broadhurst, D., Jarvis, R., Stephens, G. M., & Goodacre, R. (2008). Global Metabolic Profiling of Escherichia coli Cultures: an Evaluation of Methods for Quenching and Extraction of Intracellular Metabolites. *Analytical Chemistry*, *80*(8), 2939–2948. https://doi.org/10.1021/ac7023409

Wishart, D. S. (2011). Advances in metabolite identification. *Bioanalysis*, *3*(15), 1769–1782. https://doi.org/10.4155/bio.11.155

Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., et al. (2022). HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research*, *50*(D1), D622–D631. https://doi.org/10.1093/nar/gkab1062

Wu, L., Mashego, M. R., van Dam, J. C., Proell, A. M., Vinke, J. L., Ras, C., et al. (2005). Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards. *Analytical Biochemistry*, *336*(2), 164–171. https://doi.org/10.1016/j.ab.2004.09.001

Wunderlich, Z., & Mirny, L. A. (2006). Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains. *Biophysical Journal*, *91*(6), 2304–2311. https://doi.org/10.1529/biophysj.105.080572

Ye, C., Wei, X., Shi, T., Sun, X., Xu, N., Gao, C., & Zou, W. (2022). Genome-scale metabolic network models: from first-generation to next-generation. *Applied Microbiology and Biotechnology*, *106*(13), 4907–4920. https://doi.org/10.1007/s00253-022-12066-y

Yoo, H., Antoniewicz, M. R., Stephanopoulos, G., & Kelleher, J. K. (2008). Quantifying reductive carboxylation flux of glutamine to lipid in a brown adipocyte cell line. *The Journal of Biological Chemistry*, *283*(30), 20621–20627. https://doi.org/10.1074/jbc.M706494200

You, L., Zhang, B., & Tang, Y. J. (2014). Application of Stable Isotope-Assisted Metabolomics for Cell Metabolism Studies. *Metabolites*, *4*(2), 142–165. https://doi.org/10.3390/metabo4020142

Young, J. D. (2014). INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics*, *30*(9), 1333–1335. https://doi.org/10.1093/bioinformatics/btu015

Zamboni, N., Fendt, S.-M., Rühl, M., & Sauer, U. (2009). 13C-based metabolic flux analysis. *Nature Protocols*, *4*(6), 878–892. https://doi.org/10.1038/nprot.2009.58

Zamboni, N., Saghatelian, A., & Patti, G. J. (2015). Defining the metabolome: size, flux, and regulation. *Molecular Cell*, *58*(4), 699–706. https://doi.org/10.1016/j.molcel.2015.04.021

Zhang, J., Ahn, W. S., Gameiro, P. A., Keibler, M. A., Zhang, Z., & Stephanopoulos, G. (2014). 13C Isotope-Assisted Methods for Quantifying Glutamine Metabolism in Cancer Cells. *Methods in enzymology*, *542*, 369–389. https://doi.org/10.1016/B978-0-12-416618-9.00019-4

Zhou, J., & Yin, Y. (2016). Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. *Analyst*, *141*(23), 6362–6373. https://doi.org/10.1039/C6AN01753C

Zhou, T., Wu, T., Wu, Q., Chen, W., Wu, M., Ye, C., & He, X. (2020). Real-Time Monitoring of 13C- and 18O-Isotopes of Human Breath CO2 Using a Mid-Infrared Hollow Waveguide Gas Sensor. *Analytical Chemistry*, *92*(19), 12943–12949. https://doi.org/10.1021/acs.analchem.0c01586

Chapitre 2

An optimization method for untargeted MS-based isotopic tracing investigations of metabolism.

Noémie Butin[1,2,3], Cécilia Bergès[2,3], Jean-Charles Portais[1,2,3], Floriant Bellvert[2,3]

[1] RESTORE, CNRS ERL5311, EFs, ENVT, Inserm U1031, UPS, Université de Toulouse, Toulouse, France

[2] Toulouse Biotechnology Institute, TBI-INSA de Toulouse INSA/CNRS 5504-UMR INSA/INRA 798, 5504 Toulouse, France

[3] MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, 31077 Toulouse, France

Metabolomics, 2022

Glossary

Benchmark clusters: Dataset containing the benchmark isotopologues automatically clustered using the two-clustering software.

Benchmark isotopologues: Dataset containing isotopologues of *reference metabolites* automatically extracted from MS data of the reference material using XCMS.

Isotopic cluster: Group of MS peaks from a unique molecular entity, i.e. with the same elemental composition but different isotopic compositions (IUPAC definition).

Isotopologues: Molecular entities that differ only in their isotopic composition (IUPAC definition).

Pascal triangle (PT) sample: biologically-produced material in which the isotopic composition of the labelled substrate is designed to obtain metabolites with tracer isotopologues distributed according to the binomial coefficients of Pascal's triangle.

Reference clusters: Reference dataset containing the benchmark isotopologues manually clustered.

Reference isotopologues: Reference dataset containing isotopologues of *reference metabolites* manually extracted from MS data of the reference material.

Reference material: Labelled sample used as a reference to optimize processing parameters.

Reference metabolites: List of metabolites identified with a level 1 confidence expressed and measurable in the reference material.

Tracer isotopologues: Isotopologues of the tracer element.

Abtract

Introduction: Stable isotope tracer studies are increasingly applied to explore metabolism from the detailed analysis of tracer incorporation into metabolites. Untargeted LC/MS approaches have recently emerged and provide potent methods for expanding the dimension and complexity of the metabolic networks that can be investigated. A number of software tools have been developed to process the highly complex MS data collected in such studies; however, a method to optimize the extraction of valuable isotopic data is lacking.

Objectives: To develop and validate a method to optimize automated data processing for untargeted MS-based isotopic tracing investigations of metabolism.

Methods: The method is based on the application of a suitable reference material to rationally perform parameter optimization throughout the complete data processing workflow. It was applied in the context of ^{13}C-labelling experiments and with two different software, namely geoRge and X13CMS. It was illustrated with the study of a *E. coli* mutant impaired for central metabolism.

Results: The optimization methodology provided significant gain in the number and quality of extracted isotopic data, independently of the software considered. Pascal Triangle samples are well suited for such purpose since they allow both the identification of analytical issues and optimization of data processing at the same time.

Conclusion: The proposed method maximizes the biological value of untargeted MS-based isotopic tracing investigations by revealing the full metabolic information that is encoded in the labelling patterns of metabolites.

Keywords: Isotope labelling experiments, untargeted analysis, parameter optimization, LC/MS

1. Introduction

Stable-isotope labelling experiments coupled with mass spectrometry (MS) are increasingly used to obtain a comprehensive understanding of metabolism in many fields of biology, biotechnology, and medicine (Wittman, 2002, Chokkathukalam et al, 2014, Zaimenko et al; 2017). In such investigations, an isotope tracer (most commonly ^{13}C in metabolic studies) is fed to a biological system of interest (cells, tissues, whole organisms). The incorporation of the labelled atom into metabolites is measured by MS and provides valuable information on metabolic pathways (pathway profiling) and metabolic fluxes (fluxomics) (Wittman, 2002; Wiechert et al, 2001; Wiechert, 2001; Zamboni et al, 2009). These approaches were initially developed by exploiting targeted MS methods in which the labelling patterns of selected metabolites – hence of selected metabolic pathways – could be measured (Chokkathukalam et al, 2014, Stuani et al; 2018). Progress in MS instrumentation and methods has led to the recent emergence of untargeted approaches with the potential to access the labelling patterns of a much larger number of metabolites, resulting in significant gains in the coverage of cellular and tissular metabolic processes (Creek et al, 2012; Zamboni et al, 2015). Similar to untargeted metabolomics, which aims at maximizing the number of detected metabolites, untargeted isotopic profiling aims at maximizing the number of isotopic data – i.e. the number of measured isotopologue abundances – collected from isotopically labelled material using appropriate analytical methods and data processing tools (Hiller et al, 2010; Chokkathukalam et al, 2012; Bueschl et al, 2014, Kluger et al, 2014; Capellades et al, 2016, Weindl et al, 2016). Data processing in untargeted isotopic tracing studies is a real challenge, firstly because the MS spectra collected on labelled material are much more complex than those of (the same) unlabelled material. Potentially all the isotopologues of each metabolite in the labelled samples can be generated and detected. Given the high molecular complexity of typical biological samples, the total number of peaks in the MS spectra is drastically increased. Moreover, since the total intensity of the MS signal from a given analyte is the same whether a compound is labelled or not, the MS spectra of labelled compounds contain more signals each with lower intensities than in the corresponding unlabelled spectra. The MS spectra of labelled material therefore contain more peaks with lower intensities than those of equivalent unlabelled samples.

The untargeted processing of MS data from labelled material is also more complex. The extraction of isotopologues from the raw MS data is basically the same process as in unlabelled metabolomics so that the same tools – such as XCMS (Kessner et al, 2008), MS-Dial (Tsugawa et al, 2015), MZmine 2 (Pluskal et al, 2010) – can be used in both cases. However, the task of regrouping isotopologues into isotopic clusters is specific to isotopic studies. A number of dedicated software tools have been developed, such as X13CMS (Patti et al, 2014), geoRge (Capellades et al, 2016), MetExtractII (Buelschl et al, 2017), mzMatchIso (Chokkathukalam et al, 2012), DynaMet (Kiefer et al, 2015) and HiResTec (Hoffmann et al, 2018). Considering the wealth of information to be exploited in untargeted isotopic studies, the processing software needs to be robust and efficient in maximizing the number and quality of the extracted data. Comparisons of these programs (Capellades et al, 2016; Dange et al, 2020) have highlighted the differences in requirements, algorithms, and parameter optimizations between the different tools, as well as inconsistencies (non-detection of known peaks, inconsistent isotopic clusters, abnormal redundancy, etc) in the results obtained. This can be explained in part by the newness of these programs, which will likely be improved in the near future. It can also be explained by the challenge that parameter optimization represents in such a complex, multi-step data processing workflow. Indeed, no rational strategy to optimize the recovery of all the available information in raw MS data has yet been proposed.

In this article, we present a method for optimizing MS-based untargeted isotopic tracing experiments by maximizing the amount and quality of the isotopic information that can be extracted from the analytical data. This method is based on the use of a suitable reference material to rationally perform parameter optimization throughout the processing workflow. It is applied here for ^{13}C-labelling experiments analysed with geoRge and X13CMS, but the approach is generic and can be used with any similar program or labelling strategy. We demonstrate it here for the study of a well-described *E.coli* mutant with altered metabolic fluxes.

2. Experimental section

2.1. Preparation of biological samples

2.1.1. Reference material: The Pascal triangle sample

The 'Pascal Triangle' (PT) sample was produced biologically as described by Millard et al. (2014). Briefly, *Escherichia coli* K-12 MG1655 was grown in minimal medium with a mixture of unlabelled + ^{13}C-labelled acetate as sole carbon source. This mixture consisted of the four different (carbon) isotopic forms of acetate in equal proportions, i.e., 25% of U-^{12}C-acetate, 25% of 1-^{13}C-acetate, 25% of 2-^{13}C-acetate, and 25% of U-^{13}C-acetate. The actual isotopic composition of this mixture was controlled by quantitative ^{1}H NMR before use. A similar culture was performed with only unlabelled acetate to produce the unlabelled PT sample. Cells were grown in a 500 mL Multifors Bioreactor (Infors HT, Bottmingen-Basel, Switzerland) under pH control (pH=7.0). Cell growth was monitored by measuring the optical density at 600 nm with a Genesys 6 spectrophotometer (Thermo, Carlsbad, CA, USA). Intracellular metabolites were sampled by fast filtration (Kiefer et al. 2007; Millard et al. 2014) from cells collected in the mid-exponential growth phase. Samples (2 mL) of cell culture were rapidly dropped on a filter (Sartolon Polyamide 0.2µm) to eliminate the culture medium. The filter was rinsed with 2 mL of washing solution (NaCl 0.9% with 5mM of acetate), quickly removed from the filtration unit, then placed in a precooled centrifuge tube containing 5 mL of ACN/MeOH/H2Omq (2/2/1) with 125 mM formic acid for metabolite extraction and incubated for 20 min at −20°C. The tubes were then centrifuged for 5 min at 2000 g and the supernatant was evaporated (Savant SC250 EXP 230 Speedvac, ThermoFisher) and resuspended in 100 µL of water before LC-MS analysis.

2.1.2. E. coli samples

Two *E.coli* BW-25113 strains from the Keio collection (Baba et al. 2006) were used: BW25113 wild type, and BW25113 Δ*zwf*. Both strains were first cultured in LB medium (10 g/L tryptone, 5 g/L yeast extract and 10 g/L NaCl) with kanamycine (25 ug/ml) at 37°C overnight and then stored in glycerol stock. The strains were then inoculated from a glycerol stock and first cultured in 48-well microplates in liquid LB medium. The LB cultures were used to inoculate preculture cells in 48-well microplates in minimal synthetic medium containing 17.4 g/L Na_2HPO_4,12$H_2$0, 3.03 g/L of KH_2PO_4, 0.51 g/L NaCl, 2.04 g/L NH_4Cl, 0.49 g/L $MgSO_4$, 4.38 mg/L $CaCl_2$, 15 mg/L Na_2EDTA $2H_2O$, 4.5 mg/L $ZnSO_4$ $7H_2O$, 0.3 mg/L $CoCl_2$

17.6H_2O, 1 mg/L $MnCl_2$ 4H_2O, 1 mg/L of H_3BO_3, 0.4 mg/L Na_2MoO_4 2H_2O, 3 mg/L $FeSO_4$ 7H_2O, 0.3 mg/L $CuSO_4$ 5H_2O, 0.1 g/L thiamine and 3 g/L glucose. The M9 precultures were used to inoculate cells grown in minimal medium containing 3.48 g/L Na_2HPO_4,12$H_2$0, 0.606 g/L KH_2PO_4, 0.51 g/L NaCl, 2.04 g/L NH_4Cl, 0.098 g/L $MgSO_4$, 4.38 mg/L $CaCl_2$, 15 mg/L Na2EDTA 2H_2O, 4.5 mg/L $ZnSO_4$ 7H_2O, 0.3 mg/L $CoCl_2$ 17. 6H_2O, 1 mg/L $MnCl_2$ 4H_2O, 1 mg/L H_3BO_3, 0.4 mg/L Na_2MoO_4 2H_2O, 3 mg/L $FeSO_4$ 7H_2O, 0.3 mg/L $CuSO_4$ 5H_2O, 0.1 g/L thiamine and 3 g/L glucose. These cultures were performed in 48 15 ml bioreactors under controlled growth conditions using a robotic platform (Freedom EVO 200, Tecan), with collection of labelled samples (biomass or cultivation medium) at defined culture times or optical densities. This cell culture robot and its operation are described in detail in Heux et al. (2014) and Bergès et al. (2021). The cultures were carried out with either unlabelled glucose or a mixture of 80% [1-^{13}C1]-D-glucose + 20% [U-^{13}C6]-D-glucose. To minimize sources of unlabelled carbon atoms from the first culture steps in the latter experiments, cells were inoculated at a starting OD of between 0.04 and 0.076 from pre-cultures grown with the same medium and the same (unlabelled or labelled) carbon sources as the cultures. All cultures were performed in 15 mL reaction vessels, at 37 °C, pH 7, a stirring speed of 2300 rpm and with 5 L/min of compressed air flowing through the culture module. Intracellular metabolites were automatically sampled in each bioreactor when OD_{600nm}=1.2 was reached. Samples (200 µL) were extracted and quenched in 2 mL of acetonitrile/methanol/water (4/4/2) with 125 mM formic acid at −20°C. These 2 mL were then evaporated in a SpeedVac and resuspend in 200 µL of water before LC-HRMS analysis. All samples were produced in five replicates.

2.2.LC/MS measurements

LC/MS analyses were performed using an ICS5000+ ion chromatography system (Dionex, CA, US) coupled to an Orbitrap Q Exactive+ mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) operated in negative electrospray ionization (ESI−) mode. Central metabolites were separated on an ionic chromatography column IonPac AS11 (250 × 2 mm i.d.; Dionex, CA, USA). The mobile phase was a KOH gradient at a flow rate of 380 µL/min, varied as follows: 0 min, 0.5 mM; 1 min, 0.5 mM; 9.5 min, 4.1 mM; 14.6 min, 4.1 mM; 24 min, 9.65 mM; 31.1 min, 90 mM; and 43 min, 90 mM. The column was then equilibrated for 5 min at the initial conditions before the next sample was analysed. The injection volume was 15 µL.

MS analyses were performed in full-scan mode at a resolution of 70 000 (at 400 m/z) over the m/z range 80 – 1000. Data were acquired with the following source parameters: the capillary temperature was 350°C, the source heater temperature, 350 °C, the sheath gas flow rate, 50 a.u. (arbitrary units), the auxiliary gas flow rate, 10 a.u., the S-Lens RF level, 65 %, and the source voltage, 2.75 kV.

The data were acquired in a single analytical batch. As in untargeted metabolomics approaches, all the biological samples were randomized in the analytical run and the five-replicates of the reference sample were injected at regular intervals throughout the experiment. Raw LC/MS data were converted into the open "mzXML" format using the software Proteowizard (Kessner et al, 2008). The raw data were cut after 42 min to retain all essential information while avoiding artefacts from the cleaning step and reducing data size. Figure S-1 shows the Graphical User Interface of MSConvert.

2.3.Data processing

2.3.1.Reference data

Twenty-five metabolites covering representative metabolite classes were selected as reference metabolites: organic acids (fumarate, succinate, malate, orotate, alpha-ketoglutarate (α-KG), citrate), phosphorylated compounds (2 and 3-phosphoglycerate (2/3-PG), phosphoenolpyruvate (PEP), glycerol-3phosphate (Gly-3P), 5-phosphoribosyl-pyrophosphate (PRPP), pentose-5-phosphate (P5P), fructose-1,6-diphosphate (FBP), sedoheptulose-7-phosphate (Sed7P), glucose-1-phosphate (G1P), glucose-6-phosphate (G6P), fructose-6-phosphate (F6P), mannose-6-phosphate (Man6P)), and nucleotides (adenosine diphosphate (ADP), adenosine triphosphate (ATP), cytidine diphosphate (CDP), cytidine triphosphate (CTP), guanosine diphosphate (GDP), uridine diphosphate (UDP), uridine monophosphate (UMP), uridine triphosphate (UTP)). All these compounds were identified in the MS data with a confidence level 1 (Creek et al, 2014), including confirmation with authentic compounds.

The isotopologues from these metabolites were extracted from the MS data collected on the reference material, and were assigned to molecular isotopic clusters in a targeted manner with the software Emzed (Kiefer et al, 2013) using a mass tolerance of 0.003 m/z. Carbon Isotopologue Distributions (CIDs) of the reference metabolites were then quantified from the corresponding mass fractions after correcting for the presence of all naturally occurring isotopes and the isotopic purity of the tracer (99%) using the software IsoCor, v2.0.4 (Millard et al. 2019). The complete dataset (including the list of reference metabolites, the isotopologues, their

analytical characterics, their abundances, the isotopic clusters and the metabolite CIDs) is detailed in Supplementary Information Table S1 and was used as reference material to evaluate the optimization of data extraction

2.3.2. Detection of LC/MS features using XCMS

LC/MS features were extracted using the XCMS package (Smith et al, 2006) in Rstudio. The isotopologue parameters optimization (IPO) tool (Libiseller et al, 2015) was first used to optimize XCMS parameters, using unlabelled samples (*E.coli*) as required. The set of parameters selected using the IPO tool are given in SI Table S2, and were used as starting settings for subsequent data processing optimization. All raw datasets (i.e. from unlabelled and labelled PT samples and *E.coli* samples) were grouped and processed in a single batch with XCMS (Smith et al, 2006) so that peaks were identified and integrated using exactly the same processing parameters. This operation was iteratively repeated after changing the parameter settings to minimize the difference between the XCMS data and a reference dataset, as explained in the Results section. The XCMS parameters and their tested range of values are described in SI Table S3. The parameters giving the optimal recovery of the reference data are given in SI Table S2.

2.3.3. Isotopologue clustering

The XCMS object containing the list of putative isotopologues was processed separately with the R packages X13CMS (Patti et al, 2014) and geoRge (Cappellades et al, 2016). The parameters for the two programs are listed in SI Table S4.

2.3.4. Calculation of CIDs

Carbon isotopologue distributions were calculated from the relevant mass fractions of isotopic clusters after correcting for naturally occurring isotopes of elements other than carbon using IsoCor (Millard et al, 2019), accounting also for the MS resolution. The CIDs of metabolites in the PT samples can be predicted from the composition of the label input and the number of carbon atoms in the metabolites. The theoretical CIDs of metabolites in the PT sample were calculated using the equation

$$M_k = \binom{n}{k} * p^k * (1-p)^{n-k}$$

where *n* is the total number of carbon atoms in a molecular entity with *k* ^{13}C atoms and *p* is the abundance of ^{13}C isotopes. Here, the molecular enrichment of ^{13}C-acetate measured by NMR was $p = 0.512$. Standard deviations of measured CIDs were determined from the analysis of five analytical replicates of the PT sample.

2.3.5. Statistical analyses

Principal Component Analysis (PCA) was applied to all the biological samples (unlabelled and labelled *E.coli* strains). PCA was performed using SIMCA, v 15.0.02.5959] to separate all the biological samples (unlabelled and labelled *E.coli* strains) into different classes. A Wilcoxon test (p value ≤ 0.05) was used to identify the most discriminating isotopologues between the two *E.coli* strains. An in-house database with 47 metabolites was then used for metabolite identification based on exact masses and standard retention times (RTs). Metabolite identification was confirmed with authentic compounds.

2.4. Evaluation criteria for processing optimization

This study is primarily based on the establishment of specific metrics to evaluate isotopic measurements and validate software parameters used to process data in untargeted MS-based isotopic tracing investigations of metabolism. We used the criteria established by Heuillet et al. (2018) to validate MS-based isotopic measurements:

- The mass accuracy, i.e., the error on isotopologue masses, estimated from the difference between the theoretical (M_{th}) and experimental (M_{exp}) mass of each isotopologue.

  $$\text{mass accuracy (ppm)} = (M_{th} - M_{exp})/M_{th} \times 10^6$$

- The RT accuracy, i.e., the error on the measured RTs, calculated from the difference between theoretical and measured RTs.

  $$\text{RT accuracy (s)} = (RT_{th} - RT_{exp})/RT_{th} \times 100$$

- The RT isotopic deviation, i.e., the measured deviation of RTs between isotopologues belonging to the same isotopic cluster.
- The area precision, i.e., the spread of measured areas, estimated from the standard deviation of measurements on PT sample replicates.
- The CID accuracy (CID mean bias), i.e., the error on measured CIDs, simply the difference between predicted and measured CIDs.

  $$\text{CID accuracy} = CID_{th} - CID_{exp}$$

We also used two further criteria to evaluate the closeness of the clustering data obtained with the two ^{13}C-clustering programs to the clusters obtained by manual analysis:

- The recall, i.e. the ability of the process to retrieve the information, calculated as follows:

$$\text{recall} = \frac{\{\text{relevant isotopic clusters}\} \cap \{\text{retrieved isotopic clusters}\}}{\{\text{relevant isotopic clusters}\}}$$

- The cluster precision, i.e, a measure of the relevance of the retrieved information, defined by:

$$\text{cluster precision} = \frac{\{\text{relevant isotopic clusters}\} \cap \{\text{retrieved isotopic clusters}\}}{\{\text{retrieved isotopic clusters}\}}$$

3. Results and discussion

3.1. Overall strategy and case study

The aim of this work was to optimize data processing in untargeted MS-based isotopic tracing studies of metabolism, which refers here to isotope-labelling experiments aiming at the identification of metabolic pathways from the detailed examination of the label incorporation into metabolites. In contrast to isotope-assisted metabolomics in which a labelled sample with known label content of metabolites is added to assist in metabolome annotation (or quantification) (de Jong et al, 2012; Wang et al, 2019), the labelling patterns of metabolites are not known – and are not predictable - in tracing studies of metabolites. Indeed, they represent the desired information to elucidate metabolic pathways. According to the isotopic composition of the labelled source and to the operating metabolic pathways, potentially any combination of isotopologues can be generated for each metabolite in such experiments, which means that the complete isotopic envelope has to be measured to get valuable metabolic information. Moreover, the isotopologue abundances are determined by the pathways activities and can be exploited to measure metabolic fluxes. Hence, untargeted MS-based isotopic tracing studies of metabolism can be defined as the quantitative measurement of the complete isotopic envelope of all detected metabolites. It currently represents a major challenge in terms of MS data processing and interpretation because both metabolites and their labelling patterns are not known. Some software tools have been recently introduced to perform automated extraction of isotopic clusters in untargeted MS-based isotopic tracing studies, but due to the high complexity of the MS data collected in such studies, specific strategies to optimize the parametrization of these tools are required. In this work, a methodology to optimize the extraction of complete isotopic envelopes of all metabolites detected in full-scan MS spectra of labelled samples is introduced. The raw MS data in these experiments are processed in two steps: 1) extraction of individual isotopologues and 2) grouping of individual isotopologues into isotopic clusters. The proposed strategy for optimizing data processing in this context is shown in Figure 2.1. The key feature is the addition to the analytical batch of a reference isotopically labelled sample to optimize data processing parameters. The labelling data of a set of metabolites are manually extracted from the MS data of the reference material to generate a reference dataset and the processing parameters are then optimized by minimizing the difference between this reference dataset and the data extracted for the reference metabolites. The same reference dataset is used to optimize the isotopologue extraction and isotopologue clustering steps.

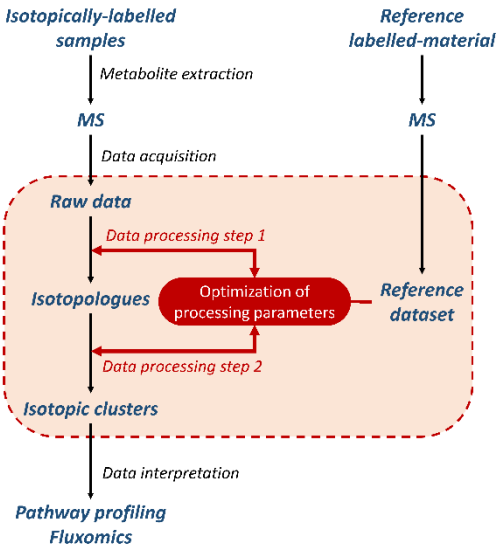


Figure 2.1 : Strategy for software parameter optimization in untargeted MS-based isotopic profiling using a reference labelled material.

The proposed optimization process is generic and can be applied to various stable isotope tracers used to investigate metabolism. Its use is demonstrated here for ^{13}C-tracing, which is the most widespread approach in isotopic studies of metabolism (Zamboni et al, 2009; Wiechert et al, 2001). As a test case to illustrate the application and relevance of the proposed optimization strategy, a ^{13}C-labelling experiment was performed in which two *E. coli* strains (wild-type BW25113 and its Δ*zwf* derivative knocked-out for the gene encoding the first committed step of the pentose-phosphate pathway) were grown in the presence of ^{13}C-labelled glucose as sole carbon source. The intracellular metabolites were sampled at mid-exponential growth and analysed by LC-MS. A reference material was analysed together with the biological samples to optimize the data processing. The reference material and its use for data optimization are described in detail in the following sections. To properly evaluate data quality throughout the optimization process, all samples (including the reference material) were produced and analysed in five replicates. In keeping with the requirements of the ^{13}C-profiling software furthermore, unlabelled samples (five replicates) of the reference material and of the *E. coli* strains were produced and analysed in the same analytical batch as the labelled samples.

3.2. Definition of reference sets for optimization

3.2.1. Choice of the reference material

Various isotopically labelled materials can be used, provided they satisfy a number of criteria related to the analytical method, the analysed samples, and the biological question to be addressed. The *reference material* should ideally have an identical or similar matrix to the samples of interest to generate the same matrix effects in the MS experiments and contain the same metabolites. It is very important for the labelling patterns of the metabolites to be known or be fully predictable to provide reliable reference data for the optimization process.

The reference material used here was a so-called Pascal triangle (PT) sample. PT samples are biologically produced materials whose isotopic composition is designed to obtain metabolites with tracer isotopologue distributions that match the binomial coefficients of Pascal's triangle. Details about these samples and their application to MS-based isotopic tracing studies can be found in Millard et al. (2014), Heuillet et al. (2018) and Schwaiger-Haber et al, (2019). PT samples were used here for several reasons. First, the fact that the sample could be produced by cultivating *E. coli* on ^{13}C-labelled acetate and collecting intracellular polar metabolites, meant that it had exactly the same matrix as the biological samples to be analysed. Second, the chosen PT sample satisfies many of the above-mentioned criteria for reference materials, including a broad metabolome coverage, fully predictable labelling patterns and broad coverage of the isotopologue space (all tracer isotopic forms of the same metabolite are present at the same abundance).

3.2.2. Definition of the reference dataset

The *reference dataset* corresponds to analytical data manually extracted from the *reference material* for a list of selected metabolites (the *reference metabolites*) and used as reference data during the optimization process. As for the *reference material*, various sets of metabolites can be used. The *reference metabolites* should be sufficient in number to cover the metabolome. They should be known compounds so that their labelling patterns can be extracted in a targeted fashion and complete isotopic clusters should be reliably detected in the reference material to optimize isotopologue recovery and isotopologue grouping. Note that the *reference metabolites* do not necessarily have to occur in the biological samples for the data optimization process itself since this depends only on the data from the *reference material*. However, they should be selected for their relevance to the objectives of the study.

For this case study, we selected 25 metabolites that are representative of the central metabolism of *E. coli* and are known to be reliably detected with our analytical method. All of them are confirmed level 1 annotated metabolites (Creek et al, 2014). The complete list of the selected *reference metabolites* is given in the Material & Methods (Section 2.3.1). This set of metabolites was consistent with the metabolite content of the labelled reference material and the biological question for the *E.coli* strains considered in this work. According to the elemental formula of the 25 *reference metabolites*, the *reference dataset* should consist of 25 isotopic clusters containing 184 tracer (carbon) isotopologues in total. By manually processing the MS data collected for the PT sample using the software Emzed (Kiefer et al, 2013), all 25 isotopic clusters were found, along with 181 tracer isotopologues (Table S1). The missing isotopologues corresponded to MS signals that were either undetected (CDP M0 and Mn) or with too low S/N ratio (G1P Mn).

The *reference dataset* was further characterized for the mass accuracy and RT isotopic deviation of individual isotopologues. Compared to their theoretical values the mean mass error was 1.42 ± 1.1 ppm for the 181 isotopologues. For the 25 reference metabolites, the RT isotopic deviation ranged from zero to 3 seconds with a mean relative error of 0.02% across the complete analytical run. These results indicate that the analytical characteristics (m/z, RT pairs) of the detected isotopologues are fully consistent with the values expected for the selected metabolites.

The experimental CIDs of the corresponding metabolites were calculated from the *reference dataset* to generate reference values (*reference CIDs*) (Table S1), which were validated by comparing them with predicted values for the PT sample. The CIDs measured manually for all 25 metabolites deviated by less than 5% on average from the predicted values (Fig S2).

These results highlight one of the benefits of using a reference material such as the PT sample to optimize processing, namely that analytical problems – limited sensitivity in this case – can be identified and considered separately from processing issues. The 181 isotopologues in the reference dataset are referred to hereafter as the *reference isotopologues.*

3.3. Optimization of isotopologue extraction

The proposed strategy for data processing optimization based on a reference material is illustrated in Figure 2.2 and involves two steps (i) optimization of isotopologue extraction and (ii) optimization of isotopologue clustering, as described in detail below. Briefly, in the first step, the *benchmark isotopologues* of the *reference metabolites* are identified automatically using extraction software (XCMS in this work) and are compared to the *reference isotopologues* using three evaluation criteria (recovery rate, analytical characteristics, and isotopologue integrals) and the extraction parameters are then iteratively modified to minimize the difference between the two isotopologue datasets.

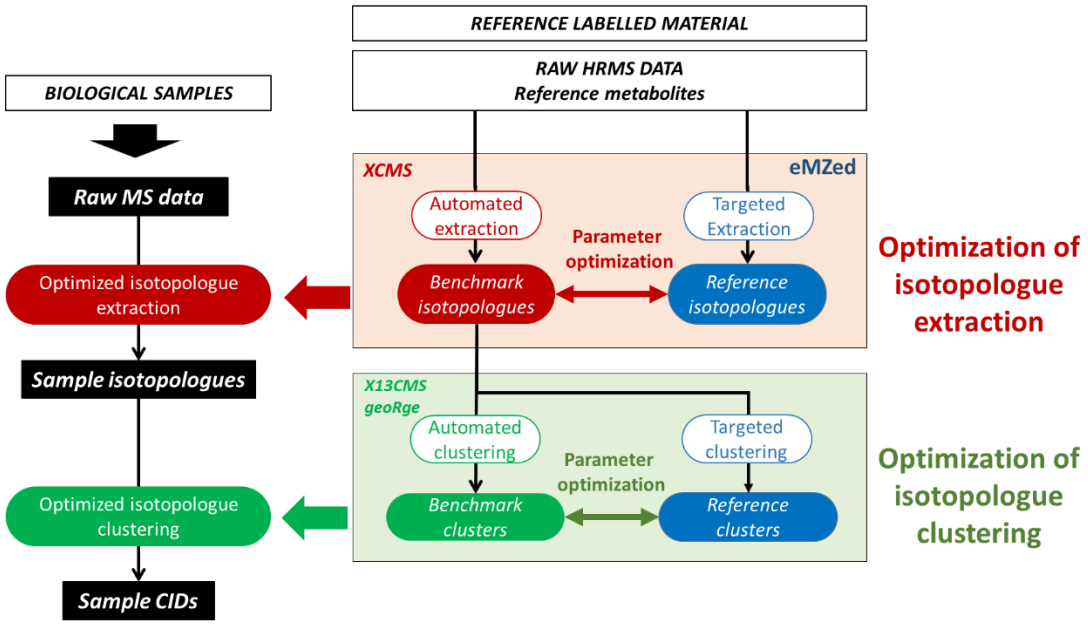


Figure 2.2 : Proposed strategy for the two-step optimization of data processing in untargeted MS-based isotopic tracing studies

3.3.1. Starting the optimization process

Some tools were recently published to perform automated parametrization of software in untargeted metabolomics (Manier et al, 2018; Libiseller et al, 2015). Because isotopologue extraction is performed with the same tools as feature extraction in untargeted metabolomics – i.e. XCMS in this study-, such tools can be also applied to untargeted isotopologue analysis. In this work we used the tool IPO (Libiseller et al. 2015; Alboniga et al, 2020), which was specifically designed to parameterize XCMS, to provide starting parameter settings for isotopologue extraction. In compliance with the IPO guidelines, this was done with MS data collected on unlabelled samples – i.e. the unlabelled PT samples. The so-obtained parameters (Table S2) were applied to extract the isotopologues from the labelled PT samples (*benchmark*

isotopologues). A total of 164 out of the 181 *reference isotopologues*, were retrieved at the end of this process (Fig S3a). Closer inspection of the data (Table S5) showed that the mass of the extracted peak differed significantly from that of the corresponding peak in the reference data (Table S4). Indeed 22 isotopologues showed mass errors above 3ppm, the error being higher than 5 ppm for 8 of them, and up to 18.8 for malate M2. This mass discrepancy, together with the fact that 17 isotopologues were not detected at all, indicates that the IPO-defined parameter settings were not optimal for isotopologue extraction. Such results can be explained because IPO was designed to optimize the processing of MS data collected on unlabeled material. The processing of MS data collected on labeled material, which are much more complex (more peaks with lower intensities), requires specific optimization tools. The tool IPO was found useful to provide a first set of parameter values which could be used to as a starting point to evaluate the benefit of the proposed optimization strategy.

3.3.2. Manual parameter selection

The XCMS parameters were next optimized using a semi-manual approach depicted in Figure 2.2. The IPO parameters (see Materials and Methods, Table S2) were used as a starting point for this process, but other tools or starting values could also have been used. In each optimization round, the isotopologues from the 25 *reference metabolites* were automatically extracted using XCMS and gathered into *benchmark isotopologues*. The *benchmark isotopologues* were then compared to the *reference isotopologues* using the evaluation criteria mentioned above. The process was then iterated after changing the extraction parameters values to maximize the agreement between the *benchmark isotopologues* and the *reference isotopologues*.

This optimization process was used for the five labelled PT samples in the analytical batch. Table S2 lists the parameter settings giving the optimal isotopologue recovery by automated extraction accross the five PT sample replicates (Fig. S3b). The optimized parameters allowed the recovery of 174 isotopologues, i.e. 10 more than with the initial parameter settings. This gain in recovery was accompanied by a gain in data quality (Table S5). The average error in mass accuracy over the common detected isotopologues (for 161 isotopologues) was 1.29 ± 1.02 ppm, to be compared to 1.74 ± 2.38 ppm in the initial data. The lower standard deviation on the mass errors indicated a higher precision of isotopologue masses after optimization. The RT accuracy for the *benchmark isotopologues* compared to the *reference isotopologues* was 0.29 ± 0.22 s on average. The results are given in full in the

Supplementary Data (Table S5) and highlight the improvement in data extraction afforded by the proposed optimization strategy.

Nevertheless, seven reference isotopologues remained undetected in the optimized dataset, indicating that the automated process was slightly less efficient than manual extraction. Five of the missing isotopologues, the M0 and Mn of ADP and GDP and the Mn_{-1} of CDP, were not detected in any of the five PT sample replicates. The two others missing isotopologues (UDP M0 and CDP M1) were detected in only one replicate. The chromatographic signal appeared more intense in this replicate than in the others. Overall, the above data indicated not only that the number missed isotopologues was decreased after optimization process, but also that the detected isotopologues were much better defined.

3.3.3. Quality of isotopologue quantification

In isotopic studies of metabolism, valuable quantitative information on biochemical pathways is obtained from isotopologue abundances. The reliability of isotopologue quantification is a major issue at the data acquisition level because ionization problems and matrix effects mean that MS is not inherently quantitative. Methods for validating MS methods for reliable isotopologue measurements – including the benefits of using PT samples for such a purpose – as discussed recently by Heuillet et al. (2018) and Schwaiger-Haber et al. (2019), are beyond the scope of this work. Isotopologue quantification can also be limited by data processing. Several factors can be problematic, but the main limitation is the capability of the processing software to properly integrate the MS signals. Optimizing data processing in untargeted isotopic tracing studies therefore also means ensuring isotopologue abundances are properly measured.

Because manual integration is somewhat arbitrary and automatic integration is imperfect regardless of the algorithm considered, the quality of isotopologue quantification was controlled and maximized throughout the optimization process by comparing isotopologue abundances in the *benchmark isotopologues* to those in the *reference isotopologues*. Two methods were used to evaluate quantification accuracy.

We first compared the absolute abundances of individual isotopologues separately. Figure S4 shows that the *benchmark isotopologues* have absolute abundances very close to those in the reference dataset. The isotopologue areas were closely similar whether integrated manually or automatically, indicating the reliability of automated isotopologue quantification after optimization. Some slight overestimations were observed for two metabolites showing

noisy peaks (succinate and malate), while individual isotopologues with very low S/N (lightest and heaviest isotopologues of G1P, Sed7P) were underestimated.

The mean SD in integrated areas across the five replicate measurements was 0.0003 for the 174 extracted isotopologues (Figure 2.3A), emphasizing the high quantitative reliability of the automatic isotopologue extraction process. We next compared the isotopologue abundances relative to the isotopic cluster of the corresponding metabolite by calculating the CIDs. Isotopologue quantification errors propagate to the entire CID vector, so that comparing CIDs calculated after automated extraction to manually measured CIDs is a sensitive method of detecting processing-induced quantification errors.

Benchmark CIDs were calculated for the *benchmark isotopologues* – before (i.e. with IPO settings) and after parameter optimization – after reconstructing molecular isotopic clusters and correcting for naturally-occurring isotopes. The *benchmark CIDs* were then compared to the *reference CIDs*. The data are shown in full in the Supplementary Data (Fig S5). The results obtained with the final, XCMS-optimized dataset are shown in Figure 2.3 B-C. For all 25 metabolites, the CIDs obtained after parameter optimization were in close agreement with the reference values (average error below 2%; Figure 2.3B). Figure 2.3C compares the CIDs of selected metabolites with reference values before and after parameter optimization. The CIDs calculated from the initial non-optimized dataset are generally biased and show significant inter-replicate variability (e.g. PRPP, Figure 2.3C). This is partly because many isotopologues go undetected with the IPO parameters, as mentioned above. After parameter optimization however, the CIDs were in good agreement with the corresponding values in the *reference dataset*, showing the benefit of the proposed optimization strategy. Interestingly, the CIDs of both malate and succinate, whose isotopologue abundances were overestimated in the optimized dataset (Figure S4), were also closely consistent (Figure 2.3C). This means that although the MS signals of the two compounds were overestimated in the automatically extracted data, the quantitative relationships between isotopologues of the same compound were preserved. This observation points to a potential bias in interpreting the abundances of individual isotopologues from different metabolites to derive quantitative metabolic information – e.g. comparing the M+5 isotopologue of citrate to the M+5 isotopologue of glutamine to determine reductive glutamine metabolism –without considering all potential data acquisition and processing problems.

Altogether, the above results clearly emphasize the significant improvement in the quality of the quantitative data achieved through the proposed optimization strategy. The results also show that data processing can be a substantial source of bias in MS-based untargeted isotopic tracing investigations, in terms of the number, correctness and quantification of the recovered isotopologues.

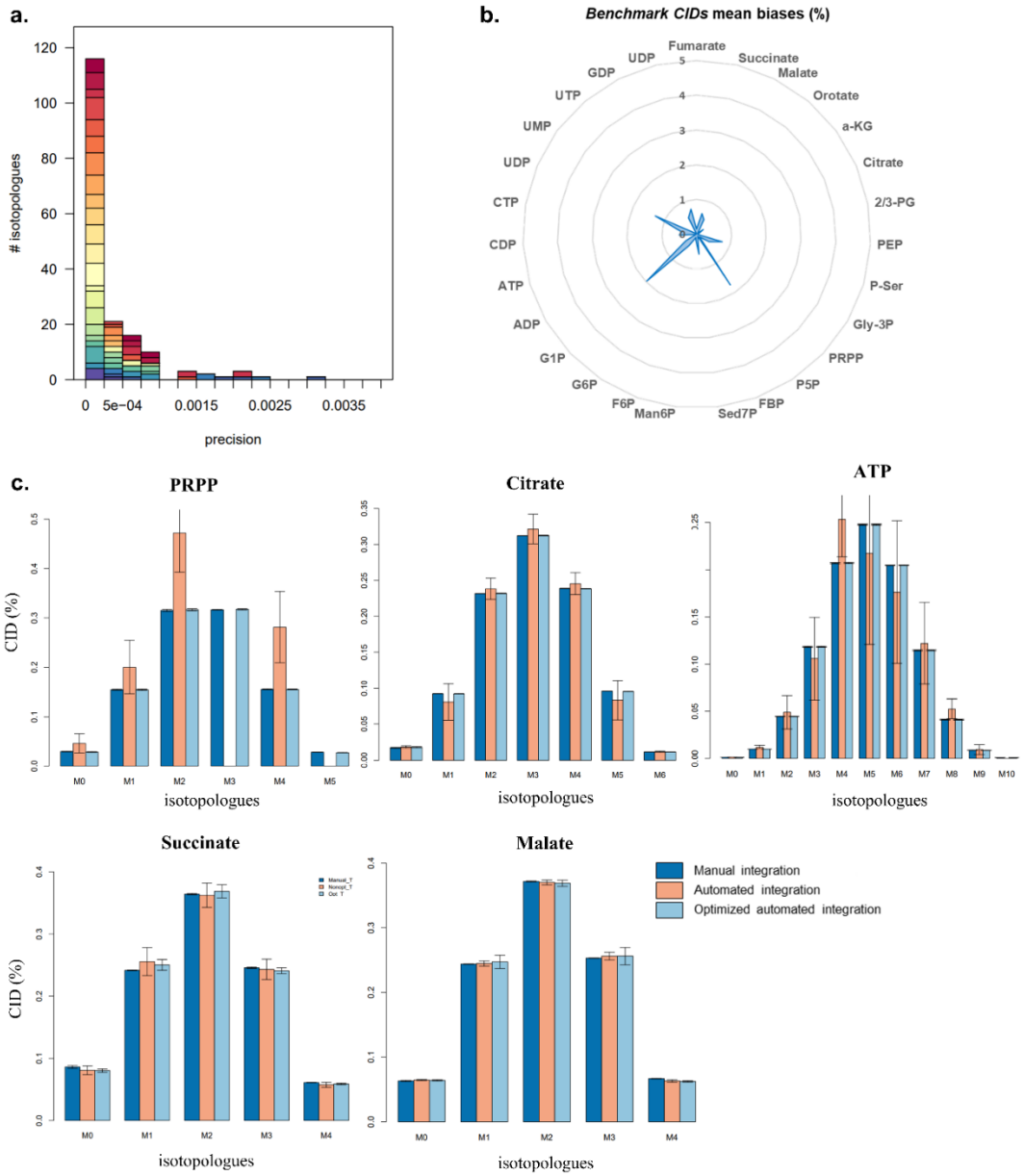


Figure 2.3 : **Impact of parameter optimization on the measurement of isotopologue abundances. a.** Distribution of precision for the 174 XCMS-extracted *benchmark isotopologues* in the PT sample. **b.** Mean biases (%) of optimized *benchmark CIDs* with respect to *reference CIDs.* **c.** Comparison of *reference CIDs* (dark blue), IPO *benchmark CIDs* and optimized *benchmark CIDs* (light blue) for PRPP, citrate, ATP, malate and succinate in the PT samples.

3.4. Optimization of isotopologue clustering

Isotopologue clustering consists in the grouping of extracted isotopologues into metabolite isotopic clusters (Figure 2.2). Two different programs, geoRge and X13CMS (Patti et al, 2014; Capellades et al, 2016; Dange et al, 2020), were used to do this. Clustering was optimized in a similar fashion as the extraction process was (Figure 2.2). The isotopic clusters of the 25 reference metabolites (*reference clusters*) were manually extracted from the optimal set of isotopologues. The software was used to automatically extract the 25 clusters (*benchmark clusters*) from the same dataset. The optimization consisted in adjusting software parameters to minimize the difference between benchmark clusters and reference clusters.

The quality of clustering was evaluated from the proportion of correct clusters that were recovered. A correct cluster was defined as containing only all the correct isotopologues. Two types of incorrect cluster were considered: incomplete clusters, missing one or more isotopologues, and corrupt clusters, with one or more spurious isotopologues. We defined two figures of merit to optimize based on the proportions of correct and incorrect clusters in the *benchmark clusters*: recall, or sensitivity, the ability to detect a cluster for all 25 reference metabolites; and precision, the number of correct clusters retrieved in the *benchmark clusters*. The software parameters were then iteratively modified to maximize the recall and the precision of the *benchmark clusters*.

Preliminary tests showed that the quality of the clustering depended mainly on two parameters (isotopologue mass deviation and RT window). The isotopologue mass deviation ("ppm" for X13CMS and "ppm.s" for geoRge) is the acceptable error in m/z measurements between successive isotopologues in the same isotopic cluster (the accuracy of isotopic distances), which should be the mass difference between ^{13}C and ^{12}C (1.00335 m/z). The RT window ("RT window" for X13CMS and "rt.win.min" for geoRge), corresponds to the tolerance on the RTs of isotopologues from the same metabolite, which should in theory be exactly the same. The RT deviation was measured to vary between 0.2 and 7.2 in the isotopologue dataset obtained after XCMS optimization (Table S5, see 3.3.2). From these values, two different RT windows (5 and 10 s) were considered and the isotopologue mass deviation was varied from 1 to 10 ppm. The noise threshold was deliberately set at a low value (5000) to maximize peak extraction.

As reported previously (Dange et al, 2020), many redundancies were observed in the clusters obtained with geoRge independently of the parameters used. This is due to the clustering algorithm of geoRge, which generates various clusters from the same set of isotopologues. The

optimization for geoRge was therefore performed after manual curation of obvious redundancies in the geoRge dataset.

For both programs, a low mass deviation threshold produced more incomplete clusters while increasing the mass deviation generated more corrupt clusters (Figure 2.4). The missing species were most often the Mn isotopologues of nucleotides (ADP, ATP, CTP, GDP, UDP, UMP, UTP), which can be explained by the lower quality of the XCMS data for these species and their larger mass deviation (see section 3.3 and Table S4). Processing the data with a larger mass tolerance allowed these isotopologues to be recovered but also tended to generate corrupt clusters.

An RT window of 10 s was found to yield a greater proportion of correct clusters than an RT window of 5 s for all isotopologue mass deviation values except 1 ppm, for which the proportion of correct clusters was the same with both. The number of corrupt clusters did not depend on the length of the RT window, regardless of the mass deviation used. These results are because while unlabelled and labelled samples were processed simultaneously, the heaviest isotopologues were only detected in the ^{13}C-enriched samples leading to a certain amount of variation in RTs.

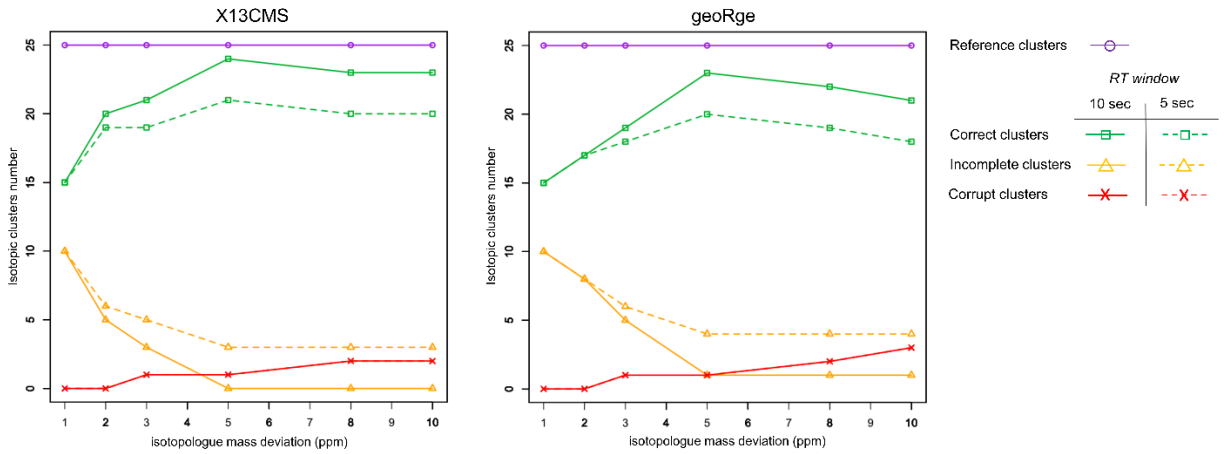


Figure 2.4 : **Quality of isotopologue clustering.** Number of correct (green line), incomplete (orange line) and corrupt (red line) isotopic clusters detected by X13CMS and geoRge depending on the isotopologue mass deviation (1, 2, 3, 5, 8 or 10 ppm) and the RT window (5 s, dotted line; or 10 s, solid line) based on the set of *reference clusters* (purple line) for the five replicates of the PT sample.

These results are based on a targeted search of 25-*benchmark clusters* and we assume that the observed errors are representative of the clustering process for the entire dataset. Table 2.1 shows that with an RT window of 10 s, the precision and recall were optimal with both programs at a mass deviation of 5 ppm. The equivalent results for an RT window of 5 s are provided in Table S6.

Table 2.1 : Cluster precision and recall for X13CMS and geoRge with a RT window of 10 s and different isotopologue mass deviations, evaluated for the 25 *reference metabolites* in the PT sample

	Isotopologue mass deviation (ppm)	**1**	**2**	**3**	**5**	**8**	**10**
X13CMS	precision	60%	80%	84%	96%	92%	92%
	recall	100%	100%	100%	100%	100%	100%
geoRge	precision	60%	68%	76%	92%	88%	84%
	recall	100%	100%	100%	100%	100%	100%

The recall of both programs was 100%, meaning that all 25 *reference metabolites* were retrieved. Almost all these clusters were correct, with a precision of 96% (1 incorrect cluster) and 92% (2 incorrect clusters) for X13CMS and geoRge, respectively. The ATP cluster was found incorrect with both programs, with some isotopologues wrongly identified by the software as M35 to M39. The second incorrect cluster for geoRge was ADP, whose M9 isotopologue was missed because of the statistical rules applied by geoRge to select potential enriched isotopologues in the labelled samples. Close inspection of the data for this isotopologue showed that some noise had been integrated for the unlabelled samples, and could be interpreted as signal by geoRge, so that the M9 peak in the labelled data was not considered as labelled.

These results show that both programs perform well despite their slightly different approaches. Briefly, geoRge compares potential isotopic peaks in the labelled and unlabelled samples with all candidate basepeaks within the vector of masses calculated for each potential isotopologue. On the other hand, X13CMS compares all potential isotopologue peak pairs within a RT bin, groups them together based on a common basepeak and discards duplicate information. In this targeted search of 25 *reference metabolites*, X13CMS generated a smaller number of incorrect clusters than geoRge and no correction for clustering redundancy was required. It has been shown that in spite of these redundancies, geoRge tends to generate fewer false positives than X13CMS, but can miss some features that X13CMS finds (Capellades et al. 2016; Dange et al. 2020). The above results show that regardless of the software chosen, independently optimizing the parameters used to group isotopic clusters is essential.

3.5. Application to the case study

To illustrate its use, the optimization workflow was applied to the study of wild type *E. coli* BW 25113 and a mutant deleted for the *zwf* gene (Δ*zwf*) that encodes glucose-6-phosphate dehydrogenase (G6PDH). This mutation has a negligible impact on the growth of the bacterium but leads to metabolic adaptations, which can be nicely revealed by using ^{13}C-

labelling experiments (Nicolas et al, 2007; Zhao et al, 2004; Bergès, Cahoreau et al, 2021). We used this example of an untargeted MS based isotopic tracing investigation to illustrate how the proposed workflow optimizes the recovery of this kind of labelling information.

The *E. coli* samples were analysed by LC-MS and first processed using the starting (IPO-derived) parameter settings (Table S2). Then data processing was repeated with the optimal parameter settings (Table S2). The gain in data quality resulting from the optimization is illustrated in for the WT strain (Table 2.2). The fact that the proposed approach yields the same cluster precision for biological samples as for the reference material, confirms the efficiency of the optimization process.

Table 2.2 : **Impact of parameter optimization on the extraction of data from the *E. coli* WT samples.** Cluster precision (%) of the two programs for the 25 *reference metabolites* before and after parameter optimization.

Clustering software	**Cluster precision (%)**	
	Starting parameter settings	Optimized parameter settings
X13CMS	40%	96%
geoRge	44%	92%

In total, 10129 isotopologues were extracted by XCMS from the two *E. coli* strains and were further processed with geoRge and X13CMS to group some of isotopologues in isotopic clusters. As already observed, geoRge generated a significant number of redundancies from this dataset compared to X13CMS (Fig 2.5A), but after manual curation the number of clusters was similar with both software (1037 and 1133 for geoRge and X13CMS, respectively) (Fig 2.5B). Over all the distinct clusters detected (1180) with both programs (Fig 2.5B), a total of 990 clusters had the same basepeak, corresponding to an overlap of 84%, of which 797 were identical in both cluster length and isotopologue composition (Table S7).

Regarding the greater number of clusters detected by X13CMS than by geoRge, the 190 isotopic clusters identified by one program and not the other (147 by X13CMS and 43 by geoRge) mostly contained a single isotopologue possibly because of instrumental noise or the presence of unlabelled metabolites or other unresolved peaks.

In isotope labelling experiments, isotopologue abundances can be interpreted either individually (*e.g.* the evolution of the M+3 peak intensity of a metabolite) or relative to the complete isotopic cluster (isotopologue distribution), the latter approach being the most common way of describing labelling patterns in ^{13}C-fluxomics. Here, mass fractions were calculated for all detected clusters in all the samples (unlabelled and labelled) from the outputs of X13CMS and geoRge, and PCA was used to explore differences in the isotopic profiles of

two *E.coli* strains (Fig 2.5C). Comparisons of isotopic profiles depend heavily on the number of features detected and how they are clustered, therefore on the quality of the data processing. The PCA plots for the two programs show a similar level of separation between the different biological conditions, confirming the repeatability of the workflow from sample preparation through to data processing. In these plots, labelled and unlabelled samples are strictly separated along the first PCA component and WT and Δ*zwf* strains along the second. The lack of separation between the two groups of unlabelled samples is expected because the unlabelled mass fractions have the same isotopomer composition (i.e natural abundance), such that the only discriminating factor is the presence or absence of peaks. On the contrary, the ^{13}C-enriched samples are closely grouped by strain on the plots according to their isotopic composition. This demonstrates the significant impact of the Δ*zwf* mutation on flux distribution, which then significantly affects the isotopic composition of the metabolites.

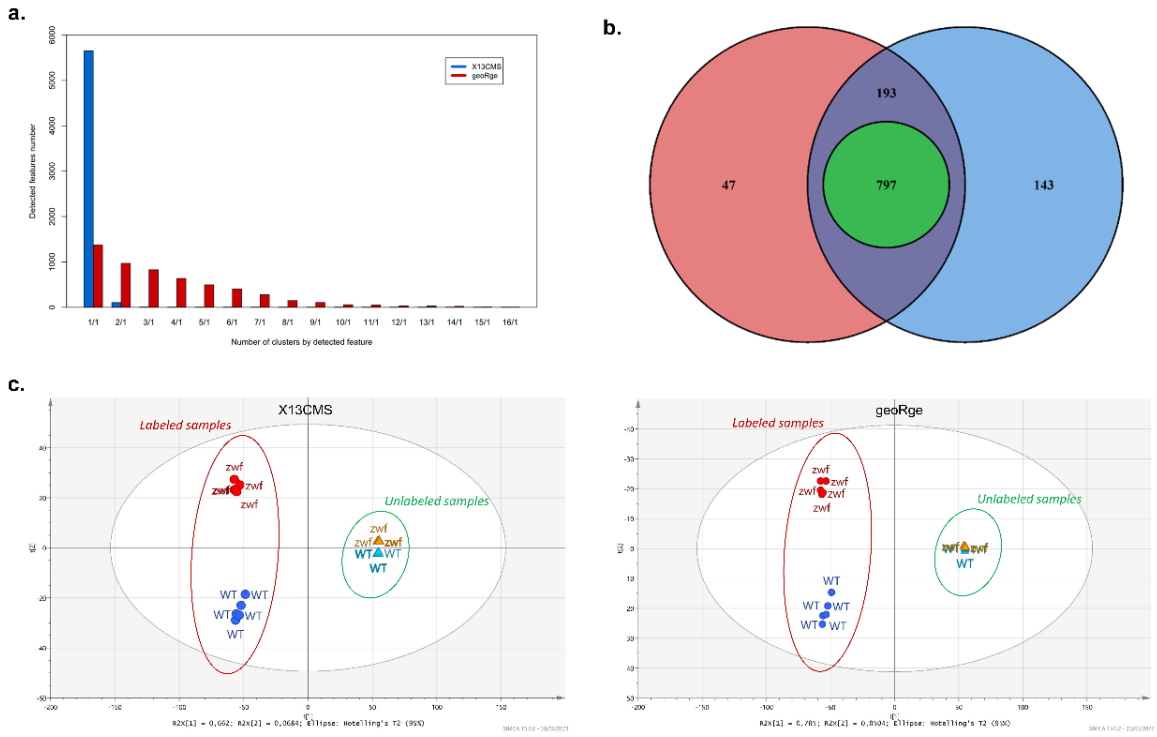


Figure 2.5 : **a.** Comparison of the number of isotopic clusters each detected isotopologue appears in for X13CMS (blue) and geoRge (red) software. **b.** Venn diagram of the number of isotopic clusters detected by X13CMS (blue) and geoRge (red) in biological samples, including clusters identical both in length and isotopologue composition (green circle). **c.** PCA plots of the extracted isotopic profiles of unlabelled and labelled wild-type and Δ*zwf* *E.coli* strains after processing using X13CMS (left) and geoRge (right).

The WT and Δ*zwf* groups were analysed to identify the most discriminating labelling data between the two strains. The corresponding isotopic clusters were compared using Wilcoxon tests, with 207 (X13CMS) and 138 (geoRge) of these clusters having more than one significantly different ($p \leq 0.025$) isotopologue between strains. By exploiting an in-house

database (containing 47 metabolites), 20 isotopic clusters could be assigned to metabolites with a level 1 confidence (Creek et al, 2014). They were related to glycolysis (Fumarate, Succinate, Malate, 2/3-PG, PEP, G6P, FBP), the PPP (Sed7P, Orotate, P5P, Shiki3P (CAS: 63959-45-5)) and nucleotide biosynthesis (ADP, ATP, CDP, CTP, UMP, UDP, UTP, UDP-Glucose (CAS: 133-89-1), UDP-Acetylglucosamine (CAS: 528-04-1)) (FigS7). Changes in the labelling patterns of these metabolites was fully consistent with the modifications expected for the Δzwf strains, which is known to significantly impact the partition between glycolysis and the PPP (Zhao et al, 2004; Nicolas et al, 2007), resulting also in differential labelling of the ribosyl moiety of nucleotides. Furthermore, the number of significantly different isotopic clusters that remain unidentified after this initial analysis demonstrates the power of the untargeted approach and the need for further identification.

4. Conclusion

This work emphasized that specific workflows have to be developed for optimal processing of the complex MS data that are generated in MS-based untargeted isotopic tracing studies of metabolism. Indeed, the results showed that significant gain in the recovery of valuable information was obtained by applying the proposed methodology for data processing optimization. The application of a suitable reference material to optimize software parametrization proved to increase not only the number of recovered isotopic data but also the quality of the data. Pascal Triangle samples are well suited for such purpose since they allow both the identification of analytical issues and optimization of data processing at the same time. Together with the progress in MS instrumentation and analytical methods, which allows to extend the metabolome – and fluxome – coverage, applying the proposed methodology is maximizing the biological value of isotopic tracing investigations by revealing the full metabolic information that is encoded in the labelling patterns of the metabolites.

References

Albóniga, O. E., González, O., Alonso, R. M., Xu, Y., & Goodacre, R. (2020). Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. Metabolomics: Official Journal of the Metabolomic Society, 16(1), 14. https://doi.org/10.1007/s11306-020-1636-9

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology, 2, 2006.0008. https://doi.org/10.1038/msb4100050

Bergès, C., Cahoreau, E., Millard, P., Enjalbert, B., Dinclaux, M., Heuillet, M., et al. (2021). Exploring the Glucose Fluxotype of the E. coli y-ome Using High-Resolution Fluxomics. Metabolites, 11(5), 271. https://doi.org/10.3390/metabo11050271

Bueschl, C., Kluger, B., Lemmens, M., Adam, G., Wiesenberger, G., Maschietto, V., et al. (2014). A novel stable isotope labelling assisted workflow for improved untargeted LC–HRMS based metabolomics research. Metabolomics, 10(4), 754–769. https://doi.org/10.1007/s11306-013-0611-0

Bueschl, C., Kluger, B., Neumann, N. K. N., Doppler, M., Maschietto, V., Thallinger, G. G., et al. (2017). MetExtract II: A Software Suite for Stable Isotope-Assisted Untargeted Metabolomics. Analytical Chemistry, 89(17), 9518–9526. https://doi.org/10.1021/acs.analchem.7b02518

Capellades, J., Navarro, M., Samino, S., Garcia-Ramirez, M., Hernandez, C., Simo, R., et al. (2016). geoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics. Analytical Chemistry, 88(1), 621–628. https://doi.org/10.1021/acs.analchem.5b03628

Chokkathukalam, A., Jankevics, A., Creek, D. J., Achcar, F., Barrett, M. P., & Breitling, R. (2013). mzMatch-ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. Bioinformatics (Oxford, England), 29(2), 281–283. https://doi.org/10.1093/bioinformatics/bts674

Chokkathukalam, A., Kim, D.-H., Barrett, M. P., Breitling, R., & Creek, D. J. (2014a). Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. Bioanalysis, 6(4), 511–524. https://doi.org/10.4155/bio.13.348

Creek, D. J., Chokkathukalam, A., Jankevics, A., Burgess, K. E. V., Breitling, R., & Barrett, M. P. (2012). Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. Analytical Chemistry, 84(20), 8442–8447. https://doi.org/10.1021/ac3018795

Creek, D. J., Dunn, W. B., Fiehn, O., Griffin, J. L., Hall, R. D., Lei, Z., et al. (2014). Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics, 10(3), 350–353. https://doi.org/10.1007/s11306-014-0656-8

Dange, M. C., Mishra, V., Mukherjee, B., Jaiswal, D., Merchant, M. S., Prasannan, C. B., & Wangikar, P. P. (2020). Evaluation of freely available software tools for untargeted quantification of 13C isotopic enrichment in cellular metabolome from HR-LC/MS data. Metabolic Engineering Communications, 10, e00120. https://doi.org/10.1016/j.mec.2019.e00120

de Jong FA, Beecher C. Addressing the current bottlenecks of metabolomics: Isotopic Ratio Outlier Analysis™, an isotopic-labeling technique for accurate biochemical profiling. Bioanalysis. 2012 Sep;4(18):2303-14. doi: 10.4155/bio.12.202. PMID: 23046270; PMCID: PMC3696345.

Heuillet, M., Bellvert, F., Cahoreau, E., Letisse, F., Millard, P., & Portais, J.-C. (2017). A methodology for the validation of isotopic analyses by mass spectrometry in stable-isotope labelling experiments. Analytical Chemistry, 90. https://doi.org/10.1021/acs.analchem.7b03886

Heux, S., Poinot, J., Massou, S., Sokol, S., & Portais, J.-C. (2014). A novel platform for automated high-throughput fluxome profiling of metabolic variants. Metabolic Engineering, 25, 8–19. https://doi.org/10.1016/j.ymben.2014.06.001

Hiller, K., Metallo, C. M., Kelleher, J. K., & Stephanopoulos, G. (2010). Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. Analytical Chemistry, 82(15), 6621–6628. https://doi.org/10.1021/ac1011574

Hoffmann, F., Jaeger, C., Bhattacharya, A., Schmitt, C. A., & Lisec, J. (2018). Nontargeted Identification of Tracer Incorporation in High-Resolution Mass Spectrometry. Analytical Chemistry, 90(12), 7253–7260. https://doi.org/10.1021/acs.analchem.8b00356

Huang, X., Chen, Y.-J., Cho, K., Nikolskiy, I., Crawford, P. A., & Patti, G. J. (2014). X13CMS: global tracking of isotopic labels in untargeted metabolomics. Analytical Chemistry, 86(3), 1632–1639. https://doi.org/10.1021/ac403384n

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics (Oxford, England), 24(21), 2534–2536. https://doi.org/10.1093/bioinformatics/btn323

Kiefer, P., Nicolas, C., Letisse, F., & Portais, J.-C. (2007). Determination of carbon labeling distribution of intracellular metabolites from single fragment ions by ion chromatography tandem mass spectrometry. Analytical Biochemistry, 360(2), 182–188. https://doi.org/10.1016/j.ab.2006.06.032

Kiefer, P., Schmitt, U., Müller, J. E. N., Hartl, J., Meyer, F., Ryffel, F., & Vorholt, J. A. (2015). DynaMet: a fully automated pipeline for dynamic LC-MS data. Analytical Chemistry, 87(19), 9679–9686. https://doi.org/10.1021/acs.analchem.5b01660

Kiefer, P., Schmitt, U., & Vorholt, J. (2013). EMZed: An open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. Bioinformatics (Oxford, England), 29. https://doi.org/10.1093/bioinformatics/btt080

Kluger, B., Bueschl, C., Neumann, N., Stückler, R., Doppler, M., Chassy, A. W., et al. (2014). Untargeted profiling of tracer-derived metabolites using stable isotopic labeling and fast polarity-switching LC-ESI-HRMS. Analytical Chemistry, 86(23), 11533–11537. https://doi.org/10.1021/ac503290j

Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., et al. (2015). IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics, 16(1), 118. https://doi.org/10.1186/s12859-015-0562-8

Mairinger, T., & Hann, S. (2017). Implementation of data-dependent isotopologue fragmentation in 13C-based metabolic flux analysis. Analytical and Bioanalytical Chemistry, 409(15), 3713–3718. https://doi.org/10.1007/s00216-017-0339-1

Manier, S., Keller, A., Meyer, M. (2018). Automated Optimization of XCMS Parameters for Improbed Peak Picking of LC/MS Data using the Coefficient of Variation and Parameter Sweeping for Untargeted Metabolomics. Drug Test Anal. 11(6):752-761. doi: 10.1002/dta.2552.

Millard, P., Delépine, B., Guionnet, M., Heuillet, M., Bellvert, F., & Létisse, F. (2019). IsoCor: isotope correction for high-resolution MS labeling experiments. Bioinformatics (Oxford, England), 35(21), 4484–4487. https://doi.org/10.1093/bioinformatics/btz209

Millard, P., Massou, S., Portais, J.-C., & Létisse, F. (2014). Isotopic studies of metabolic systems by mass spectrometry: using Pascal's triangle to produce biological standards with fully controlled labeling patterns. Analytical Chemistry, 86(20), 10288–10295. https://doi.org/10.1021/ac502490g

Nicolas, C., Kiefer, P., Letisse, F., Krömer, J., Massou, S., Soucaille, P., et al. (2007). Response of the central metabolism of Escherichia coli to modified expression of the gene encoding the glucose-6-phosphate dehydrogenase. FEBS letters, 581(20), 3771–3776. https://doi.org/10.1016/j.febslet.2007.06.066

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC bioinformatics, 11, 395. https://doi.org/10.1186/1471-2105-11-395

Schwaiger-Haber, M., Hermann, G., El Abiead, Y., Rampler, E., Wernisch, S., Sas, K., et al. (2019). Proposing a validation scheme for 13C metabolite tracer studies in high-resolution mass spectrometry. Analytical and Bioanalytical Chemistry, 411(14), 3103–3113. https://doi.org/10.1007/s00216-019-01773-7

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical Chemistry, 78(3), 779–787. https://doi.org/10.1021/ac051437y

Stuani, L., Riols, F., Millard, P., Sabatier, M., Batut, A., Saland, E., et al. (2018). Stable Isotope Labeling Highlights Enhanced Fatty Acid and Lipid Metabolism in Human Acute Myeloid Leukemia. International Journal of Molecular Sciences, 19(11), 3325. https://doi.org/10.3390/ijms19113325

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., et al. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nature Methods, 12(6), 523–526. https://doi.org/10.1038/nmeth.3393

Wang L, Naser FJ, Spalding JL, Patti GJ. A Protocol to Compare Methods for Untargeted Metabolomics. Methods Mol Biol. 2019;1862:1-15. doi: 10.1007/978-1-4939-8769-6_1. PMID: 30315456; PMCID: PMC6482454.

Weindl, D., Cordes, T., Battello, N., Sapcariu, S., Dong, X., Wegner, A., & Hiller, K. (2016). Bridging the gap between non-targeted stable isotope labeling and metabolic flux analysis. Cancer & Metabolism, 4. https://doi.org/10.1186/s40170-016-0150-z

Wiechert, W. (2001). 13C metabolic flux analysis. Metabolic Engineering, 3(3), 195–206. https://doi.org/10.1006/mben.2001.0187

Wiechert, Wolfgang, Möllney, M., Petersen, S., & de Graaf, A. A. (2001). A Universal Framework for 13C Metabolic Flux Analysis. Metabolic Engineering, 3(3), 265–283. https://doi.org/10.1006/mben.2001.0188

Wittmann, C. (2002). Metabolic flux analysis using mass spectrometry. Advances in Biochemical Engineering/Biotechnology, 74, 39–64. https://doi.org/10.1007/3-540-45736-4_3

Zaimenko, I., Lisec, J., Stein, U., & Brenner, W. (2017). Approaches and techniques to characterize cancer metabolism in vitro and in vivo. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1868(2), 412–419. https://doi.org/10.1016/j.bbcan.2017.08.004

Zamboni, N., Fendt, S.-M., Rühl, M., & Sauer, U. (2009). (13)C-based metabolic flux analysis. Nature Protocols, 4(6), 878–892. https://doi.org/10.1038/nprot.2009.58

Zamboni, N., Saghatelian, A., & Patti, G. J. (2015). Defining the metabolome: size, flux, and regulation. Molecular Cell, 58(4), 699–706. https://doi.org/10.1016/j.molcel.2015.04.021

Zhao, J., Baba, T., Mori, H., & Shimizu, K. (2004). Effect of zwf gene knockout on the metabolism of Escherichia coli grown on glucose or acetate. Metabolic Engineering, 6(2), 164–174. https://doi.org/10.1016/j.ymben.2004.02.004

Supplementary Data

Figure S-1. Conversion of raw data using MSConvert

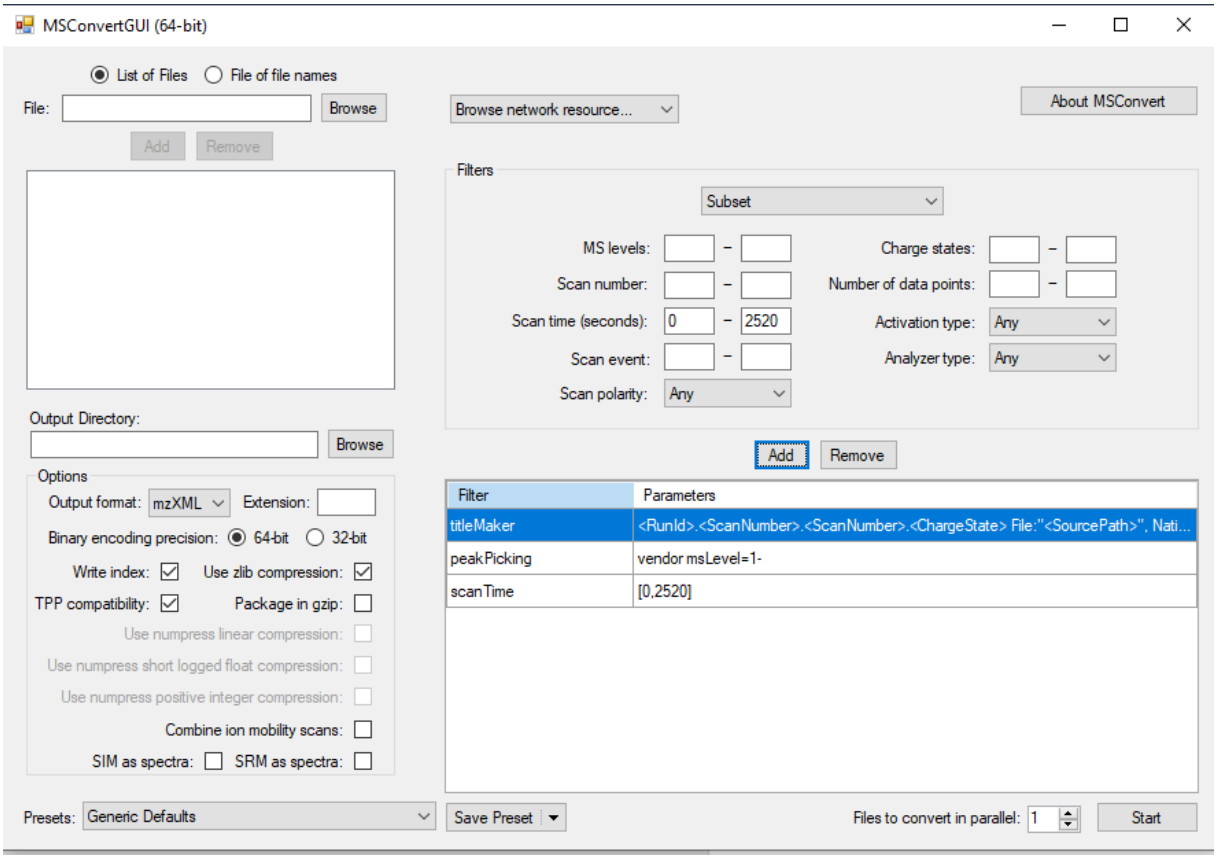


Table S-1. Reference datasets.

Reference metabolites	Reference Isotopologues	m/z	RT (sec)	area	Reference CIDs
Fumarate	M0	115.00350	981	56030235	0.07
	M1	116.00685	981	186005277	0.24
	M2	117.01019	981	286367284	0.37
	M3	118.01353	981	195796893	0.25
	M4	119.01686	981	50076785	0.06
Succinate	M0	117.01914	667	282803625	0.09
	M1	118.02248	667	793899995	0.24
	M2	119.02576	667	1195649421	0.36
	M3	120.02911	667	806966256	0.25
	M4	121.03250	667	202284443	0.06
Malate	M0	133.01417	670	424342051	0.06
	M1	134.01748	670	1626757443	0.24
	M2	135.02079	670	2478580090	0.37
	M3	136.02404	670	1688263347	0.25
	M4	137.02743	670	447848485	0.07
Orotate	M0	155.00991	1542	307245998	0.03
	M1	156.01325	1542	1523469272	0.16
	M2	157.01648	1542	3020822605	0.31
	M3	158.01978	1542	3019378130	0.31
	M4	159.02310	1542	1505195603	0.16
	M5	160.02658	1542	287627182	0.03
a-KG	M0	145.01417	882	30474657	0.05
	M1	146.01753	882	96355684	0.15
	M2	147.02088	882	195160824	0.30
	M3	148.02424	882	200443971	0.31
	M4	149.02753	882	102259920	0.16
	M5	150.03093	882	21077798	0.03
Citrate	M0	191.01994	1813	418678922	0.02
	M1	192.02325	1813	2201958564	0.09
	M2	193.02646	1813	5529085713	0.23
	M3	194.02973	1813	7452410371	0.31
	M4	195.03301	1813	5705670648	0.24
	M5	196.03628	1813	2288719058	0.10
	M6	197.03998	1813	280685227	0.01
2/3-PG	M0	184.98577	1741	1606633559	0.12
	M1	185.98906	1741	4951837448	0.37
	M2	186.99233	1741	5048634498	0.38
	M3	187.99559	1741	1678812830	0.13
PEP	M0	166.97505	1889	338965384	0.12
	M1	167.97833	1889	1060010283	0.38
	M2	168.98164	1889	1080910744	0.38
	M3	169.98506	1889	341973019	0.12
Gly-3P	M0	171.00653	514	60073480	0.21
	M1	172.00990	514	101324423	0.35
	M2	173.01317	514	99069549	0.34
	M3	174.01661	514	30466742	0.10
PRPP	M0	388.94405	2154	3773252	0.03
	M1	389.94740	2154	19526130	0.15
	M2	390.95072	2154	39788198	0.32
	M3	391.95403	2154	39994054	0.32
	M4	392.95738	2154	19646953	0.16
	M5	393.96044	2154	3620865	0.03
P5P	M0	229.01225	1087	15790399	0.03
	M1	230.01552	1087	80560741	0.16
	M2	231.01884	1087	164024774	0.32
	M3	232.02220	1087	165122627	0.32

	M4	233.02558	1087	79802609	0.15
	M5	234.02910	1087	12746690	0.02
FBP	M0	338.98810	2082	60911534	0.02
	M1	339.99131	2082	372220250	0.09
	M2	340.99460	2082	929801620	0.23
	M3	341.99782	2082	1247455449	0.31
	M4	343.00107	2082	939764667	0.24
	M5	344.00425	2082	375932623	0.09
	M6	345.00732	2082	545633560	0.01
Sed7P	M0	289.03269	1194	4238381	0.01
	M1	290.03604	1194	30245296	0.05
	M2	291.03938	1194	92560049	0.16
	M3	292.04273	1194	156438036	0.28
	M4	293.04608	1194	156671855	0.28
	M5	294.04944	1194	92460730	0.16
	M6	295.05165	1194	28587991	0.05
	M7	296.05471	1194	2996829	0.01
Man6P	M0	259.02304	1022	19907444	0.02
	M1	260.02636	1022	118046331	0.10
	M2	261.02961	1022	293341582	0.25
	M3	262.03289	1022	380574277	0.33
	M4	263.03623	1022	263394462	0.23
	M5	264.03964	1022	86609808	0.07
	M6	265.04309	1022	6272360	0.01
F6P	M0	259.02307	953	16279740	0.02
	M1	260.02637	953	99556729	0.09
	M2	261.02961	953	255419536	0.24
	M3	262.03289	953	345140729	0.32
	M4	263.03621	953	256612959	0.24
	M5	264.03961	953	96965877	0.09
	M6	265.04305	953	12107622	0.01
G6P	M0	259.02303	892	37644409	0.01
	M1	260.02632	892	232144699	0.09
	M2	261.02955	892	600703579	0.24
	M3	262.03284	892	813250693	0.32
	M4	263.03613	892	610245636	0.24
	M5	264.03951	892	232058201	0.09
	M6	265.04295	892	29220183	0.01
G1P	M0	259.02294	521	2188513	0.02
	M1	260.02630	521	11459761	0.10
	M2	261.02963	521	28785602	0.25
	M3	262.03295	521	38072046	0.33
	M4	263.03632	521	26003880	0.23
	M5	264.03973	521	7928553	0.07
	M6	-	-	-	0.00
ADP	M0	426.02220	2046	2457757	0.001
	M1	427.02540	2046	16922886	0.01
	M2	428.02871	2046	74325863	0.04
	M3	429.03196	2046	199432722	0.12
	M4	430.03521	2046	349509378	0.21
	M5	431.03846	2046	417235612	0.25
	M6	432.04172	2046	343721466	0.20
	M7	433.04502	2046	192069385	0.11
	M8	434.04845	2046	68002000	0.04
	M9	435.05177	2046	13419219	0.01
	M10	436.05448	2046	1474728	0.0008
ATP	M0	505.98885	2184	16246587	0.001
	M1	506.99234	2184	121367108	0.01
	M2	507.99573	2184	534430940	0.04
	M3	508.99902	2184	1422686675	0.12

	M4	510.00231	2184	2486164802	0.21
	M5	511.00559	2184	2972343098	0.25
	M6	512.00881	2184	2458060594	0.20
	M7	513.01208	2184	1378870250	0.11
	M8	514.01539	2184	497063414	0.04
	M9	515.01844	2184	104015344	0.01
	M10	516.02078	2184	9917869	0.0008
CDP	M0	-	-	-	0.00
	M1	403.01368	1818	6378057	0.02
	M2	404.01712	1818	27717271	0.07
	M3	405.02042	1818	64362251	0.17
	M4	406.02372	1818	95223930	0.25
	M5	407.02705	1818	93347270	0.25
	M6	408.03040	1818	61069389	0.16
	M7	409.03379	1818	24949899	0.07
	M8	410.03702	1818	4902167	0.01
	M9	-	-	-	0.00
CTP	M0	481.9776	2106	8538776	0.002
	M1	482.98096	2106	65463726	0.02
	M2	483.98428	2106	256077353	0.07
	M3	484.98759	2106	592625022	0.17
	M4	485.99086	2106	884008506	0.25
	M5	486.99417	2106	870996338	0.25
	M6	487.99749	2106	567941167	0.16
	M7	489.00077	2106	233553217	0.07
	M8	490.00390	2106	54518770	0.02
	M9	491.00624	2106	5699163	0.002
GDP	M0	442.01707	2258	1268796	0.004
	M1	443.02045	2258	2757797	0.01
	M2	444.02364	2258	12302899	0.05
	M3	445.02699	2258	32572041	0.12
	M4	446.03027	2258	56954948	0.21
	M5	447.03359	2258	67914235	0.25
	M6	448.03690	2258	55575927	0.20
	M7	449.04030	2258	30545046	0.11
	M8	450.04364	2258	10876038	0.04
	M9	451.04667	2258	2043296	0.01
	M10	452.04925	2258	61934	0.0002
UDP	M0	402.99462	2202	21775720	0.002
	M1	403.99807	2202	183070962	0.02
	M2	405.00128	2202	737410758	0.07
	M3	406.00453	2202	1728622632	0.17
	M4	407.00778	2202	2580480660	0.25
	M5	408.01098	2202	2561842283	0.25
	M6	409.01421	2202	1684070366	0.16
	M7	410.01752	2202	691542759	0.07
	M8	411.02084	2202	155155050	0.01
	M9	412.02337	2202	13935933	0.001
UMP	M0	323.02759	1931	31372647	0.002
	M1	324.03098	1931	262454720	0.02
	M2	325.03428	1931	1050573013	0.07
	M3	326.03757	1931	2416270973	0.17
	M4	327.04080	1931	3587714566	0.25
	M5	328.04402	1931	3551766195	0.25
	M6	329.04727	1931	2346046321	0.16
	M7	330.05055	1931	984698697	0.07
	M8	331.05388	1931	219836705	0.02
	M9	332.05669	1931	18514492	0.001
UTP	M0	482.96185	2208	9341187	0.002
	M1	483.96505	2208	81137802	0.02

	M2	484.96834	2208	319754515	0.07
	M3	485.97165	2208	744990764	0.17
	M4	486.97494	2208	1110932982	0.25
	M5	487.97820	2208	1103056897	0.25
	M6	488.98148	2208	723196293	0.16
	M7	489.98478	2208	300497734	0.07
	M8	490.98783	2208	70319340	0.02
	M9	491.99024	2208	6951686	0.002

Table S-2. Optimized XCMS automated peak extraction and integration parameters from the Pascal triangle sample.

	xcmsSet()				group()			retcor()
Parameters	**method**	**mass error**	**peakwidth**	**mzdiff**	**mzwid**	**minfrac**	**bw**	**method**
Starting (IPO) parameters	centWave	10	(24.86)	−0.036	0.013	0.5	5	-
Manually optimized parameters	centWave	10	(28.99)	−0.001	0.0015	0.2	10	peakgroups

Table S-3. Ranges for the parameter optimization. The XCMS parameters and their range of values tested starting with the parameters provided by IPO.

Parameter	**Minimum**	**Maximum**
peakwidth minimum	10	30
peakwidth maximum	80	120
ppm	10	10
mzdiff	-0.001	0.002
mzwid	0.001	0.01
bw	5	15

Table S-4. Optimized parameters for 13C-clustering software

Program	**Routine**	**Settings**	**Optimized value**
X13CMS	getIsoLabelReport()	RTwin	10
		ppm	5
		noise	5000
geoRge	PuIncSeeker()	fc threshold	1.2
		p-value threshold	0.05
		Basepeak min intensity	5000
	Basepeak_finder()	Basepeak mass error	5
		RT min win	10

Figure S-2. Reference CIDs mean biases. Mean biases (%) of *reference CIDs* with respect to the predicted values for the 25 *reference metabolites* in the PT samples.

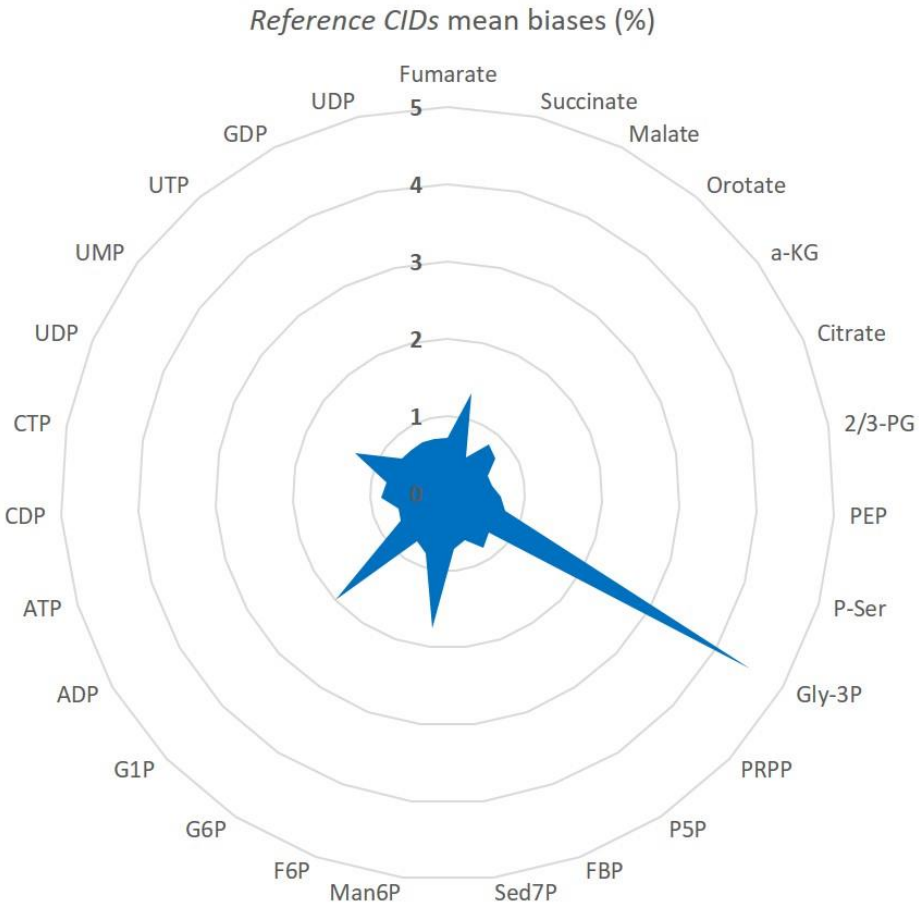


Figure S-3. Percentage of recovery in datasets. Number of isotopologues from *reference metabolites* that were detected or missing in the *reference isotopologues* and the *benchmark isotopologues* after **a.** XCMS processing using starting parameters and **b.** XCMS processing using optimized parameters in the PT samples.

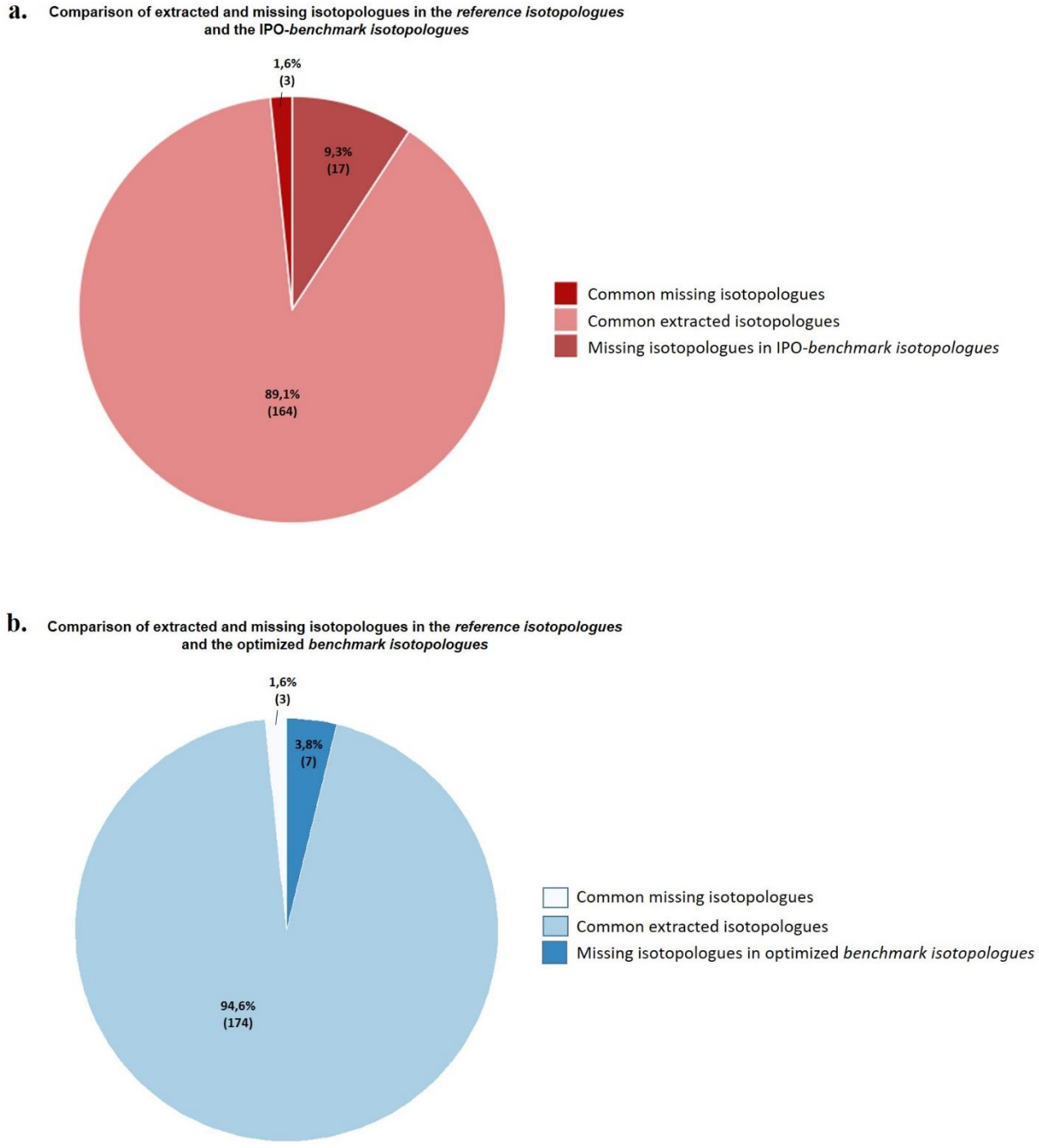


Table S-5. Mass accuracies of XCMS datasets. Mass accuracies for the 25 *reference metabolites* extracted in the five-labelled PT samples using XCMS compared to the theoretical masses.

				Starting parameter settings		Optimized parameter settings		
Metabolite	**Isotopologue**	**m/z theo**	**RT theo (sec)**	**m/z**	**Δm/z (ppm)**	**m/z**	**Δm/z (ppm)**	**RT (sec)**
Fumarate	M0	115.00368	981	115.00345	1.98	115.00345	1.98	980
	M1	116.00704	981	116.00682	1.86	116.00682	1.86	980
	M2	117.01039	981	117.01015	2.08	117.01019	1.69	979
	M3	118.01375	981	118.01356	1.57	118.01356	1.57	979
	M4	119.01710	981	119.01686	1.99	119.01687	1.94	979
Succinate	M0	117.01933	667	117.01911	1.90	117.01911	1.90	668
	M1	118.02269	667	118.02248	1.72	118.02248	1.72	668
	M2	119.02604	667	119.02571	2.75	119.02575	2.41	668
	M3	120.02940	667	120.02914	2.09	120.02913	2.18	668
	M4	121.03275	667	121.03254	1.77	121.03254	1.76	668
Malate	M0	133.01425	670	133.01408	1.24	133.01409	1.24	677
	M1	134.01761	670	134.01748	0.93	134.01748	0.93	677
	M2	135.02096	670	135.01841	18.87	135.02082	1.06	677
	M3	136.02432	670	136.02412	1.45	136.02413	1.37	677
	M4	137.02767	670	137.02750	1.25	137.02750	1.24	676
Orotate	M0	155.00983	1542	155.00978	0.31	155.00978	0.31	1 526
	M1	156.01319	1542	156.01312	0.39	156.01312	0.40	1 526
	M2	157.01654	1542	157.01631	1.48	157.01637	1.09	1 525
	M3	158.01990	1542	158.01968	1.35	158.01968	1.33	1 525
	M4	159.02325	1542	159.02309	1.03	159.02309	1.01	1 525
	M5	160.02661	1542	160.02655	0.36	160.02655	0.36	1 524
a-KG	M0	145.01425	882	145.01418	0.47	145.01418	0.48	884
	M1	146.01761	882	146.01756	0.32	146.01756	0.31	884
	M2	147.02096	882	147.02091	0.33	147.02091	0.33	883
	M3	148.02432	882	148.02428	0.21	148.02429	0.16	883
	M4	149.02767	882	149.02756	0.71	149.02756	0.71	883
	M5	150.03103	882	150.03097	0.35	150.03097	0.38	883
Citrate	M0	191.01973	1813	191.01987	0.75	191.01987	0.75	1 816
	M1	192.02309	1813	192.02324	0.79	192.02324	0.78	1 816
	M2	193.02644	1813	193.02399	12.70	193.02651	0.39	1 816
	M3	194.02980	1813	194.02984	0.25	194.02985	0.26	1 816
	M4	195.03315	1813	195.03317	0.12	195.03318	0.14	1 816
	M5	196.03651	1813	196.03662	0.57	196.03663	0.65	1 816
	M6	197.03986	1813	197.04026	2.03	197.04027	2.09	1 816
2/3-PG	M0	184.98566	1741	184.98571	0.25	184.98571	0.25	1 742
	M1	185.98902	1741	185.98908	0.36	185.98908	0.37	1 742
	M2	186.99237	1741	186.99010	12.13	186.99269	1.70	1 741
	M3	187.99573	1741	187.99577	0.23	187.99577	0.24	1 741
PEP	M0	166.97510	1889	166.97509	0.04	166.97509	0.04	1 890
	M1	167.97846	1889	167.97847	0.11	167.97847	0.11	1 890
	M2	168.98181	1889	168.98178	0.17	168.98190	0.54	1 890
	M3	169.98517	1889	169.98525	0.51	169.98524	0.45	1 890
Gly-3P	M0	171.0064	514	171.00642	0.10	171.00642	0.14	516
	M1	172.00976	514	172.00985	0.54	172.00985	0.54	516
	M2	173.01311	514	173.01318	0.38	173.01326	0.88	516
	M3	174.01647	514	174.01653	0.37	174.01655	0.47	516
PRPP	M0	388.94454	2154	388.94404	1.29	388.94404	1.28	2 162
	M1	389.94790	2154	389.94740	1.27	389.94740	1.27	2 162
	M2	390.95125	2154	390.95069	1.42	390.95067	1.48	2 160
	M3	391.95461	2154	-	-	391.95395	1.66	2 161
	M4	392.95796	2154	392.95726	1.78	392.95728	1.72	2 159

	M5	393.96132	2154	-	-	393.96043	2.24	2 158
P5P	M0	229.01188	1087	229.01215	1.17	229.01215	1.18	1 090
	M1	230.01524	1087	230.01552	1.25	230.01552	1.22	1 089
	M2	231.01859	1087	231.01755	4.50	231.01889	1.29	1 087
	M3	232.02195	1087	232.02228	1.43	232.02227	1.42	1 088
	M4	233.02530	1087	233.02564	1.45	233.02564	1.45	1 087
	M5	234.02866	1087	234.02907	1.79	234.02907	1.79	1 086
FBP	M0	338.98877	2082	338.98787	2.66	338.98784	2.73	2 087
	M1	339.99213	2082	339.99120	2.73	339.99120	2.72	2 087
	M2	340.99548	2082	340.99223	9.53	340.99454	2.76	2 087
	M3	341.99884	2082	341.99669	6.28	341.99769	3.34	2 087
	M4	343.00219	2082	343.00114	3.07	343.00114	3.05	2 087
	M5	344.00555	2082	344.00438	3.39	344.00436	3.43	2 087
	M6	345.00890	2082	345.00784	3.08	345.00782	3.12	2 087
Sed7P	M0	289.03301	1194	289.03281	0.68	289.03282	0.67	1 191
	M1	290.03637	1194	290.03619	0.59	290.03618	0.64	1 191
	M2	291.03972	1194	291.03719	8.69	291.03956	0.54	1 190
	M3	292.04308	1194	292.04291	0.56	292.04291	0.56	1 189
	M4	293.04643	1194	293.04627	0.53	293.04627	0.55	1 189
	M5	294.04979	1194	294.04965	0.45	294.04965	0.45	1 189
	M6	295.05314	1194	295.05189	4.22	295.05190	4.21	1 188
	M7	296.05650	1194	296.05520	4.37	296.05513	4.60	1 187
Man6P	M0	259.02244	1022	259.02288	1.70	259.02287	1.64	1 023
	M1	260.02580	1022	260.02629	1.90	260.02629	1.90	1 022
	M2	261.02915	1022	261.02963	1.82	261.02968	2.03	1 022
	M3	262.03251	1022	262.03302	1.98	262.03303	2.01	1 022
	M4	263.03586	1022	263.03640	2.04	263.03640	2.04	1 022
	M5	264.03922	1022	264.03962	1.52	264.03965	1.64	1 021
	M6	265.04257	1022	265.04312	2.08	265.04312	2.08	1 020
F6P	M0	259.02244	953	259.02283	1.49	259.02283	1.51	957
	M1	260.02580	953	260.02628	1.88	260.02629	1.89	956
	M2	261.02915	953	261.02954	1.50	261.02969	2.07	955
	M3	262.03251	953	262.03294	1.67	262.03285	1.30	952
	M4	263.03586	953	263.03642	2.14	263.03637	1.92	953
	M5	264.03922	953	264.03965	1.65	264.03966	1.67	953
	M6	265.04257	953	265.04310	2.01	265.04311	2.02	954
G6P	M0	259.02244	892	259.02281	1.41	259.02281	1.41	890
	M1	260.02580	892	260.02622	1.62	260.02624	1.72	889
	M2	261.02915	892	261.02832	3.16	261.02965	1.92	888
	M3	262.03251	892	262.03283	1.23	262.03283	1.22	888
	M4	263.03586	892	263.03640	2.05	263.03640	2.06	888
	M5	264.03922	892	264.03978	2.13	264.03977	2.10	887
	M6	265.04257	892	265.04305	1.83	265.04306	1.84	887
G1P	M0	259.02244	521	259.02295	1.96	259.02293	1.88	523
	M1	260.02579	521	260.02631	2.02	260.02631	2.01	523
	M2	261.02914	521	261.02974	2.31	261.02979	2.49	523
	M3	262.03249	521	-	-	262.03309	2.28	521
	M4	263.03584	521	-	-	263.03653	2.64	521
	M5	264.03919	521	264.03998	2.98	264.03996	2.90	523
	M6	265.04254	521	-	-	-	-	
ADP	M0	426.02214	2046	-	-	-	-	2 051
	M1	427.02550	2046	427.02544	0.12	427.02544	0.12	2 051
	M2	428.02885	2046	428.02845	0.94	428.02864	0.48	2 051
	M3	429.03221	2046	429.03199	0.49	429.03199	0.49	2 051
	M4	430.03556	2046	430.03533	0.54	430.03533	0.54	2 051
	M5	431.03892	2046	431.03867	0.56	431.03868	0.56	2 051
	M6	432.04227	2046	432.04203	0.56	432.04203	0.56	2 051
	M7	433.04563	2046	433.04547	0.36	433.04547	0.36	2 051
	M8	434.04898	2046	-	-	434.04849	1.14	2 051
	M9	435.05234	2046	-	-	435.05183	1.17	2 051

	M10	436.05569	2046	-	-	-	-	
ATP	M0	505.98847	2184	505.98900	1.06	505.98900	1.05	2 187
	M1	506.99183	2184	506.99234	1.02	506.99234	1.02	2 187
	M2	507.99518	2184	507.99487	0.62	507.99573	1.08	2 187
	M3	508.99854	2184	508.99885	0.61	508.99900	0.92	2 187
	M4	510.00189	2184	510.00232	0.84	510.00234	0.88	2 187
	M5	511.00525	2184	511.00559	0.68	511.00564	0.78	2 187
	M6	512.00860	2184	512.00900	0.79	512.00906	0.89	2 187
	M7	513.01196	2184	513.01234	0.74	513.01225	0.58	2 187
	M8	514.01531	2184	514.01565	0.65	514.01563	0.63	2 187
	M9	515.01867	2184	515.01883	0.32	515.01883	0.32	2 187
	M10	516.02202	2184	-	-	516.02078	2.41	2 185
CDP	M0	402.01090	1818	-	-	-	-	
	M1	403.01426	1818	403.01388	0.92	-	-	
	M2	404.01761	1818	404.01722	0.97	404.01727	0.84	1 818
	M3	405.02097	1818	405.02057	0.97	405.02060	0.91	1 818
	M4	406.02432	1818	406.02389	1.07	406.02388	1.07	1 818
	M5	407.02768	1818	407.02723	1.10	407.02723	1.08	1 818
	M6	408.03103	1818	408.03056	1.14	408.03059	1.08	1 818
	M7	409.03439	1818	-	-	409.03367	1.75	1 817
	M8	410.03774	1818	-	-	-	-	
	M9	411.04110	1818	-	-	-	-	
CTP	M0	481.97723	2106	481.97750	0.56	481.97752	0.60	2 116
	M1	482.98059	2106	482.98099	0.83	482.98099	0.84	2 117
	M2	483.98394	2106	483.98266	2.65	483.98428	0.69	2 116
	M3	484.98730	2106	484.98749	0.41	484.98751	0.44	2 116
	M4	485.99065	2106	485.99090	0.51	485.99088	0.47	2 116
	M5	486.99401	2106	486.99424	0.48	486.99422	0.44	2 116
	M6	487.99736	2106	487.99754	0.37	487.99754	0.37	2 116
	M7	489.00072	2106	489.00088	0.34	489.00089	0.35	2 116
	M8	490.00407	2106	490.00414	0.14	490.00412	0.11	2 116
	M9	491.00743	2106	-	-	491.00621	2.47	2 111
GDP	M0	442.01705	2258	442.01707	0.03	-	-	
	M1	443.02040	2258	443.02067	0.61	443.02067	0.61	2 266
	M2	444.02375	2258	444.02356	0.42	444.02373	0.04	2 266
	M3	445.02710	2258	445.02709	0.02	445.02708	0.03	2 266
	M4	446.03045	2258	446.03040	0.11	446.03039	0.14	2 266
	M5	447.03380	2258	447.03374	0.14	447.03374	0.14	2 266
	M6	448.03715	2258	448.03708	0.15	448.03708	0.15	2 266
	M7	449.04050	2258	449.04039	0.25	449.04039	0.26	2 262
	M8	450.04385	2258	-	-	450.04372	0.28	2 260
	M9	451.04720	2258	-	-	451.04674	1.02	2 259
	M10	452.05055	2258	-	-	-	-	
UDP	M0	402.99492	2202	402.99449	1.07	-	-	
	M1	403.99826	2202	403.99806	0.50	403.99805	0.51	2 208
	M2	405.00160	2202	405.00110	1.23	405.00128	0.78	2 208
	M3	406.00494	2202	406.00457	0.91	406.00457	0.90	2 208
	M4	407.00828	2202	407.00797	0.76	407.00797	0.75	2 207
	M5	408.01162	2202	408.01130	0.77	408.01128	0.84	2 208
	M6	409.01496	2202	409.01466	0.73	409.01466	0.73	2 208
	M7	410.01830	2202	410.01766	1.55	410.01754	1.85	2 203
	M8	411.02164	2202	411.02089	1.81	411.02095	1.68	2 205
	M9	412.02498	2202	-	-	412.02337	3.92	2 203
UMP	M0	323.02859	1931	323.02754	3.25	323.02753	3.29	1 936
	M1	324.03195	1931	324.03099	2.94	324.03099	2.95	1 936
	M2	325.03530	1931	325.03311	6.73	325.03430	3.08	1 936
	M3	326.03866	1931	326.03607	7.93	326.03760	3.25	1 935
	M4	327.04201	1931	327.04089	3.41	327.04090	3.40	1 935
	M5	328.04537	1931	328.04421	3.53	328.04421	3.53	1 935
	M6	329.04872	1931	329.04755	3.56	329.04755	3.55	1 935

	M7	330.05208	1931	330.05090	3.57	330.05090	3.56	1 935
	M8	331.05543	1931	331.05424	3.60	331.05424	3.58	1 935
	M9	332.05879	1931	332.05715	4.91	332.05708	5.13	1 934
UTP	M0	482.96125	2208	482.96167	0.86	482.96167	0.87	2 209
	M1	483.96460	2208	483.96510	1.03	483.96510	1.04	2 210
	M2	484.96795	2208	484.96781	0.29	484.96837	0.86	2 210
	M3	485.97130	2208	485.97146	0.34	485.97165	0.73	2 209
	M4	486.97465	2208	486.97495	0.61	486.97497	0.65	2 209
	M5	487.97800	2208	487.97833	0.68	487.97834	0.70	2 209
	M6	488.98135	2208	488.98168	0.68	488.98169	0.69	2 209
	M7	489.98470	2208	489.98491	0.43	489.98492	0.44	2 209
	M8	490.98805	2208	490.98793	0.25	490.98792	0.27	2 210
	M9	491.99140	2208	-	-	491.99023	2.38	2 203

Figure S-4. Impact of parameter optimization on the measurement of isotopologue abundances. Comparison of the integrated areas of benchmark isotopologues (red bars) and reference isotopologues (black line) in a log scale.

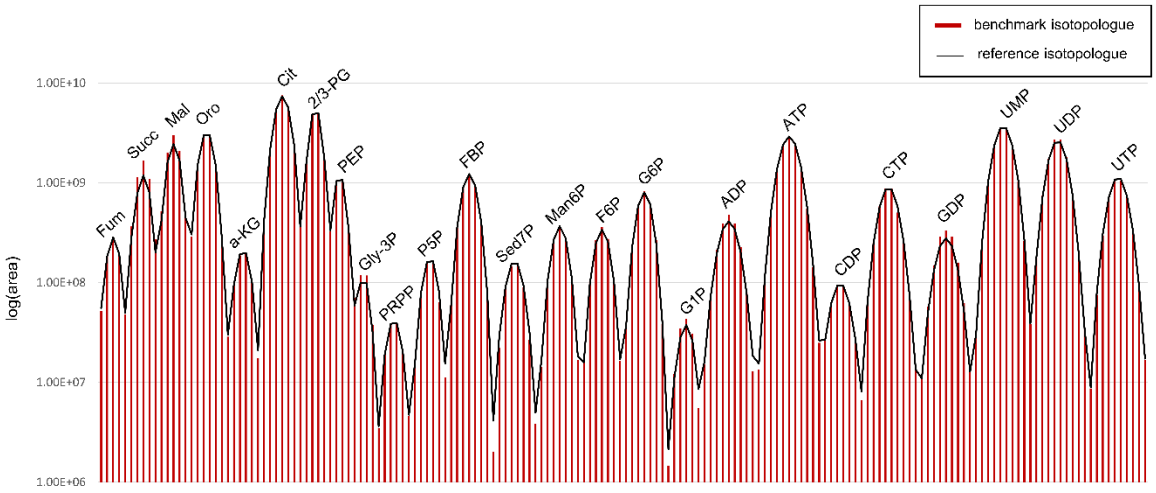
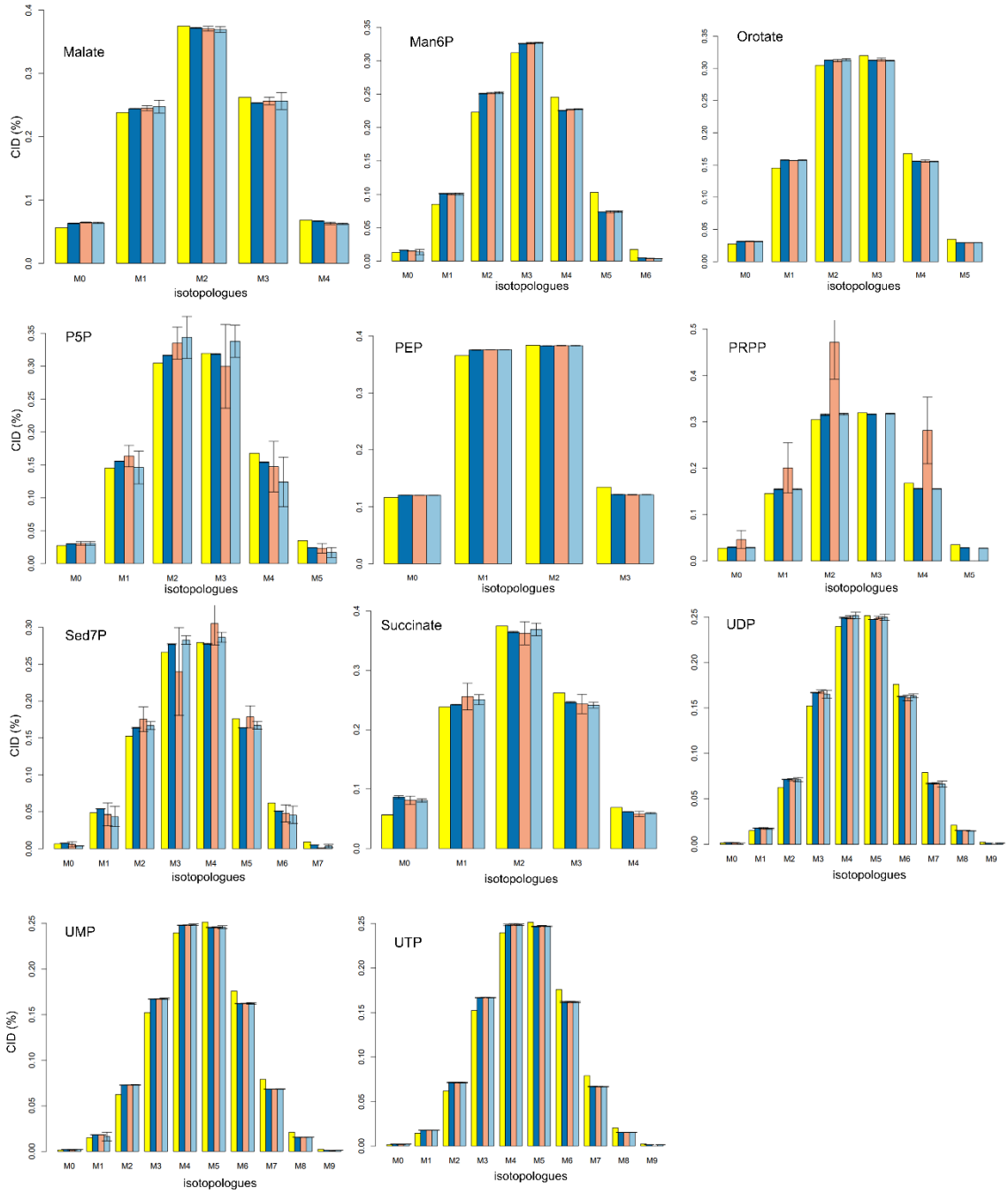


Figure S-5. CIDs comparison between optimized and non-optimized datasets. CIDs comparison between theoretical distribution (yellow), *reference CIDs* (dark blue), IPO *benchmark CIDs* (light red) and optimized *benchmark CIDs* (light blue) for the 25-*reference metabolites.*



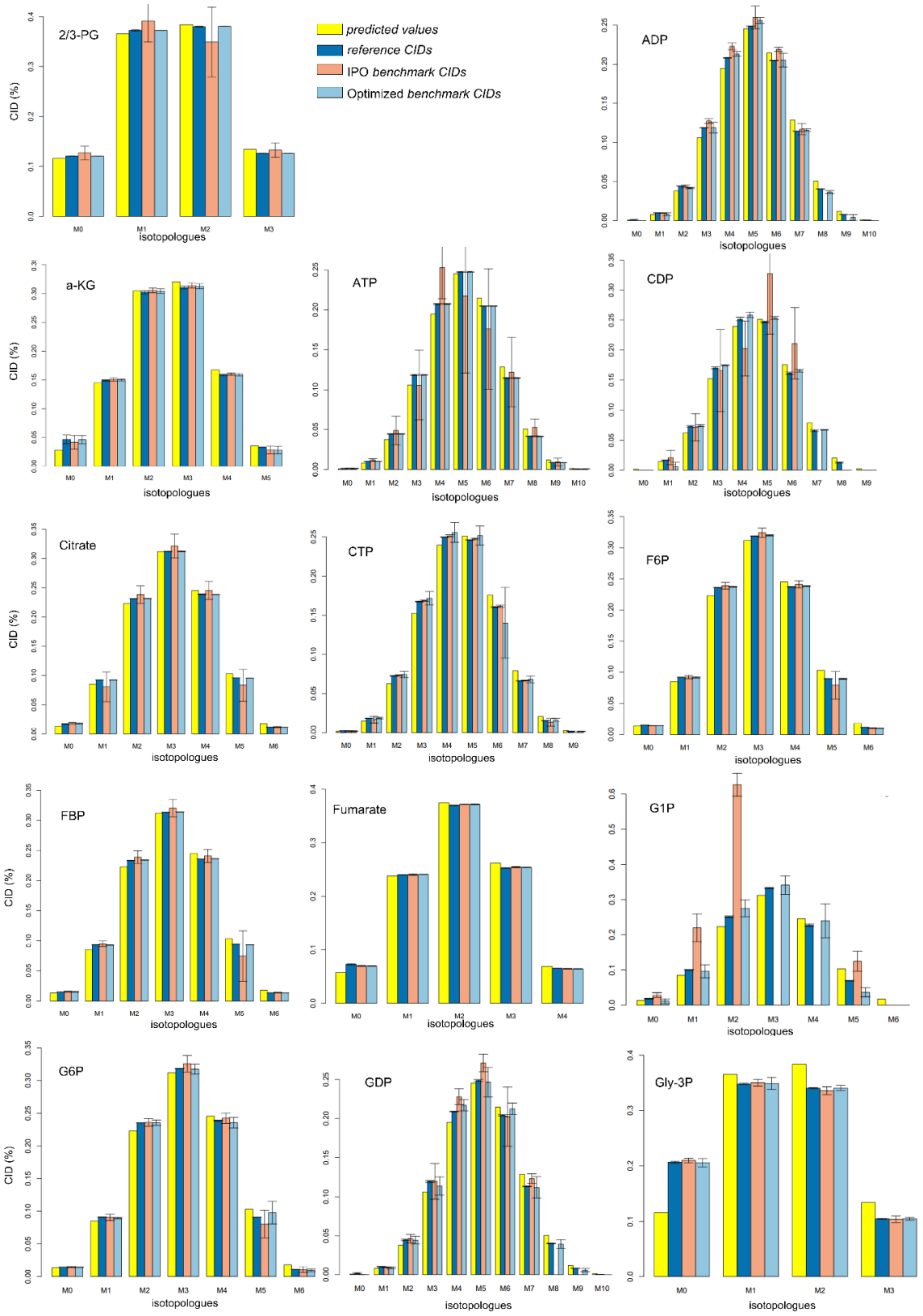

predicted values
reference CIDs
IPO benchmark CIDs
Optimized benchmark CIDs
2/3-PG
ADP
a-KG
ATP
CDP
Citrate
CTP
F6P
FBP
Fumarate
G1P
G6P
GDP
Gly-3P
CID (%)
isotopologues

Figure S-6. Clustering software redundancies after manual curation. Comparison of the number of isotopic clusters each detected isotopologue appears in for X13CMS (blue) and geoRge (red) software after manual curation of obvious redundancies.

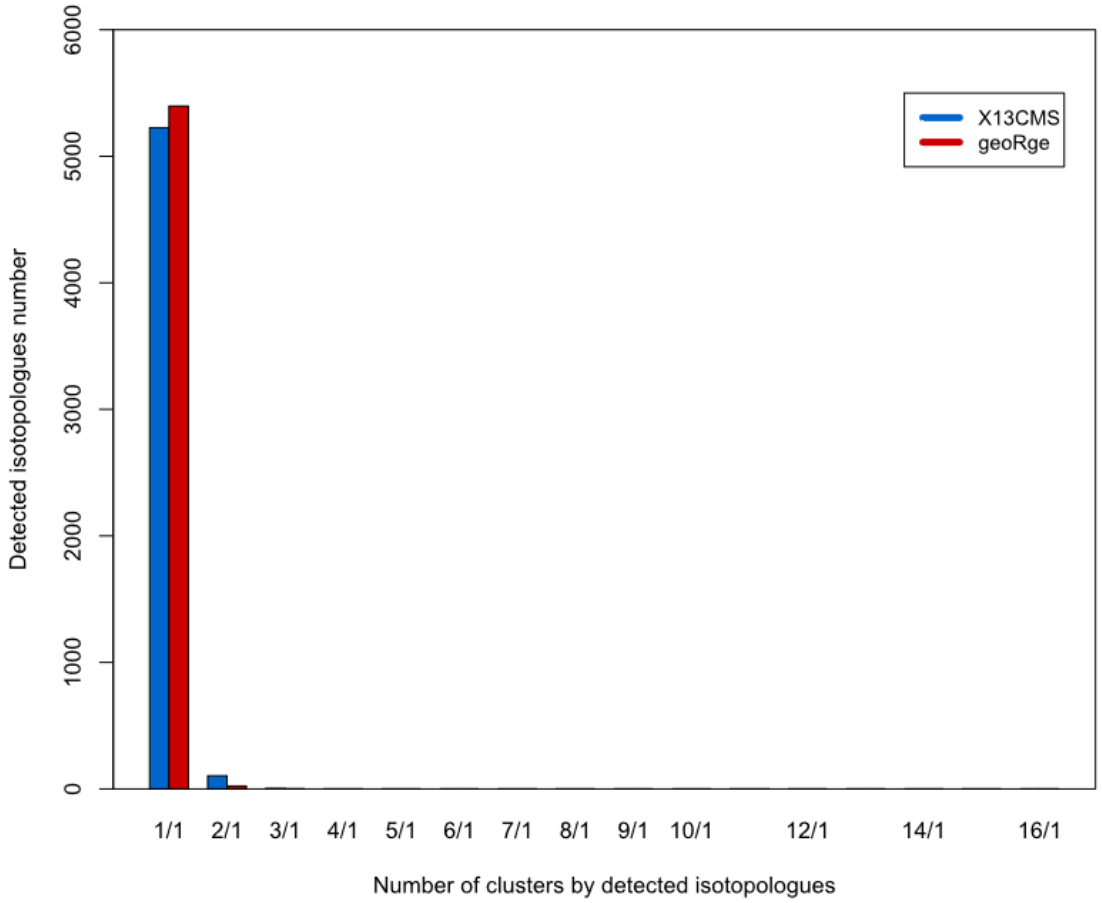
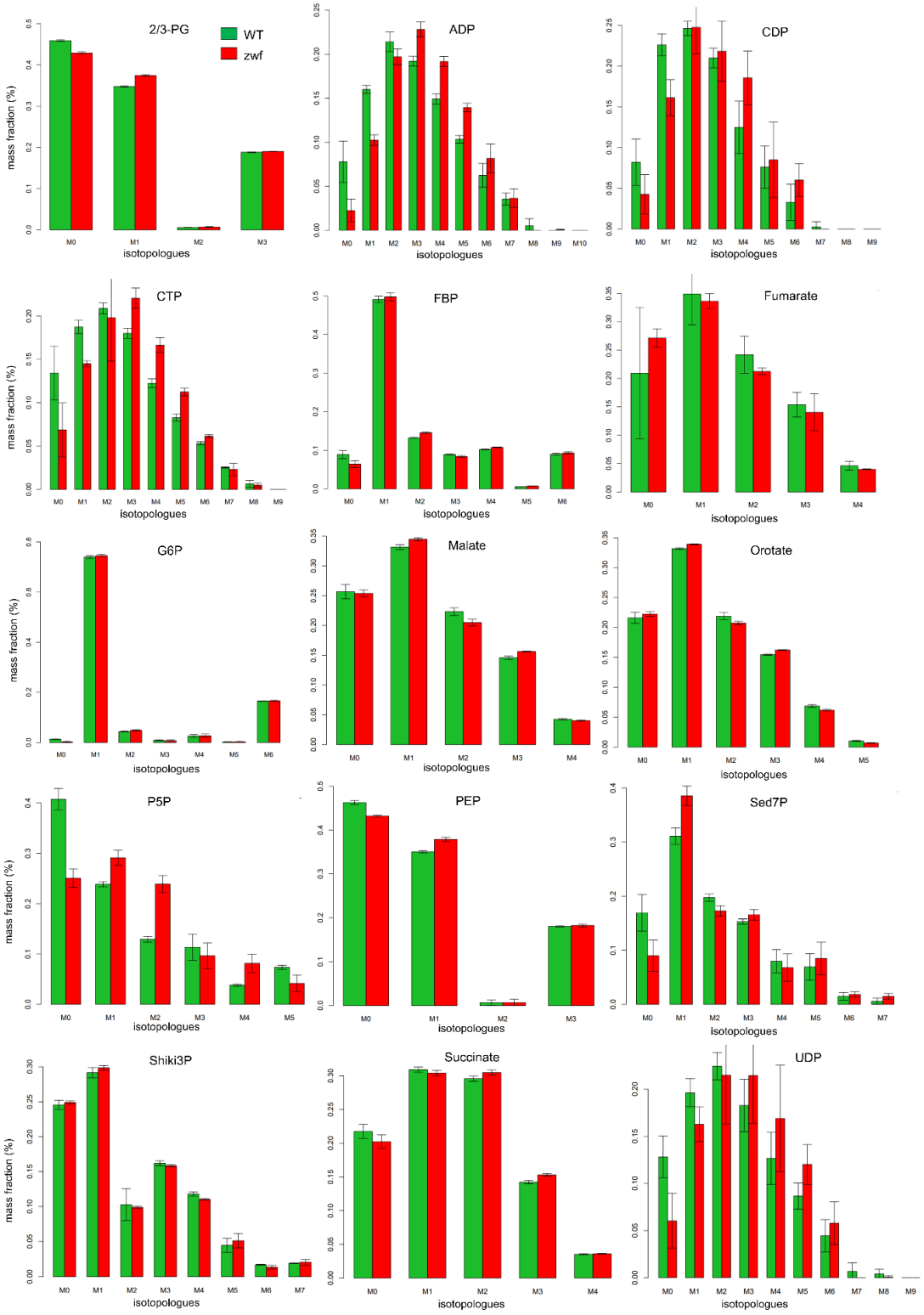


Table S-6. Cluster precision and recall for clustering software. Cluster precision and recall for X13CMS and geoRge with a RT window of 5 s and different isotopologue mass deviations, evaluated for the 25 reference metabolites in the PT sample.

	Isotopologue mass deviation (ppm)	1	2	3	5	8	10
X13CMS	precision	60%	76%	76%	84%	80%	80%
	recall	100%	100%	100%	100%	100%	100%
geoRge	precision	60%	68%	72%	80%	74%	72%
	recall	100%	100%	100%	100%	100%	100%

Figure S-7. Carbon mass fractions for significant metabolites between two *E.coli* strains. Carbon mass fractions comparison between the wild-type (green) and the Δzwf strains (red) for significant metabolites identified with a level 1 confidence.



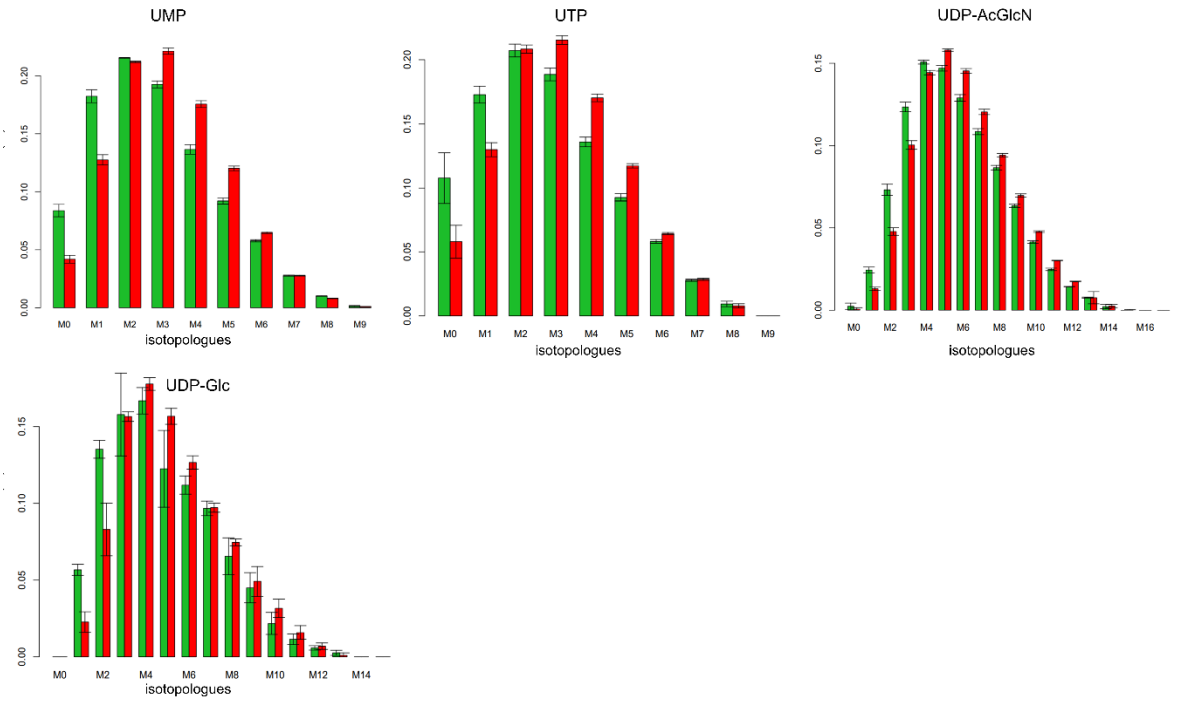


Table S-7. Overlapping isotopic clusters. The m/z and RT pairs of the 797 overlapping isotopic clusters between X13CMS and geoRge software. The online version contains supplementary material available at https://doi.org/10.1007/s11306-022-01897-5

Chapitre 3

Data-driven ^{13}C-fluxomics towards *ab initio* reconstruction of metabolic networks

Noémie Butin[1,2,3], Pierre Millard[4], Clément Frainay[4], Loïc Legregam[2,3], Fabien Jourdan[3,4], Uwe Schmitt[5], Floriant Bellvert[2,3], Patrick Kiefer[5], Jean-Charles Portais[1,2,3]

[1] RESTORE, CNRS ERL5311, EFs, ENVT, Inserm U1031, UPS, Université de Toulouse, Toulouse, France

[2] Toulouse Biotechnology Institute, TBI-INSA de Toulouse INSA/CNRS 5504-UMR INSA/INRA 798, 5504 Toulouse, France

[3] MetaboHUB-MetaToul, National Infrastructure of Metabolomics and Fluxomics, 31077 Toulouse, France

[4] INRAE UMR1331 Toxalim, 180 chemin de Tournefeuille St-Martin-du-Touch, BP 3 31931 TOULOUSE CEDEX France

[5] ETH Zürich, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland

Manuscrit in preparation

Glossary

Local label input: the substrate labeling dynamics in a minimal subnetwork.

Minimal subnetwork: minimal set of reactions that contains sufficient data required to simulate the labeling dynamics of a given metabolite, *i.e* (i) the reactions that produce it with tracer atom transitions and (ii) the labeling dynamics of its substrate (LLIs).

Complete minimal subnetwork: minimal subnetwork for which all local label inputs are available.

Incomplete minimal subnetwork: minimal subnetwork in which one or more reaction(s) is not conserved during the construction process.

Empty subnetwork: subnetwork containing only a sink reaction.

Minimal identifiable subnetwork: a minimal subnetwork is defined as identifiable if the data available are sufficient to estimate and provide a unique solution of fluxes.

Abbreviations

α-KG : alpha-ketoglutarate

AKGDH : oxogluterate dehydrogenase

FBA : Flux Balance Analysis

GLNS : glutamine synthetase

GLUDy : glutamate dehydrogenase

GLUSy : glutamate synthase

GSAM : Genome Scale Atom Mapping

GSM : Genome scale model

GSMN : Genome scale metabolic network

(HR)MS : (High resolution) mass spectrometry

ICDHyr : Isocitrate dehydrogenase

LC : Liquid chromatography

LLI : Local label input

Mal : Malate

NMR : Nuclear magnetic resonnance

OAA : Oxaloacetate

ODE : Ordinary differential equation

PEP : Phosphoenolpyruvate

Pyr : Pyruvate

SBML : System Biology Markup Language

1. Introduction

Systems biology is the study of how the components of a living organism interact with each other and with their environment to give rise to biological functions. The reconstruction of metabolic networks is a key step in systems biology that it provides an in-depth understanding of the molecular mechanisms of a particular organism. Its analysis allows to investigate the features that identify the topology of a metabolic network and the relative activities of metabolic pathways. These informations allows to identify the functional capacities of the organism, such as, its operation, or its adaptation and regulation in response to metabolic perturbations. It can be applied as a large variety of applications from biotechnologies to increase the production of a compound of interest to health for therapeutic screening [Gu et al, 2019; Cesur et al, 2020].

The first reconstruction approaches are based on the genome annotation [Edwards and Palsson, 2000]. In brief, on the basis of available genomic data and using homology studies, the biochemical reactions that can occur in an organism are identified [Chen et al, 2012]. However, genomic information does not consider the actual expression of each gene. The model based on genomic annotation can be curated using databases, such as KEGG or CheBI. Some « -omics » methods (transcriptomic, proteomic, metabolomic of fluxomic) can also be used to reaffine the genome scale metabolic network by providing data and describing metabolic pathways in specific organisms and conditions [Ryu et al, 2015]. However, this approach to metabolic reconstruction is limited as it is a time-consuming and indirect process to identify the active network in a given context. Moreover, it strongly relies on the correct annotation of the genome.

An alternative approach is to reconstruct an *ab initio* metabolic network based on experimental data alone, to identify the active reactions in an organism. Breitling et al, proposed the first reconstruction of metabolic networks based only on the observed metabolites obtained using unbiased metabolomics and independently of genome annotation [Breitling et al, 2006]. But they are far to be generic and are generally limited to analysis of a few topological motifs. Here, we propose a novel *ab initio* approach for metabolic network reconstruction based on non-targeted isotopic labeling dynamics experiments, in a way that is orthogonal and complementary to *in silico* reconstruction approaches. The dynamics of label propagation in a stationary metabolic network during an isotope labeling experiment can provide highly valuable information on the network topology, metabolic fluxes, and on metabolite pool sizes. In this work, we initiated this approach by developing a data-driven ^{13}C-fluxomic approach. It is based on untargeted labeling dynamics data obtained using high resolution-mass spectrometry and on

strategies of label propagation simulations and flux calculation to identify network topologies that can explain experimental data.

2. Results

2.1. General strategy

The general strategy of the proposed approach, named Isotopic driven Metabolic reconstruction (IsoMet) approach, is illustrated in the Figure 3.1. It is based on dynamics ^{13}C-labeling experiments in which the incorporation of the label in the metabolites can be monitored over time. Following automated data extraction and annotation, metabolic subnetworks are constructed for each metabolite, and the consistency of the assessed topology is tested by fitting their time-course label incorporation. The core of the approach consists of constructing – for each metabolite in the data – the metabolic subnetwork in which the labeling dynamics will be fitted to assess whether the network topology is consistent with experimental data. The proposed strategy aims to obtain a set of active metabolic subnetworks for an organism.

The strategy includes the following steps : (i) an experimental part to perform instationnary ^{13}C-labeling experiments, untargeted profiling of intracellular metabolites and isotopic data processing, and (ii) a mathematical part containing atom mapping and construction and evaluation of metabolic subnetworks.

First of all, this approach is based on instationnary ^{13}C-labeling experiments. In these types of experiments, the organism of interest is grown in a medium containing an unlabeled carbon source. After the cells reach the metabolic steady state, the culture medium is switched from unlabeled to ^{13}C-labeled medium. Samples are taken at different times to monitor the incorporation of the isotopic tracer in the intracellular metabolites, which are progressively labeled as a function of metabolite concentrations and fluxes. In order to maximize the metabolome coverage and study the metabolic network on a larger scale, samples are analyzed with untargeted metabolomics approaches. The incorporation of the label can be followed using high-resolution mass spectrometry (HRMS). The advantages of using HRMS is its high sensitivity and its large metabolic coverage when its coupled with liquid chromatography (LC). LC/HRMS analyses provides features (mz, RT, intensity). In our work, the features must be extracted, clustered and annotated to provide a list of metabolites with complete labeling information and isotopologues distributions over time. This step of the process requires the use

of robust and automatic tools to ensure the quality of isotopic data for use in metabolic reconstruction [Butin et al, 2022].

The next steps in the process consists to test the ability of different network topologies to explain the observed labeling dynamics data, thereby driving the identification and reconstruction of active metabolic subnetworks. For that we follow the guidelines defined in ScalaFlux [Millard et al, 2020]. To simulate the label propagation of a metabolite, two types of informations are required: (i) the set of reactions that participate to its labeling dynamics with carbon atom transitions and (ii) the labeling dynamics of its substrate. To do this, subnetworks are necessary and must be constructed for each metabolite in the data. We defined as *minimal subnetwork* the smallest subnetwork that contains sufficient data required to simulate the labeling dynamics of a given metabolite. In contrast to classical fluxomics, the label propagation of the metabolite in the subnetwork is modeled directly from the labeling dynamics of its metabolic precursor(s), which are defined as *local label inputs* (LLIs).

The first step (Step 4 in Figure 3.1) consists on getting a list of potential metabolic reactions with carbon atom transitions. Since this strategy is orthogonal to genome-based reconstruction approaches, any available (bio)chemical and metabolic knowledge can be optionally included to support the construction process. This may include all possible known biochemical reactions and chemical rules from databases (as done in this study) or literature for any type of organism. The objective is to obtain the most complete list of reactions that could occur in a living system. This knowledge can also be enriched with orthogonal approaches such as the establishment of pairwise mass-mass difference from data [Breitling et al, 2006]. Then, minimal metabolic subnetworks are constructed for each metabolite in the dataset (Step 5 in Figure 3.1). The construction process is the essence of the approach and will be detailed in the next section (section 2.2). Once the minimal subnetwork has been obtained, we can use the scalaflux approach to simulate label propagation and fit the labeling dynamics of the metabolite of interest. This step (Step 6 in Figure 3.1) aims at evaluating the ability of subnetwork topology to explain the labeling pattern of a metabolite in the experimental data [Millard et al, 2020]. The goodness-of-fit between simulated and experimental labeling dynamics can be exploited to validate or exclude reactions in minimal subnetworks. When the data are accurately fitted, the flux values can be obtained. Although labeling dynamics are sufficient to fit experimental data and give access to the turnover of metabolites or to relative fluxes (in case of convergent reactions), measure of metabolites concentration can be optionally included to give access to absolute fluxes. A minimal subnetwork is defined as *identifiable* if the data available are sufficient to estimate and provide a unique solution of fluxes.

At the end of the process, a set of active metabolic subnetworks can be obtained. The proposed approach is iterative and can be refined or completed to identify gaps or validate sets of reactions by adjusting the experimental design. Although illustrated in this study with ^{13}C-isotopes and MS data, the proposed approach is generic and can be applied to different organisms using any isotopic tracer (e.g. ^{2}H, ^{15}N or ^{18}O) and analytical methods (NMR or MS/MS) according to the biological question and the context of the study.

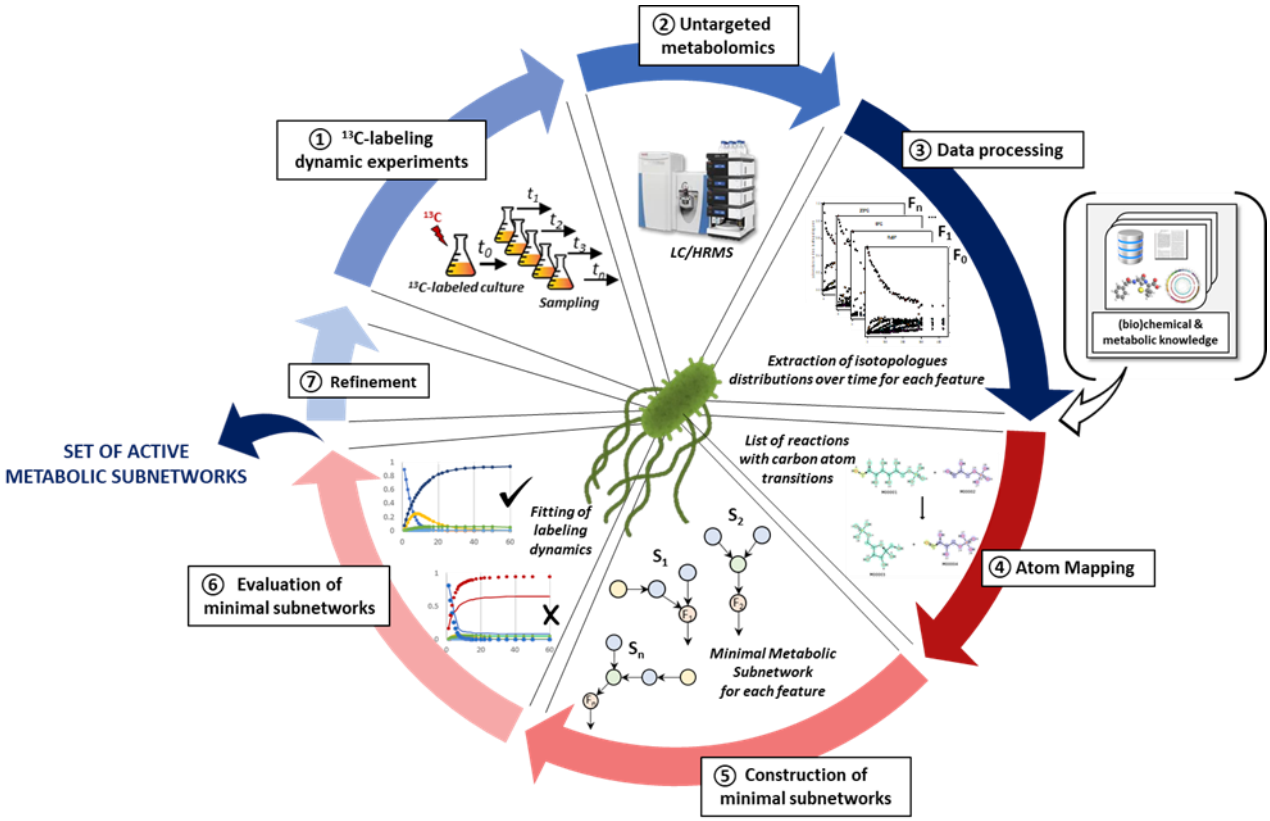


Figure 3.1 *: General strategy for ab initio metabolic subnetwork construction based on untargeted ^{13}C-labeling dynamic experiments.*

2.2.Construction of minimal subnetworks

The proposed strategy uses a set of available tools and methods for ^{13}C-labeling experiments and data extraction. Specific developments had to be made for some key steps such as generation of a list of reactions with carbon atom transition and the construction of subnetworks. The essence of the proposed approach lies in the way we construct and analyze metabolic subnetworks required to evaluate the consistency between the labeling dynamics of a given metabolite and a set of reactions. This section is focused on the step five of the workflow (Figure 3.1). Based on the list of reactions and metabolites, subnetworks are constructed for each metabolite in the dataset. From a given metabolite in the data, the subnetwork is extended iteratively until getting a minimal subnetwork containing all the data required to simulate. The extension is done by combining minimal subnetworks following the rules established in ScalaFlux [Millard et al, 2020]. We defined as *complete minimal subnetwork,* the minimal subnetwork for which all LLIs are available. An *incomplete minimal subnetwork* is a minimal subnetwork in which one or more reaction(s) is not conserved because of missing LLIs. Missing LLIs can be either that no information at all is available (*i.e*: the labeling dynamics of the substrate is not present in the dataset) or that at least one critical information is missing (*i.e:* we don't have access to the elementary metabolite unit which participates to the formation of the metabolite). This last case can occur in cleveage reaction.

The process is illustrated for the construction of a minimal subnetwork centered on a given metabolite Y (Figure 3.2A). First, a sink reaction consuming Y has to be included to avoid its accumulation, in keeping with the metabolic steady-state assumption. The reaction(s) and the LLIs which are involved in its production are searched in the list of reactions and the list of metabolites, respectively. Two cases can be distinguished. In the first case, all data are available (*i.e* the LLIs A and B producing Y via the reactions r_1,r_2). In the second case, information on the LLIs are missing.

The subnetwork is then extended by searching for an upstream substrate for the part of the subnetwork where LLI is missing. We define the extension levels according to the number of searches performed upstream of the metabolite of interest during the construction process. For each level of extension, the initial precursor becomes a metabolic intermediate. At each iteration step, we check if the subnetwork is minimal. If not, the subnetwork continues to be extended on the part of the network for which the LLIs are missing. The process of construction stops when all data are available or when the level of extension defined by the experimenter is reached. At the end of the process, the part of the subnetwork for which the LLIs are still

missing, is not conserved in the minimal subnetwork. For example, at the third level, information on the LLIs (F and E) are missing. The set of reactions (r_2, r_3, r_5, r_6) is deleted from the subnetwork (step 3 in Figure 3.2). If after several extensions, none LLIs could be identified, the resulting subnetwork is an *empty* subnetwork, containing only the sink reaction (step 4 in Figure 3.2).

Once the minimal subnetwork has been obtained, simulation of label propagation and fitting of experimental labeling dynamics of the metabolite of interest are performed, as detailed previously [Millard et al, 2020]. In the minimal subnetwork S_Y, the labeling dynamics of Y are directly simulated from the labeling dynamics of its precursors G, F and E (Figure 3.2B). The goodness-of-fit between the simulated and experimental data aims to determine if the topology of the minimal subnetwork is consistent with the observed labeling dynamics of the metabolite.

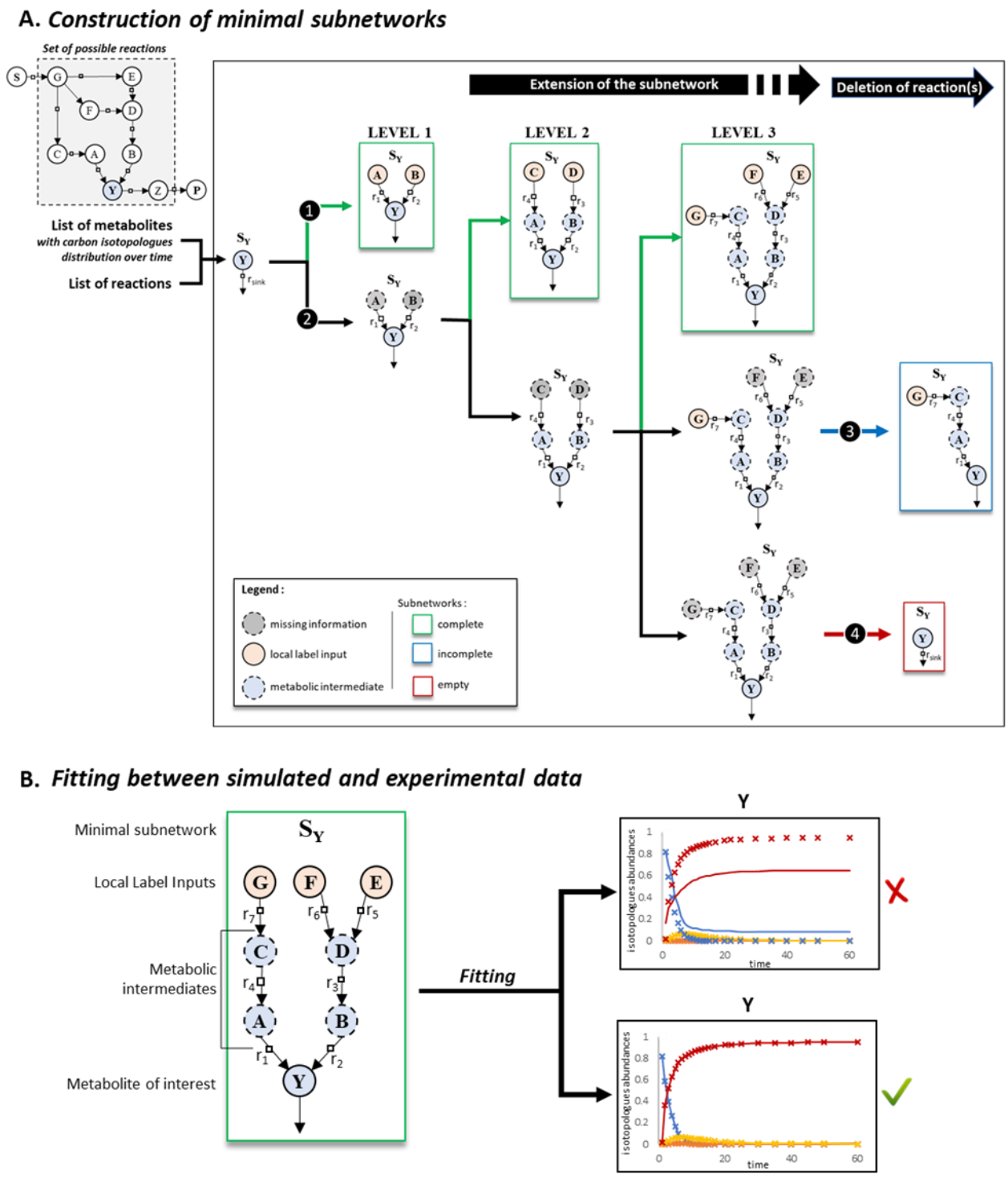


Figure 3.2 : Construction and evaluation of minimal subnetworks. *A. Subnetwork construction and extension process. Minimal subnetworks are constructed for each metabolite in the data by searching for the reactions and the LLIs that produce it. If all data are available, the minimal subnetwork is complete (in green). If not, it is extended (black arrows) until a minimal subnetwork containing all the data needed for the simulation is obtained. At the end of the process, the part of the subnetwork for which LLIs are still missing is not conserved. The resulting subnetworks may be incomplete (in blue) or empty (in red) depending on the final reactions included. B. The labeling dynamics of Y are directly simulated from the labeling dynamics of its precursors G, F and E. The fit between the simulated and experimental data is used to estimate if the experimental data can be explained by the proposed reactions.*

2.3.Implementation of IsoMet : Isotopic driven Metabolic reconstruction

The computational workflow for the IsoMet approach is divided into three steps that contains a set of tools developed to construct metabolic subnetworks from HRMS data (Figure 3.3).

First, the *Feature Collection* aims to provide a list of metabolites containing the labeling incorporation and isotopologues distributions over time from HRMS data obtained using untargeted analysis strategies. Some software specific for untargeted isotopic profiling studies are developed [Huang et al, 2014 ; Capellades et al, 2016 ; Bueschl et al, 2017 ; Dange et al, 2020] but there are still few available for integrate dynamic data [Kiefer et al, 2015, Dange et al, 2020]. This step of isotopic labeling dynamics extraction is performed using the DynaMet tool [Kiefer et al, 2015], a software developed on eMZed [Kiefer et al, 2013] for fully automated investigations of isotope labeling experiments from LC-HRMS raw data. DynaMet extracts metabolite labeling profiles from ^{13}C-time course labeling experiments. Special attention to data extraction is required. The use of a specifically labeled sample can help to ensure and validate the data quality [Butin et al, 2022].

Second, for the *Atom Mapping* step, we developed GSAM to perform Genome Scale Atom Mapping. As discussed above, many sources can be used to create a list of reactions. Initially, to develop the tools to implement the *ab initio* approach, we relied on a Genome Scale Model (GSM).

The atom mapping is a procedure that establishes a correspondence between the atoms of reactants and products [Litsa et al, 2018]. The results can be exploited to build carbon skeleton networks for topological analysis of genome-scale metabolic networks (GSMN), and analyse the fate of carbon atoms. GSAM requires a metabolic network in SBML (Systems Biology Markup Language) file format, which is a machine-readable format for representing models, and a table containing SMILES (Simplified molecular-input line-entry system) representing the chemical structure of the compounds in the model [Hucka et al, 2003]. From these input files, GSAM creates a list of reactions with consistent carbon atom transitions at the scale of the network. SBML is oriented towards describing systems where biological entities are involved in, and modified by, processes that occur over time. It can be found in various databases such as BiGG [King et al, 2016], BioModels [Glont et al, 2018] or Metexplore [Cottret et al, 2010]. Some reactions in SBML are not suited for atom mapping (such as biomass production) and GSAM tool provides flexible filtering utilities for adapting a GSMN to atom mapping tasks.

The cofactors and non-carbon compounds can be conserved as reactants to ensure the reaction stoechiometry and metabolic balancing but can be excluded for the atom mapping.

The *Network Analyzer* is the tool developed to construct and evaluate minimal subnetworks. Starting from the list of metabolites with carbon isotopologues distribution over time and the list of reactions, the Network Analyzer first constructs minimal subnetworks centered of each metabolite in the data (see section 2.2). Then, it simulates the label propagation in each minimal subnetwork according to the scalaflux approach [Millard et al, 2020] and fits the labeling dynamics of the metabolites by adjusting fluxes using *influx_i* [Sokol et al, 2012]. The goodness-of-fit is evaluated using a chi-square test [Antoniewicz et al, 2006]. Measures of metabolite concentrations can be optionally included and adjusted to fit experimental data and provide absolute flux measurements. The Network Analyzer returns evaluated subnetworks, *i.e.* minimal subnetworks and the results of simulations (see Methods for details). Finally, the interpretation of the goodness-of-fit can provide a set of active metabolic subnetworks in the studied organism.

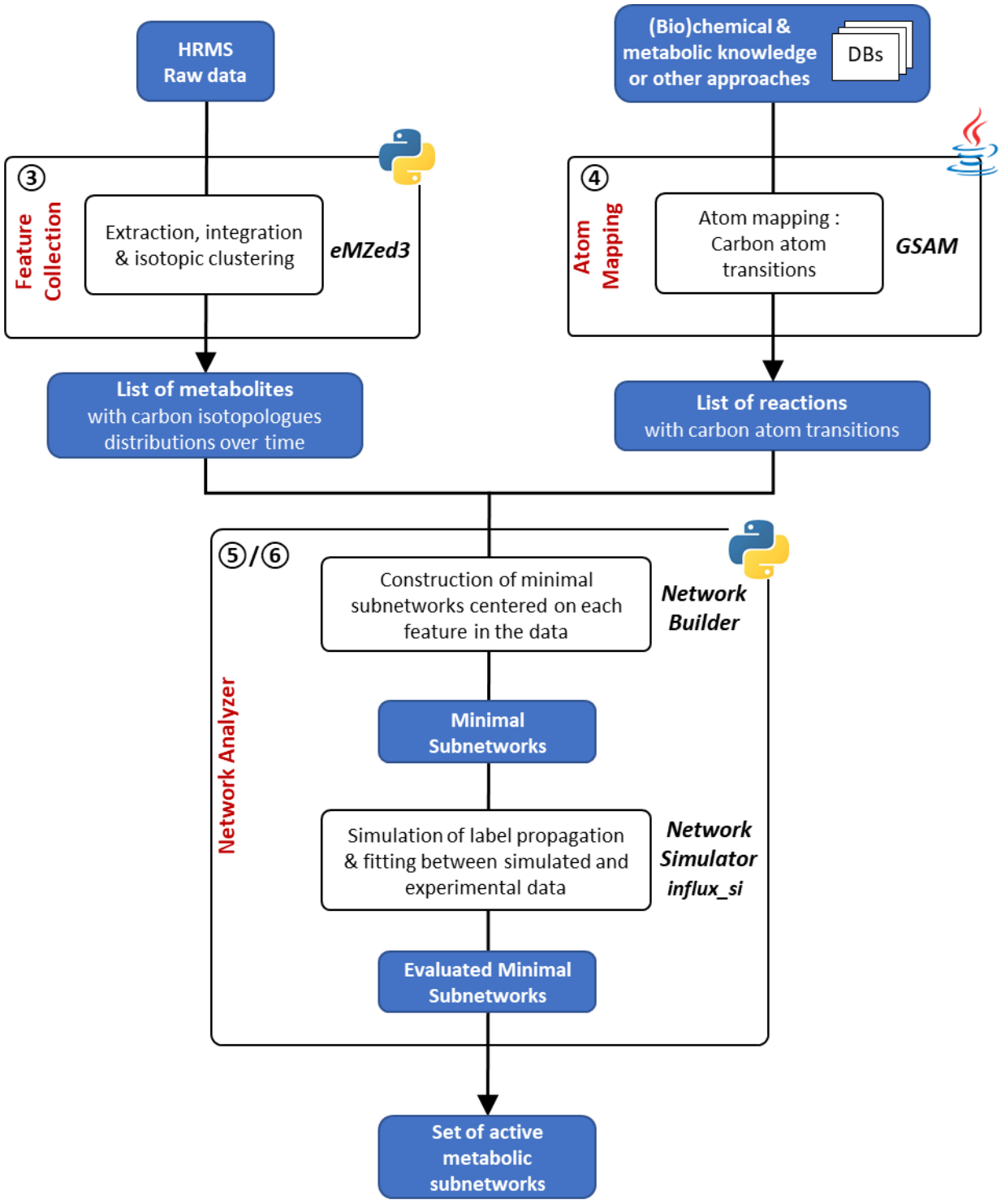


***Figure 3.3 : Computational workflow of IsoMet approach**. The IsoMet approach is composed of three steps to reconstruct a set of active metabolic subnetworks from HRMS data and (bio)chemical knowledge.*

2.4.In silico validation of the Network Analyzer

In this section we evaluate the core of the IsoMet approach: the construction of minimal subnetworks and the fit of labeling dynamics. The Network Analyzer was tested from a network obtained from a GSM and a theoretical dataset to validate its ability to construct minimal subnetworks and test if they can explain experimental labeling dynamics. The evaluation of the robustness of the approach is currently being assessed, so we will only focus here on the validation of the tool.

The validation process is illustrated in Figure 3.4A. The Network Analyzer was tested on a *Escherichia coli* core model from the BiGG database (http://bigg.ucsd.edu/models/e_coli_core), representative of a known metabolic network with a set of topological motif that occurs in a « real » metabolic network (such as condensation and cleavage reactions, cycles, mono and bimolecular reactions). The model was adapted to correspond to a classical *E. coli* culture under aerobic condition grown in a minimum medium with glucose as sole carbon source (Table S1) and converted into an isotopic model using GSAM. Metabolite concentrations and fluxes were initialized at the values listed in the Supporting information (Table S1) and label propagation through this network was simulated to create a theoretical dataset containing isotopologues distributions over time. From the list of reactions and the list of metabolites, the Network Analyzer was applied to construct minimal subnetworks and test if they can explain the theoretical isotopic labeling dynamics. The minimal subnetworks have been evaluated both on their topology and their functional characteristics. The topology of subnetworks is evaluated based on (i) the number of minimal subnetworks (the Network Analyzer is expected to construct a minimal subnetwork for each metabolite in the theoretical dataset) and (ii) their characteristics (complete or incomplete). Their functionnal characteristics are evaluated based on (i) the goodness-of-fit (it is based on the result value of the $\chi 2$ test. The fitting quality is considered as correct for a $\chi 2$ pvalue < 0.05), (ii) their identifiability and (iii) the compliance of flux values. This last parameter looks for the closeness between the theoretical fluxes used to simulate the theoretical dataset and the estimated fluxes measured in the minimal subnetworks. Fluxes are considered as accurate when the calculated difference is less than 10%. It can only be measured measured for identifiable subnetworks.

From the 32 metabolites in the theoretical dataset (see Section 4.1), 29 minimal subnetworks could be constructed with the Network Analyzer (Table 3.1). The three subnetworks centered on the sedoheptulose, the glyoxylate and the erythrose-4-phosphate do

not have sufficient data to obtain minimal subnetworks. Cleavage reactions directly contribute to the formation of these three metabolites and the corresponding LLIs are not available. Among the 29 constructed minimal subnetworks, 14 are complete and 16 are incomplete. For these last ones, some reactions were not conserved at the end of the construction process due to a missing LLI. There are two causes of missing LLIs (i) the CO_2 measurements are missing in the data and/or (ii) the elementary metabolite unit which participates to the formation of the metabolite is not known (because of a cleavage reaction). An interesting result is the "level 3 extended" minimal subnetwork centered on the pyruvate (S_{PYR}) (Figure 3.5B). At the first level, the phosphoenopyruvate (PEP) and the malate (Mal) are the direct LLIs of the pyruvate (Pyr). However, isotopes of the "$C_1C_2C_3$ block" of the malate are a missing LLI as the elementary metabolite unit producting the pyruvate is lost in cleavage reactions (ME1, ME2). The subnetwork was extended from this part of the subnetwork to identify the minimal subnetwork. At the third level, the information on the elementary metabolite unit involved in the production of the pyruvate is retrieved. The final LLI of this minimal subnetwork is the phosphoenopyruvate.

Table 3.1 : ***Results of validation process.*** *The minimal subnetworks have been validated both on their topology and their functionnal characteristics.*

		Evaluation criteria	Number	% of recovery
TOPOLOGY	①	**Total**	29	91
	②	**Complete**	14	44
		Incomplete	15	47
FUNCTIONNAL CHARACTERISTICS	③	**Goodness-of-fit**	21	66
	④	**Identifiability**	9	22
	⑤	**Compliance of flux values**	6	19

The labeling dynamics of metabolites are accurately fitted for 21 minimal subnetworks (Table 3.1 and Table S3). In case of incomplete minimal subnetworks (e.g: S_{DHAP}, Table S2 and Table S3), an accurate fit indicates that the reactions that are not conserved during the construction process do not strongly influence the labeling dynamics of the metabolite. For example, two disctinct minimal subnetworks are illustrated in the Figure 3.4C. The first one is the complete minimal subnetwork centered on the glutamine S_{GLN}. As shown in the fitting

results, the simulated labeling dynamics fit accurately the theoretical measurements. The label incorporation into this metabolite is slow. Indeed, glutamine, like glutamate, is known to have a particularly slow labeling dynamic, due to its high pool size [Bennett et al, 2009]. In opposition to the S_{GLN}, the minimal subnetwork centered on the alpha-ketoglutarate ($S_{\alpha\text{-}KG}$) is an incomplete subnetwork. The final subnetwork used to simulate label propagation is only composed of the Glutamine as metabolic precursor that produce the α-KG via the glutamate dehydrogenase ($GLUD_Y$). In this subnetwork, the topology of the subnetwork can not explain the isotopic dynamic ($\chi 2$ pvalue > 0.05), which is clearly visible on the graph.

Among the 29 minimal subnetworks, 9 are identified as identifiable subnetworks and provide a unique solution of fluxes. Regarding the estimated flux values, 6 of them are in good agreement with the theoretical fluxes used to simulate the theoretical dataset, with a mean accuracy of 5.9% +/- 3%. Regarding the S_{GLN}, (Figure 3.4C) this minimal subnetwork is an identifiable subnetwork and provides a unique solution of fluxes for the glutamine synthetase reaction (GLNS), that is on good agreement with the value used to simulate the theoretical dataset with a relative error of 0.87%. The good agreement between simulations and theoretical measurements indicate that the topology of the subnetwork is consistent with the isotopic data. The detailed results for all minimal subnetworks are provided in Supplementary Data (Table S3). For other subnetworks, the measurements provided are not sufficient to unambiguously determine fluxes.

This test case demonstrates the strength of our approach. The IsoMet approach aims to highlight active subnetworks in the organism. If a given minimal subnetwork fits accurately the experimental labeling dynamics, this highlights the active reactions involved in the production of the metabolite of interest, even if some reactions that contribute little to the observed dynamics are not included in the network. If not, the fit result can inform on the existence of potential other reactions that contribute significantly in the production of the metabolite but not included because of missing data. Typically, the glutamate dehydrogenase alone does not explain the labeling dynamics of the α-Kg labeling as other reactions are involved in α-KG production (Isocitrate dehydrogenase (ICDHyr), oxogluterate dehydrogenase (AKGDH) and glutamate synthase (GLUSy)) (Figure 3.4C).

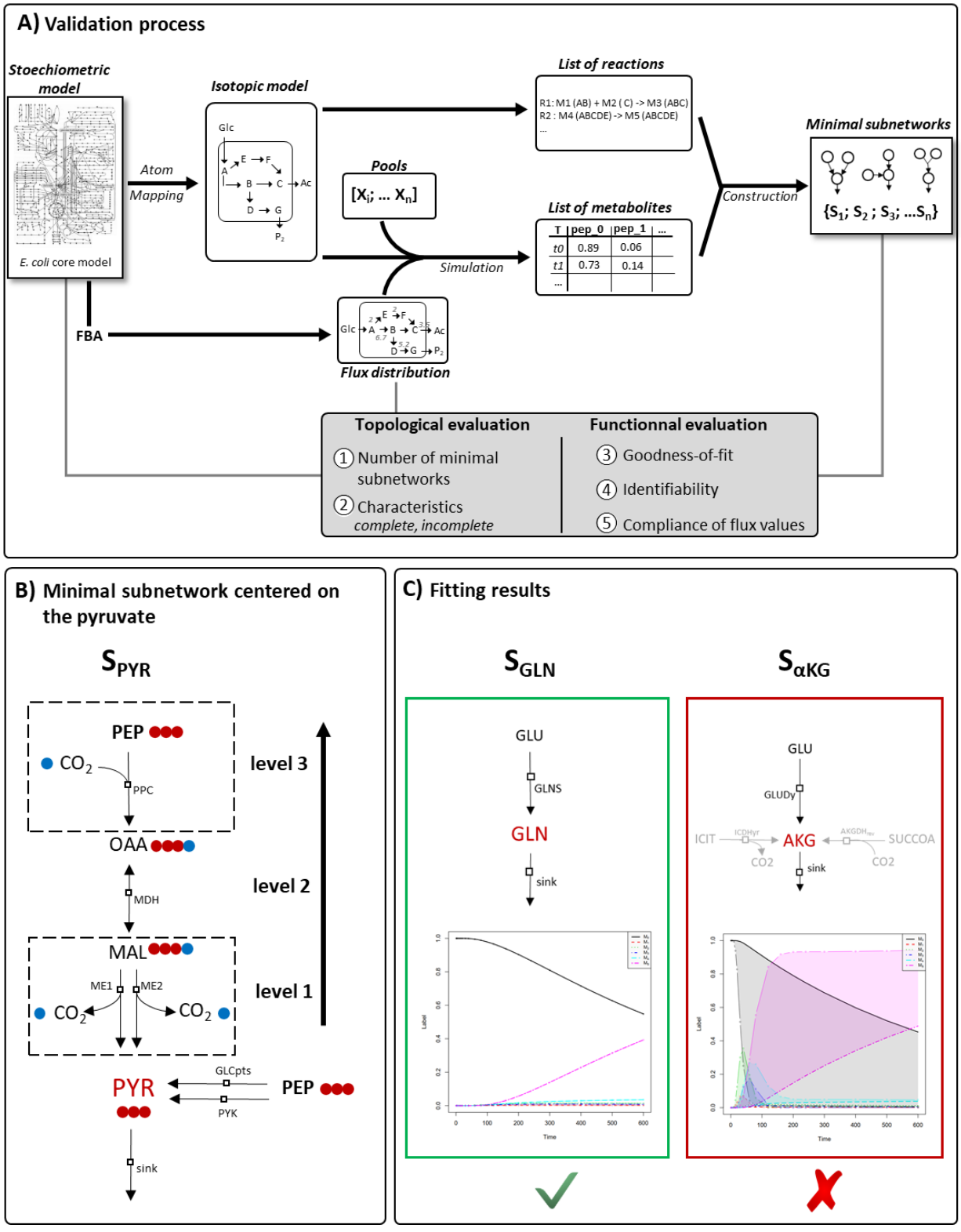


Figure 3.4 : ***Validation of the Network Analyzer.*** *A. Validation process. The Network Analyzer is validated using a GSM and a theoretical dataset based on topological and functionnal criteria. B. Minimal subnetwork centered on the pyruvate. The elementary metabolite unit producting the three atoms of carbon of the pyruvate is retrieve at the third level of extension. The final LLI of the pyruvate is the phosphoenolpyruvate. C. Fitting results of two minimal subnetworks (S_{GLN} and $S_{\alpha\text{-}KG}$). The reactions that were not conserved during the construction process are in grey.*

3. Discussion

In this work we proposed a data-driven ^{13}C-fluxomic approach based on dynamic isotope-labeling experiments to identify set of active subnetworks in a organism.

It's important to note that the relevance of resulting metabolic subnetworks depends on the quality of the experimental setup, including experimental design (choice of nutrient, isotopic tracer, analytical techniques …) and data processing (extraction of isotopic data) and annotation. To ensure the quality of isotopic data with untargeted strategies, robust tools and approaches are thus needed [Butin et al, 2022].

The tools developed for this work allow for direct access to other sources of information, such as analytical sources (MS/MS data) or in terms of biochemical knowledge. In this work we used a GSM as source of biochemical knowledge but many other sources can be implemented, such as a metaGSM [Zorilla et al, 2021], additional biological databases, chemical rules or literature. Further development towards *ab initio* fluxomics will require enriching this step with the *ab initio* generation of a reaction list, using e.g. the establishment of pairwise mass-mass difference from data [Breitling et al, 2006] or spectral correlations [Amara et al, 2022].

From a computational point of view, the proposed approach shares elements with the scalaflux approach to simulate label propagation, combine subnetworks and fit labeling dynamics [Millard et al, 2020]. The process of (sub)network extension allows incomplete datasets to be exploited to identify the active subnetworks. Interpretation of the fit between simulated and experimental data allows to evaluate and validate the topology of the metabolic subnetworks and identify the main (actives) reactions which are involved in the formation of a given metabolite. However, this does not mean that there are no other undescribed reactions that are not significantly active yet are involved in the production of this metabolite.

Currently, our approach allows to highlight the active parts of a metabolic network and requires more development and additionnal information to combine these subnetworks to reconstruct a complete network at the cellular level. Combination with traditional *in silico* approaches can be useful to complete metabolic pathways identification or to connect metabolic subnetworks. With respect to gaps in data, combination of different analytical techniques can also be valuable. For example, in LC/HMRS coupling, the use of complementary columns can cover another part of the metabolome.

Although illustrated in this study with ^{13}C-isotopes and MS data, the proposed approach is generic and can be applied using any isotopic tracer (e.g. ^{2}H, ^{15}N or ^{18}O) and analytical

methods (NMR or MS/MS). For example, relevant and additional information can be obtained by optimizing the labeling of the nutrient or using combination of multiple ^{13}C-carbon sources [Millard et al, 2014]. In a similar way, other analytical techniques can be used to obtain additional data, such as positionnal label information using NMR or MS/MS analyses, which can be helpful for specific topological motifs such as cleavage reactions. Our approach should also be useful to study complexes or poorly characterized organisms.

4. Methods

4.1. Generation of the theoretical dataset

The metabolic network implemented in the model was derived from the *E. coli* core model from the BiGG Database (http://bigg.ucsd.edu/models/e_coli_core). The model was adapted to correspond to a classical *E. coli* culture under aerobic condition grown in a minimum medium with glucose as sole carbon source and was converted into a isotopic model using GSAM. The final isotopic model contains 58 biochemical reactions (including the biomass equation), 32 intracellular metabolites and all the stoechiometric reactions of the metabolic network as well as the carbon transitions between substrates and products of each reactions. The label propagation was simulated using *influx_i* [Sokol et al, 2012] to create a theoretical dataset with isotopologues distribution over time for the 32 intracellular metabolites. The set of initial flux values used to simulate the data was calculated using an *in silico* Flux Balance Analysis (FBA) based on Enjalbert et al., 2017 to optimize the production of biomass. The constraints on extracellular fluxes were fixed with a consumption of glucose of -8.7 ± 0.1 mmol.g_{DW}^{-1}.h^{-1} and a production of acetate of -1.5 ± 0.1 mmol.g_{DW}^{-1}.h^{-1}. The time points used to simulate the labeling dynamics were the following : t1, t2, t5, t10, t15, t20, t30, t40, t60, t80, t120, t160, t200, t240, t300, t400, t500 and t600 seconds and the metabolites pools were provided by Bennett et al., 2009.

4.2. GSAM

GSAM was developed with JAVA language and performs Genome Scale Atom Mapping. The results can be exploited to build carbon skeleton networks for topological analysis of genome-scale metabolic networks (GSMN), and analyse atom tracking results. GSAM is a built around the Reaction Decoder Toolkit, a library which compute atom mappings

from reactions represented as RxnSmiles. It uses the JSBML library to import a GSMN in standard format, then automatically generate all the RxnSmiles from provided compounds structures, launch the atom mapping and finally parse the results to render meaningful substrate-products transitions. GSAM was applied to create a list of reactions with carbon atom transitions from a SBML of *E. coli* core (http://bigg.ucsd.edu/models/e_coli_core).

4.3.Network Analyzer

The Network Analyzer was developed with python language and runs on all platforms. It was used to construct minimal subnetworks for each metabolite in the theoretical dataset. The maximum level for network extension was set to 4. The Network Analyzer performs simulation of label propagation according to the scalaflux approach [Millard et al, 2020] implemented in influx_si [Sokol et al, 2012]. At metabolic steady state, the LLIs can be formalized mathematically by computing a signal in the form of continuous linear piece-wise functions in time points [Sokol and Portais, 2015]. The label propagation of the metabolite of interest can then be simulated from the LLIs by resolving a system of ordinary differential equations (ODEs). It uses *influx_si* to fit labeling dynamics using flux calculation routines by minimizing the difference between experimental labeling data and the simulated labeling profiles. It computes fitting quality metrics ($\chi 2$) to evaluate consistency score of the corresponding minimal subnetwork and evaluate if they can explain experimental isotopic labeling dynamics [Antoniewicz et al, 2006].

References

Amara, A., Frainay, C., Jourdan, F., Naake, T., Neumann, S., Novoa-Del-Toro, E. M., et al. (2022). Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Frontiers in Molecular Biosciences*, *9*, 841373. https://doi.org/10.3389/fmolb.2022.841373

Antoniewicz, M. R., Kelleher, J. K., & Stephanopoulos, G. (2006). Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metabolic Engineering*, *8*(4), 324–337. https://doi.org/10.1016/j.ymben.2006.01.004

Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., & Rabinowitz, J. D. (2009). Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in Escherichia coli. *Nature chemical biology*, *5*(8), 593–599. https://doi.org/10.1038/nchembio.186

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., & Barrett, M. P. (2006). Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics: Official Journal of the Metabolomic Society*, *2*(3), 155–164. https://doi.org/10.1007/s11306-006-0029-z

Butin, N., Bergès, C., Portais, J.-C., & Bellvert, F. (2022). An optimization method for untargeted MS-based isotopic tracing investigations of metabolism. *Metabolomics: Official Journal of the Metabolomic Society*, *18*(7), 41. https://doi.org/10.1007/s11306-022-01897-5

Cesur, M. F., Siraj, B., Uddin, R., Durmuş, S., & Çakır, T. (2020). Network-Based Metabolism-Centered Screening of Potential Drug Targets in Klebsiella pneumoniae at Genome Scale. *Frontiers in Cellular and Infection Microbiology*, *9*. https://www.frontiersin.org/articles/10.3389/fcimb.2019.00447. Accessed 26 April 2023

Chen, N., del Val, I. J., Kyriakopoulos, S., Polizzi, K. M., & Kontoravdi, C. (2012). Metabolic network reconstruction: advances in in silico interpretation of analytical information. *Current Opinion in Biotechnology*, *23*(1), 77–82. https://doi.org/10.1016/j.copbio.2011.10.015

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M.-F., & Jourdan, F. (2010). MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research*, *38*(Web Server issue), W132-137. https://doi.org/10.1093/nar/gkq312

Dange, M. C., Mishra, V., Mukherjee, B., Jaiswal, D., Merchant, M. S., Prasannan, C. B., & Wangikar, P. P. (2020). Evaluation of freely available software tools for untargeted quantification of 13C isotopic enrichment in cellular metabolome from HR-LC/MS data. *Metabolic Engineering Communications*, *10*, e00120. https://doi.org/10.1016/j.mec.2019.e00120

Edwards, J. S., & Palsson, B. O. (2000). The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5528–5533.

Glont, M., Nguyen, T. V. N., Graesslin, M., Hälke, R., Ali, R., Schramm, J., et al. (2018). BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Research*, *46*(D1), D1248–D1253. https://doi.org/10.1093/nar/gkx1023

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*(1), 121. https://doi.org/10.1186/s13059-019-1730-3

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, *19*(4), 524–531. https://doi.org/10.1093/bioinformatics/btg015

Kiefer, P., Schmitt, U., Müller, J. E. N., Hartl, J., Meyer, F., Ryffel, F., & Vorholt, J. A. (2015). DynaMet: A Fully Automated Pipeline for Dynamic LC–MS Data. *Analytical Chemistry*, *87*(19), 9679–9686. https://doi.org/10.1021/acs.analchem.5b01660

Kiefer, P., Schmitt, U., & Vorholt, J. A. (2013). eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. *Bioinformatics (Oxford, England)*, *29*(7), 963–964. https://doi.org/10.1093/bioinformatics/btt080

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., et al. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515-522. https://doi.org/10.1093/nar/gkv1049

Litsa, E. E., Peña, M. I., Moll, M., Giannakopoulos, G., Bennett, G. N., & Kavraki, L. E. (2018). Machine Learning Guided Atom Mapping of Metabolic Reactions. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.8b00434

Millard, P., Schmitt, U., Kiefer, P., Vorholt, J. A., Heux, S., & Portais, J.-C. (2020). ScalaFlux: A scalable approach to quantify fluxes in metabolic subnetworks. *PLoS Computational Biology*, *16*(4), e1007799. https://doi.org/10.1371/journal.pcbi.1007799

Millard, P., Sokol, S., Letisse, F., & Portais, J.-C. (2014). IsoDesign: a software for optimizing the design of 13C-metabolic flux analysis experiments. *Biotechnology and Bioengineering*, *111*(1), 202–208. https://doi.org/10.1002/bit.24997

Sokol, S., Millard, P., & Portais, J.-C. (2012). influx_s: increasing numerical stability and precision for metabolic flux analysis in isotope labelling experiments. *Bioinformatics (Oxford, England)*, *28*(5), 687–693. https://doi.org/10.1093/bioinformatics/btr716

Sokol, S., & Portais, J.-C. (2015). Theoretical Basis for Dynamic Label Propagation in Stationary Metabolic Networks under Step and Periodic Inputs. *PloS One*, *10*(12), e0144652. https://doi.org/10.1371/journal.pone.0144652

Zorrilla, F., Buric, F., Patil, K. R., & Zelezniak, A. (2021). metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Research*, *49*(21), e126. https://doi.org/10.1093/nar/gkab815

Supplementary Data

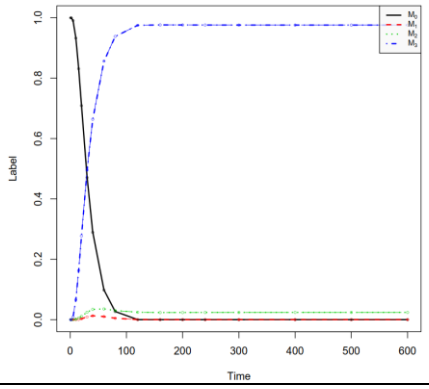
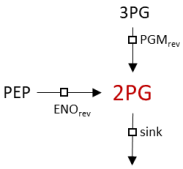
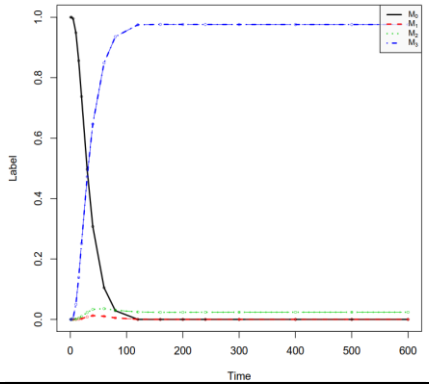
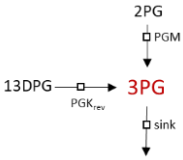
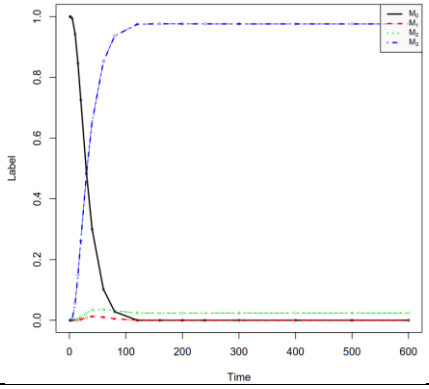
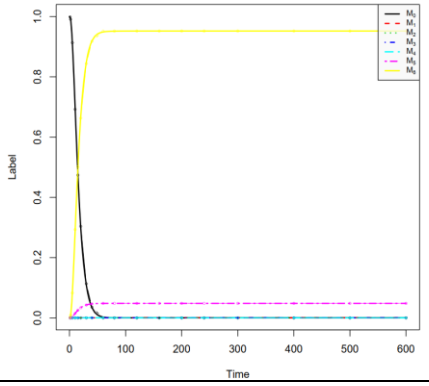
Table S1 : The isotopic model. All the stoichiometric reactions with metabolic pathways and flux values used to simulate the theoretical dataset. The present list of reactions was obtained from the *E. coli* core model and the flux values were obtained using an *in silico* Flux Balance Analysis.

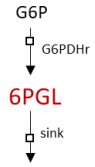
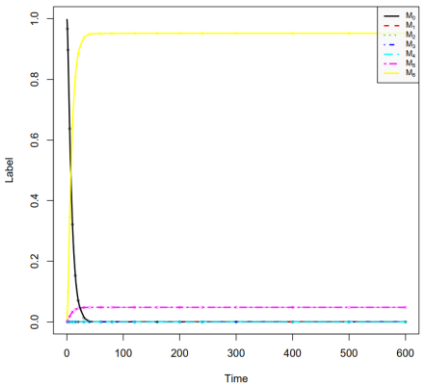
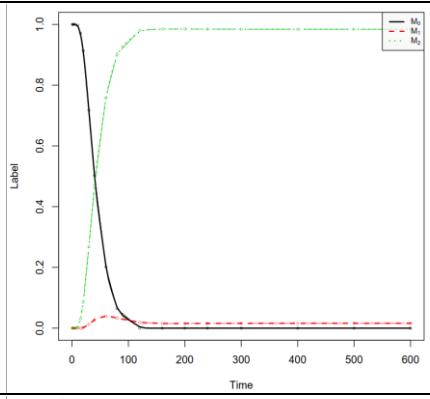
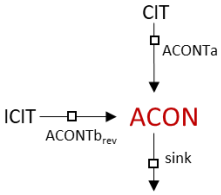
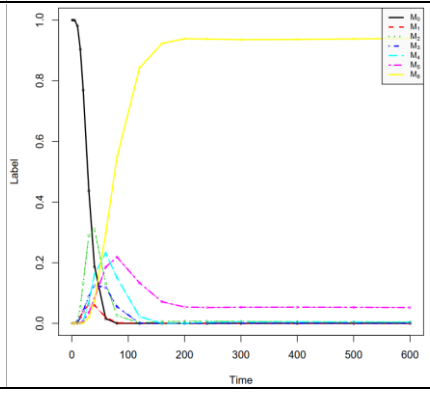
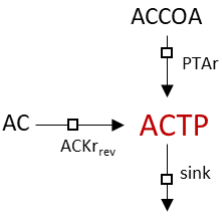
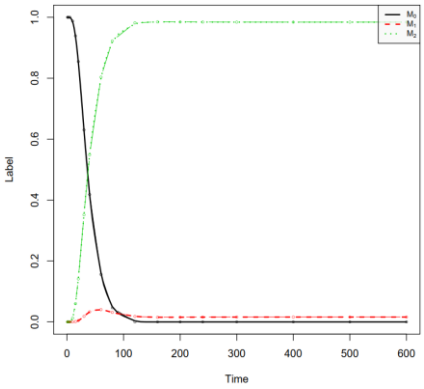
Reaction	Stoechiometry	Pathway	Flux values (mmol/L/s)
R_ACKr	M_actp_c (AB) + M_adp_c () <-> M_ac_c (AB) + M_atp_c ()	Acetate	2.35E-01
R_PTAr	M_accoa_c (AB) + M_pi_c () <-> M_coa_c () + M_actp_c (AB)	Acetate	2.35E-01
R_ME1	M_mal__L_c (ABCD) + M_nad_c () -> M_co2_c (D) + M_pyr_c (ABC) + M_nadh_c ()	Anapleros	-2.60E-04
R_ME2	M_nadp_c () + M_mal__L_c (VWXY) -> M_nadph_c () + M_co2_c (Y) + M_pyr_c (VWX)	Anapleros	1.00E-04
R_PPC	M_co2_c (A) + M_pep_c (BCD) + M_h2o_c () -> M_oaa_c (BCDA) + M_h_c () + M_pi_c ()	Anapleros	3.34E-01
R_PPCK	M_oaa_c (ABCD) + M_atp_c () -> M_co2_c (D) + M_pep_c (ABC) + M_adp_c ()	Anapleros	1.00E-05
R_PPS	M_pyr_c (ABC) + M_atp_c () + M_h2o_c () -> M_pep_c (ABC) + M_amp_c () + 2.0*M_h_c () + M_pi_c ()	Anapleros	-7.53E-05
R_ADK1	M_amp_c () + M_atp_c () <-> 2.0*M_adp_c ()	AXP	-7.53E-05
R_ATPM	M_atp_c () + M_h2o_c () -> M_adp_c () + M_h_c () + M_pi_c ()	AXP	1.32E+00
R_ATPS4r	M_adp_c () + 4.0*M_h_e () + M_pi_c () <-> M_atp_c () + M_h2o_c () + 3.0*M_h_c ()	AXP	6.03E+00
R_ENO	M_2pg_c (ABC) <-> M_pep_c (ABC) + M_h2o_c ()	Glycolysis	2.09E+00
R_FBA	M_fdp_c (ABCDEF) <-> M_g3p_c (DEF) + M_dhap_c (ABC)	Glycolysis	1.09E+00
R_FBP	M_fdp_c (ABCDEF) + M_h2o_c () -> M_f6p_c (ABCDEF) + M_pi_c ()	Glycolysis	-5.13E-05
R_GAPD	M_g3p_c (ABC) + M_nad_c () + M_pi_c () <-> M_13dpg_c (ABC) + M_nadh_c () + M_h_c ()	Glycolysis	2.26E+00
R_PFK	M_atp_c () + M_f6p_c (KLMNOP) -> M_adp_c () + M_fdp_c (KLMNOP) + M_h_c ()	Glycolysis	1.09E+00
R_PGI	M_g6p_c (ABCDEF) <-> M_f6p_c (ABCDEF)	Glycolysis	8.62E-01
R_PGK	M_atp_c () + M_3pg_c (KLM) <-> M_adp_c () + M_13dpg_c (KLM)	Glycolysis	-2.26E+00
R_PGM	M_2pg_c (ABC) <-> M_3pg_c (ABC)	Glycolysis	-2.09E+00
R_PYK	M_adp_c () + M_pep_c (KLM) + M_h_c () -> M_atp_c () + M_pyr_c (KLM)	Glycolysis	3.26E-01
R_TPI	M_dhap_c (ABC) <-> M_g3p_c (ABC)	Glycolysis	1.09E+00
R_ICL	M_icit_c (ABCDEF) -> M_glx_c (AB) + M_succ_c (CDEF)	Glyoxylate Shunt	-6.00E-05

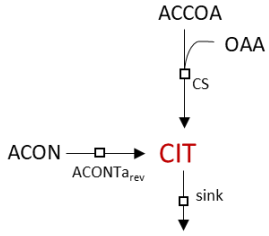
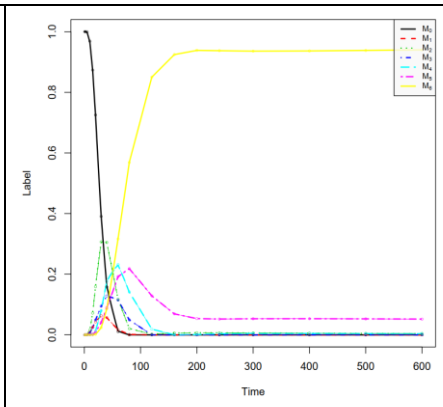
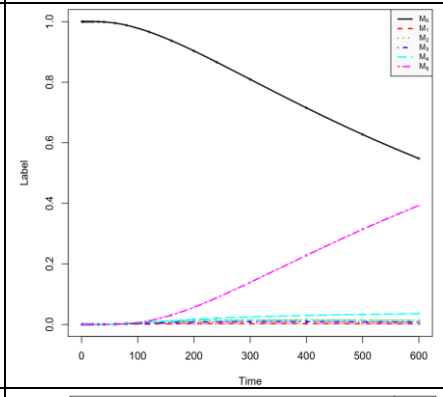
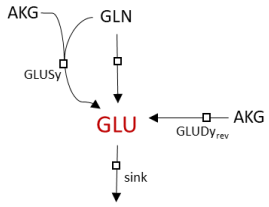
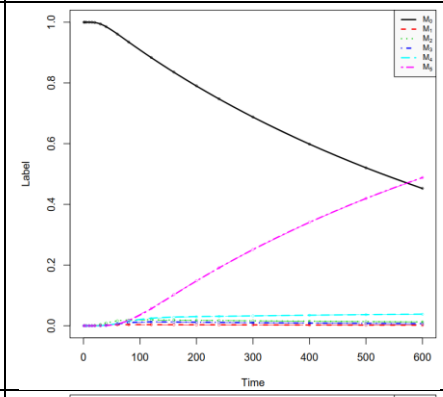
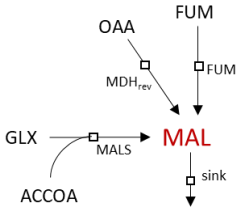
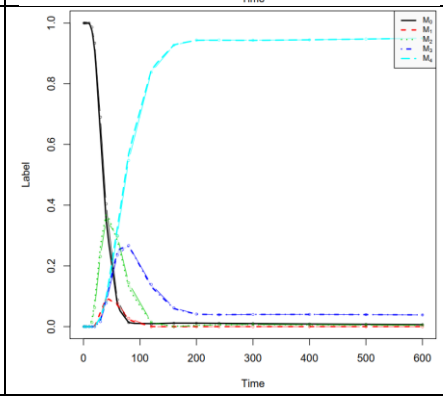
R_MALS	M_glx_c (AB) + M_accoa_c (CD) + M_h2o_c () -> M_mal__L_c (ABCD) + M_coa_c () + M_h_c ()	Glyoxylate Shunt	-6.00E-05
R_GLNS	M_glu__L_c (ABCDE) + M_atp_c () + M_nh4_c () -> M_gln__L_c (ABCDE) + M_adp_c () + M_h_c () + M_pi_c ()	Nitrogen metabolism	2.98E-02
R_GLUDy	M_nadp_c () + M_glu__L_c (VWXYZ) + M_h2o_c () <-> M_nadph_c () + M_akg_c (VWXYZ) + M_h_c () + M_nh4_c ()	Nitrogen metabolism	-1.26E-01
R_GLUN	M_gln__L_c (ABCDE) + M_h2o_c () -> M_glu__L_c (ABCDE) + M_nh4_c ()	Nitrogen metabolism	-1.00E-04
R_GLUSy	M_nadph_c () + M_gln__L_c (VWXYZ) + M_akg_c (abcde) + M_h_c () -> M_nadp_c () + M_glu__L_c (VWXYZ) + M_glu__L_c (abcde)	Nitrogen metabolism	4.99E-05
R_G6PDH2r	M_nadp_c () + M_g6p_c (VWXYZa) <-> M_nadph_c () + M_6pgl_c (VWXYZa) + M_h_c ()	PPP	4.80E-01
R_GND	M_nadp_c () + M_6pgc_c (aVWXYZ) -> M_nadph_c () + M_co2_c (a) + M_ru5p__D_c (VWXYZ)	PPP	4.80E-01
R_PGL	M_6pgl_c (ABCDEF) + M_h2o_c () -> M_6pgc_c (ABCDEF) + M_h_c ()	PPP	4.80E-01
R_RPE	M_ru5p__D_c (ABCDE) <-> M_xu5p__D_c (ABCDE)	PPP	2.36E-01
R_RPI	M_r5p_c (ABCDE) <-> M_ru5p__D_c (ABCDE)	PPP	-2.44E-01
R_TALA	M_g3p_c (ABC) + M_s7p_c (DEFGHIJ) <-> M_e4p_c (GHIJ) + M_f6p_c (DEFABC)	PPP	1.39E-01
R_TKT1	M_xu5p__D_c (ABCDE) + M_r5p_c (FGHIJ) <-> M_g3p_c (CDE) + M_s7p_c (ABFGHIJ)	PPP	1.39E-01
R_TKT2	M_e4p_c (ABCD) + M_xu5p__D_c (EFGHI) <-> M_f6p_c (EFABCD) + M_g3p_c (GHI)	PPP	9.69E-02
R_CYTBD	M_q8h2_c () + 2.0*M_h_c () + 0.5*M_o2_c () -> M_q8_c () + M_h2o_c () + 2.0*M_h_e ()	Redox	5.72E+00
R_NADH16	M_q8_c () + M_nadh_c () + 4*M_h_c () -> M_q8h2_c () + M_nad_c () + 3*M_h_e ()	Redox	5.16E+00
R_NADTRHD	M_nadph_c () + M_nad_c () -> M_nadh_c () + M_nadp_c ()	Redox	1.00E-04
R_THD2	M_nadp_c () + M_nadh_c () + 2*M_h_e () -> M_nad_c () + M_nadph_c () + 2*M_h_c ()	Redox	1.00E-05
R_Act2r	M_ac_c (AB) + M_h_c () -> M_ac_e (AB) + M_h_e ()	Release	2.35E-01
R_CO2_sink	M_co2_c (A) -> M_co2_sink (A)	Release	2.75E+00
R_H2Ot	M_h2o_c () -> M_h2o_e ()	Release	3.60E+00
R_H_out	M_h_e -> M_h_out	Release	2.58E+00
R_ACONTa	M_cit_c (ABCDEF) <-> M_acon_C_c (ABCDEF) + M_h2o_c ()	TCA Cycle	6.87E-01
R_ACONTb	M_acon_C_c (ABCDEF) + M_h2o_c () <-> M_icit_c (ABCDEF)	TCA Cycle	6.87E-01
R_AKGDH	M_coa_c () + M_nad_c () + M_akg_c (qrstu) -> M_succoa_c (rstu) + M_nadh_c () + M_co2_c (q)	TCA Cycle	5.61E-01
R_CS	M_accoa_c (AB) + M_oaa_c (XYZa) + M_h2o_c () -> M_cit_c (aZYXBA) + M_coa_c () + M_h_c ()	TCA Cycle	6.87E-01
R_FUM	M_fum_c (ABCD) + M_h2o_c () <-> M_mal__L_c (ABCD)	TCA Cycle	5.61E-01

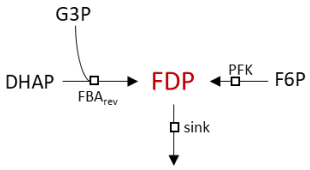
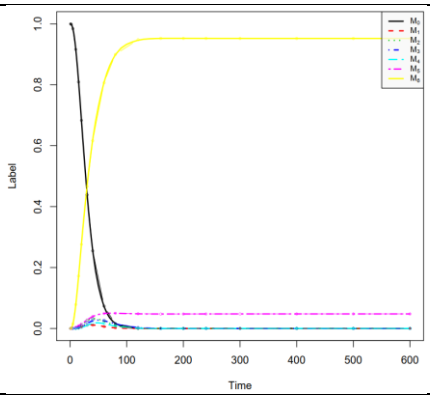
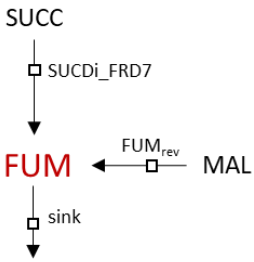
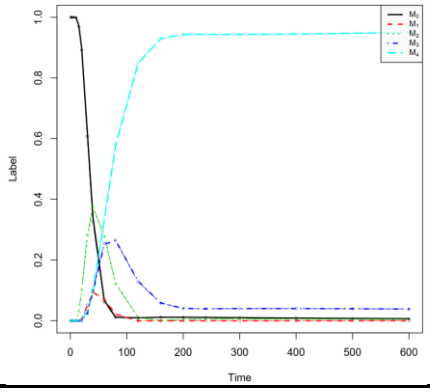
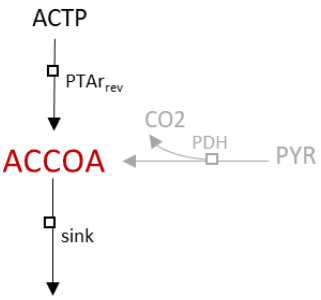
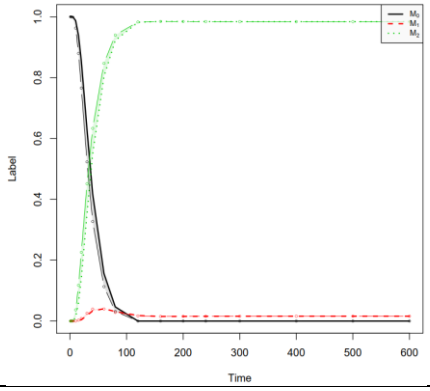
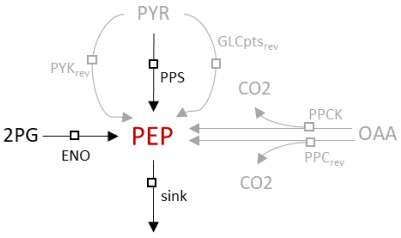
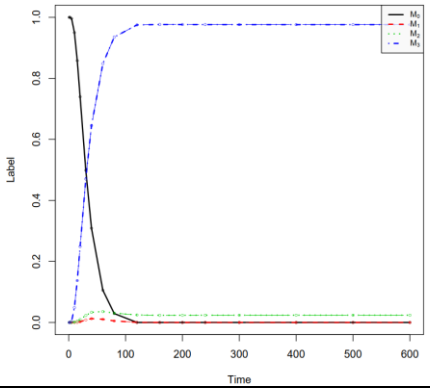
R_ICDHyr	M_nadp_c () + M_icit_c (VWXYZa) <-> M_nadph_c () + M_co2_c (Y) + M_akg_c (VWXZa)	TCA Cycle	6.87E-01
R_MDH	M_mal__L_c (ABCD) + M_nad_c () <-> M_oaa_c (ABCD) + M_nadh_c () + M_h_c ()	TCA Cycle	5.61E-01
R_PDH	M_pyr_c (ABC) + M_coa_c () + M_nad_c () -> M_co2_c (A) + M_accoa_c (BC) + M_nadh_c ()	TCA Cycle	1.36E+00
R_SUCDi_FRD7	M_q8_c () + M_succ_c (ABCD) <-> M_q8h2_c () + M_fum_c (ABCD)	TCA Cycle	5.61E-01
R_SUCOAS	M_coa_c () + M_succ_c (VWXY) + M_atp_c () <-> M_succoa_c (VWXY) + M_adp_c () + M_pi_c ()	TCA Cycle	-5.61E-01
R_CO2_atm	M_co2_atm (A) -> M_co2_c (A)	Uptake	6.54E-05
R_GLCpts	M_pep_c (ABC) + M_glc__D_e (DEFGHI) -> M_pyr_c (ABC) + M_g6p_c (DEFGHI)	Uptake	1.37E+00
R_NH4t	M_nh4_e () -> M_nh4_c ()	Uptake	1.56E-01
R_O2t	M_o2_e () -> M_o2_c ()	Uptake	2.86E+00
R_Plt2r	M_h_e () + M_pi_e () -> M_h_c () + M_pi_c ()	Uptake	4.29E-01
BM	1.496*M_3pg_c + 3.7478*M_accoa_c + 59.81*M_atp_c + 0.361*M_e4p_c + 0.0709*M_f6p_c + 0.129*M_g3p_c + 0.205*M_g6p_c + 0.2557*M_gln__L_c + 0.8232*M_glu__L_c + 59.81*M_h2o_c + 3.547*M_nad_c + 13.0279*M_nadph_c + 1.7867*M_oaa_c + 0.5191*M_pep_c + 2.8328*M_pyr_c + 0.8977*M_r5p_c -> 59.81*M_adp_c + 3.7478*M_coa_c + 59.81*M_h_c + 3.547*M_nadh_c + 13.0279*M_nadp_c + 59.81*M_pi_c	Biomass equation	1.17E-01

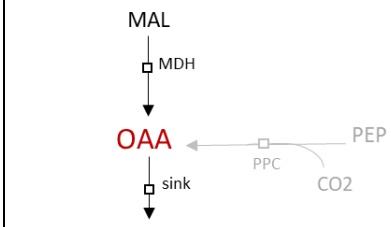
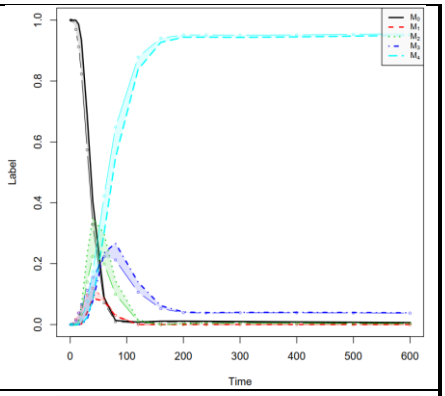
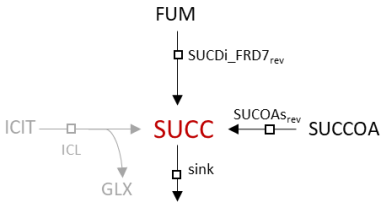
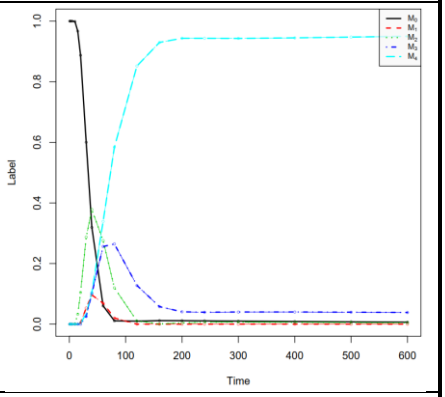
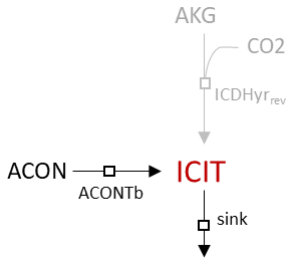
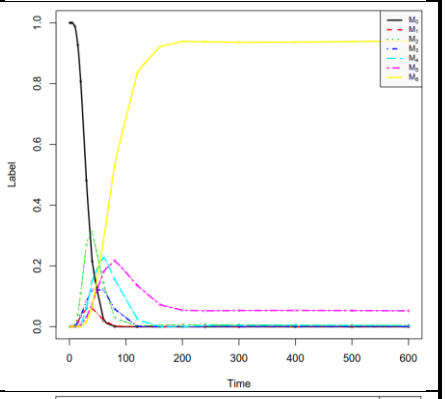
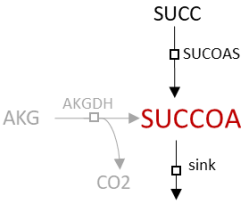
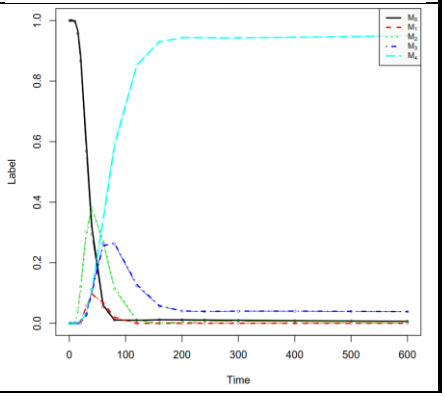
Table S2 : Minimal Subnetworks. Topology of the 29 minimal subnetworks constructed using the Network analyzer and fit results. The reactions at level 1 which were not conserved in incomplete minimal subnetworks are illustrated in grey**.**

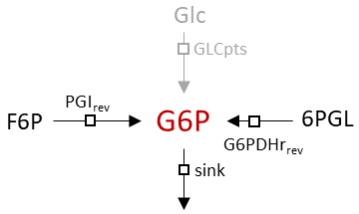
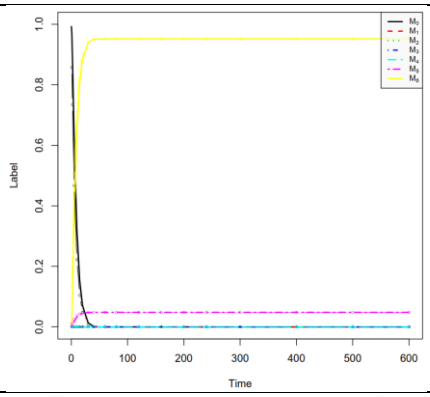
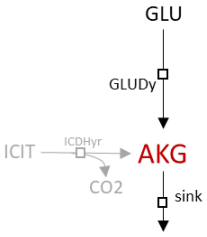
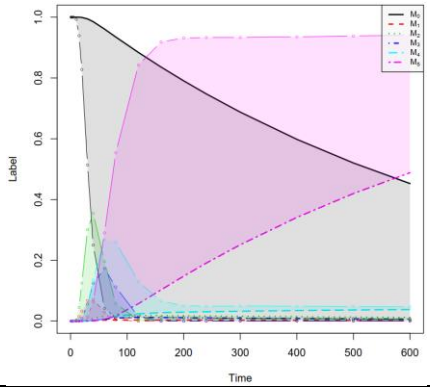
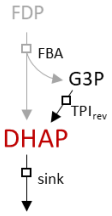
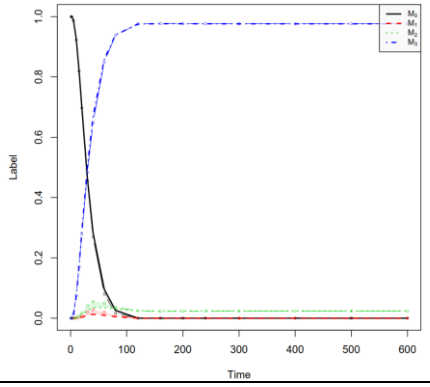
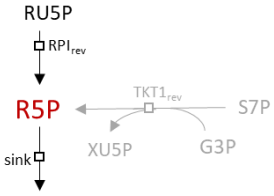
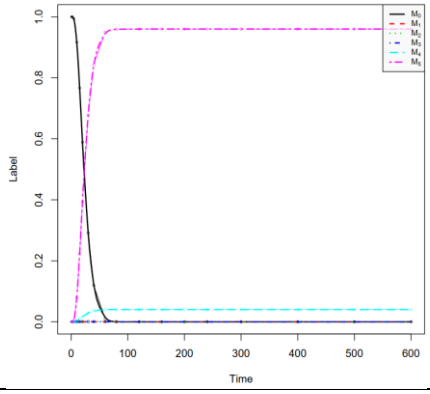
	Minimal subnetwork	Topology	Fit
Complete minimal subnetworks	S_{13DPG}		
	S_{2PG}		
	S_{3PG}		
	S_{6PGC}		

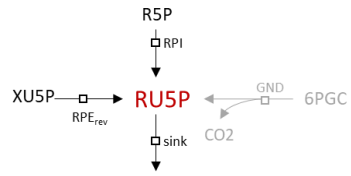
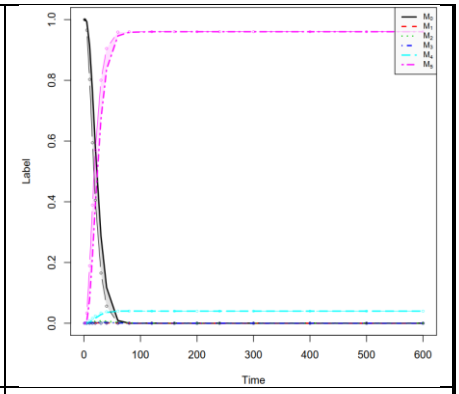
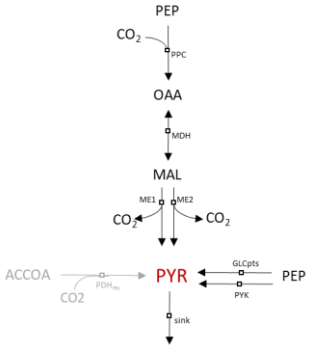
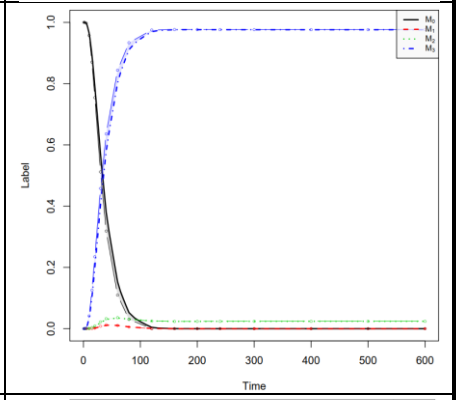
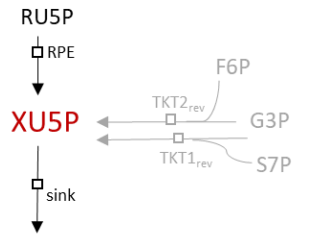
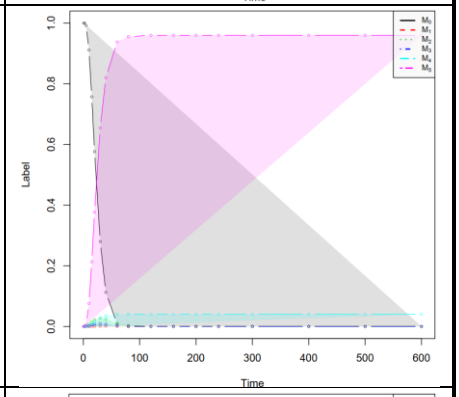
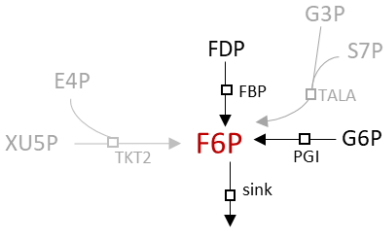
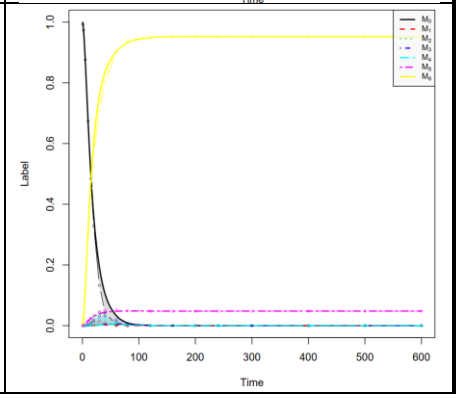
	S_{6PGL}		
	S_{AC}		
	S_{ACON}		
	S_{ACTP}		

	S_{CIT}		
	S_{GLN}		
	S_{GLU}		
	S_{MAL}		

	S_{FDP}		
	S_{FUM}		
Incomplete minimal subnetworks	S_{ACCOA}		
	S_{PEP}		

	S_{OAA}		
	S_{SUCC}		
	S_{ICIT}		
	S_{SUCCOA}		

	S_{G6P}		
	$S_{\alpha KG}$		
	S_{DHAP}		
	S_{R5P}		

	S_{RU5P}		
	S_{PYR}		
	S_{XU5P}		
	S_{F6P}		

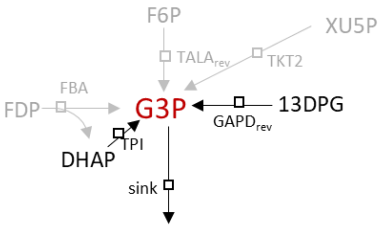
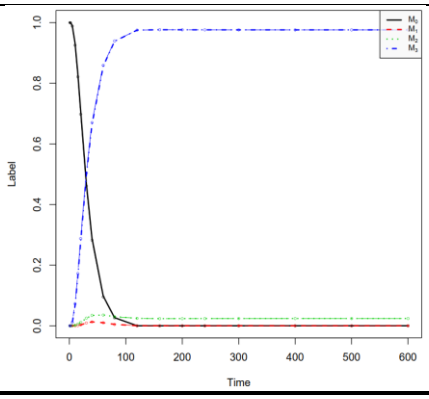
	S_{G3P}	 F6P XU5P TALA$_{rev}$ TKT2 FBA FDP G3P 13DPG GAPD$_{rev}$ TPI DHAP sink	 Label Time 0.0 0.2 0.4 0.6 0.8 1.0 0 100 200 300 400 500 600

Table S3 : Results of the *in silico* validation of the Network Analyzer. n_reactions_tot is the number of reactions (with the sink reaction) included in the constructed minimal subnetworks. This number is compared to the number of reactions in the theoretical subnetworks, *i.e* if the subnetworks were complete at level 1 of extension. Minimal subnetworks are evaluated (i) on their topology (characteristics: complete or incomplete) and (ii) their functionnal characteristics: the goodness-of-fit based on the χ2 pvalue; the identifiability if they provide a unique solution of flux values ; the compliance of flux values between estimated and theoretical fluxes.

		TOPOLOGYCAL EVALUATION			FUNCTIONNAL CHARACTERISTICS		IDENTIFIABILITY	
		Theoretical Subnetworks	Minimal Subnetworks		Minimal Subnetworks		Minimal Subnetworks	
ID	Subnetwork	n_reactions_tot	n_reactions_tot	Characteristics	Goodness-of-fit	pvalue Chi2	Identifiability	μ Fluxes Relative Errors (%)
1	S_{13DPG}	3	3	Complet	Correct	<0.05	False	/
2	S_{2PG}	3	3	Complet	Correct	<0.05	False	/
3	S_{3PG}	3	3	Complet	Correct	<0.05	False	/
4	S_{ACON}	3	3	Complet	Correct	<0.05	False	/
5	S_{ACTP}	3	3	Complet	Correct	<0.05	False	/
6	S_{CIT}	3	3	Complet	Correct	<0.05	False	/
7	S_{MAL}	4	4	Complet	Correct	<0.05	False	/
8	S_{FDP}	3	3	Complet	Correct	<0.05	False	/
9	S_{FUM}	3	3	Complet	Correct	<0.05	False	/
10	S_{6PGC}	2	2	Complet	Correct	<0.05	True	2.53
11	S_{6PGL}	2	2	Complet	Correct	<0.05	True	8.23
12	S_{AC}	2	2	Complet	Correct	<0.05	True	8.1
13	S_{GLN}	2	2	Complet	Correct	<0.05	True	0.86
14	S_{GLU}	4	4	Complet	Correct	<0.05	True	15048
15	S_{RSP}	3	2	Incomplet	Correct	<0.05	True	6.86
16	S_{ICIT}	3	2	Incomplet	Correct	<0.05	True	8.59
17	S_{DHAP}	3	2	Incomplet	Correct	<0.05	True	256
18	S_{F6P}	5	3	Incomplet	Correct	<0.05	True	239041
19	S_{PEP}	4	3	Incomplet	Correct	<0.05	False	/
20	S_{SUCC}	4	3	Incomplet	Correct	<0.05	False	/
21	S_{SUCCOA}	3	2	Incomplet	Correct	<0.05	False	/
22	S_{ACCOA}	3	2	Incomplet	Incorrect	>0.05	False	/

23	S_{RU5P}	**4**	**3**	Incomplet	Incorrect	>0.05	False	**/**
24	S_{G6P}	**4**	**3**	Incomplet	Incorrect	>0.05	False	**/**
25	S_{AKG}	**3**	**2**	Incomplet	Incorrect	>0.05	False	**/**
26	S_{OAA}	**3**	**2**	Incomplet	Incorrect	>0.05	False	**/**
27	S_{G3P}	**7**	**3**	Incomplet	Incorrect	>0.05	False	**/**
28	S_{PYR}	**6**	**8**	Incomplet	Incorrect	>0.05	False	**/**
29	S_{XU5P}	**4**	**2**	Incomplet	Incorrect	>0.05	False	**/**
30	S_{S7P}	**3**	**1**	X	**/**	**/**	**/**	**/**
31	S_{GLX}	**2**	**1**	X	**/**	**/**	**/**	**/**
32	S_{E4P}	**3**	**1**	X	**/**	**/**	**/**	**/**

Chapitre 4

Exploring the Glucose Fluxotype of the E.coli y-ome Using High-Resolution Fluxomics.

Cécilia Bergès[1,2,†], Edern Cahoreau[1,2,†], Pierre Millard[1], Brice Enjalbert[1] , Mickael Dinclaux[1] , Maud Heuillet[1,2], Hanna Kulyk[1,2], Lara Gales[1,2], **Noémie Butin[1,2,3]**, Maxime Chazalviel[4] , Tony Palama[1,2], Matthieu Guionnet[1,2] , Sergueï Sokol[1] , Lindsay Peyriga[1,2] , Floriant Bellvert[1,2], Stéphanie Heux[1] and Jean-Charles Portais[1,2,3,*]

1 Toulouse Biotechnology Institute (TBI), Université de Toulouse, CNRS, INRAE, INSA, 31077 Toulouse, France;

2 MetaToul-MetaboHUB, National Infrastructure of Metabolomics & Fluxomics (ANR-11-INBS-0010), 31077 Toulouse, France

3 RESTORE, Université de Toulouse, Inserm U1031, CNRS 5070, UPS, EFS, 31100 Toulouse, France

4 Toxalim (Research Centre in Food Toxicology), UMR1331, Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, 31300 Toulouse, France; maxime.chazalviel@gmail.com

* Correspondence: jean-charles.portais@insa-toulouse.f

Metabolites, 2021

Abstract

We have developed a robust workflow to measure high-resolution fluxotypes (metabolic flux phenotypes) for large strain libraries under fully controlled growth conditions. This was achieved by optimizing and automating the whole high-throughput fluxomics process and integrating all relevant software tools. This workflow allowed us to obtain highly detailed maps of carbon fluxes in the central carbon metabolism in a fully automated manner. It was applied to investigate the glucose fluxotypes of 180 *Escherichia coli* strains deleted for y-genes. Since the products of these y-genes potentially play a role in a variety of metabolic processes, the experiments were designed to be agnostic as to their potential metabolic impact. The obtained data highlight the robustness of *E. coli*'s central metabolism to y-gene deletion. For two y-genes, deletion resulted in significant changes in carbon and energy fluxes, demonstrating the involvement of the corresponding y-gene products in metabolic function or regulation. This work also introduces novel metrics to measure the actual scope and quality of high-throughput fluxomics investigations.

1. Introduction

Despite the progress that has been made in sequencing technology and genome annotation, a non-negligible percentage of genes remain uncharacterized [1]. Even for very well-known organisms such as the bacterium *E. coli*, 35-40% of genes (the 'y-ome') are of unknown function [2,3]. The products of these genes (y-genes), are likely to have highly diverse functions among the various cellular processes. Metabolism is the basic process that sustains all the energetic and biosynthetic needs of living organisms to survive and grow. Not surprisingly, attempts have been made to investigate the potential role of the y-genes in metabolic processes [4,5]. Accordingly, several high throughput (HT) methods have recently been developed to explore the potential metabolic function of y-genes. This includes the incubation of purified proteins with metabolite cocktails and the identification by mass spectrometry of metabolites with varying abundances [6] and the application of untargeted metabolomics to map gene-metabolite interactions [5].

Alongside studies aiming to directly identify gene product function, an alternative strategy to study the role of y-genes consists in detailing the phenotypic consequences of a loss-of-function mutation or a gene deletion [7,8]. Together with the generation of mutant libraries - such as the Keio *E. coli* mutant collection [9] -, which has facilitated genome-wide investigations, this approach has been used for large-scale investigations of gene essentiality [9–12] and of molecular [13], morphological [14] and fitness [15] phenotypes in bacteria. Metabolic phenotyping approaches have also been developed in which comparative metabolomics is applied to reveal gene functions in yeast [4].

Metabolic phenotypes are most accurately revealed by fluxomics, which aims to measure the actual rates of biochemical reactions in metabolic networks [16]. Fluxomics measures the actual output of the integrated response of the gene-protein-metabolite interaction network [17] and provides direct access to the cellular phenotype in a quantitative manner [18]. It has thereby become a major tool in comprehensive investigations of cellular metabolism in many fields, ranging from biotechnology [19] to the medical sciences [20]. The measurement of metabolic fluxes is based on 13C-labelling experiments coupled with detailed mathematical models of metabolism (13C-fluxomics). This is a complex and tedious process involving many steps [16,21] and requiring significant expertise in data collection and interpretation. Obtaining the detailed flux information on hundreds of strains needed to explore the *E. coli* y-ome is therefore a motivation to improve current HT fluxomics workflows [22].

In this work, we developed a robust workflow to obtain high-resolution fluxotypes under fully controlled growth conditions for large strain libraries. A fluxotype is defined here as the particular distribution of metabolic fluxes measured for a given strain under given physiological conditions. Resolution refers to the level of flux information that can be generated and is high when a significant number of fluxes can be measured [21, 23, 24]. This workflow was applied to investigate the fluxotypes of 180 *E. coli* strains deleted for y-genes and grown on glucose as sole carbon source. The data show that the central metabolism of *E. coli* is highly robust to y-gene deletion. For two y-genes, deletion resulted in significant alterations of metabolic fluxes, pointing to the role of the corresponding y-gene products in metabolic function and/or regulation.

2. Results

2.1. Selection of E.coli y-ome strains

With the aim of investigating the metabolic phenotypes of the *E. coli* y-ome during growth on glucose as sole carbon source, we first selected y-genes as follows (Figure 4.1):

i) We first considered all genes with a single-deletion mutant in the Keio mutant library [9]. This library contains single-gene deletion mutants able to grow on glucose, meaning the mutated genes are not essential for growth on glucose as sole carbon source. The Keio collection contains 3985 mutants.

ii) We then selected the genes in the Keio collection lacking evidence of function. This represented a total of 1563 y-genes

iii) We verified that each y-gene was duly expressed and translated during growth on glucose as the sole carbon source. This selection step was based on the extensive proteomic investigation performed by Schmidt *et al*., in 2016 [25]. These authors measured the functional expression of 55% *E. coli* genes (> 2300 genes) by quantitative proteomics in 22 experimental conditions, including growth in minimal medium with glucose as carbon source. Among the y-genes identified in step 2, we further selected the 218 y-genes that were experimentally shown to be translated under these conditions.

iv) The y-gene status – *i.e.* the lack of annotation or of experimental evidence of function – was manually verified in two complementary databases, namely Biocyc (https://biocyc.org/version 19.5) and Uniprot (https://www.uniprot.org/, release 2016_02).

This process yielded a group of 180 y-genes with unknown or unclear function, dispensable but duly expressed in a medium with glucose as sole carbon source (Supplementary data 1). The corresponding single deletion mutants were obtained from the Keio collection and considered for further metabolic investigations, which were designed to optimize the measurement of growth parameters and fluxotypes across hundreds of *E. coli* mutants (Figure 4.1). Fluxotypes can be visualized using a so-called flux map, a graphical representation of the metabolic network showing the flux values for the various reactions or pathways.

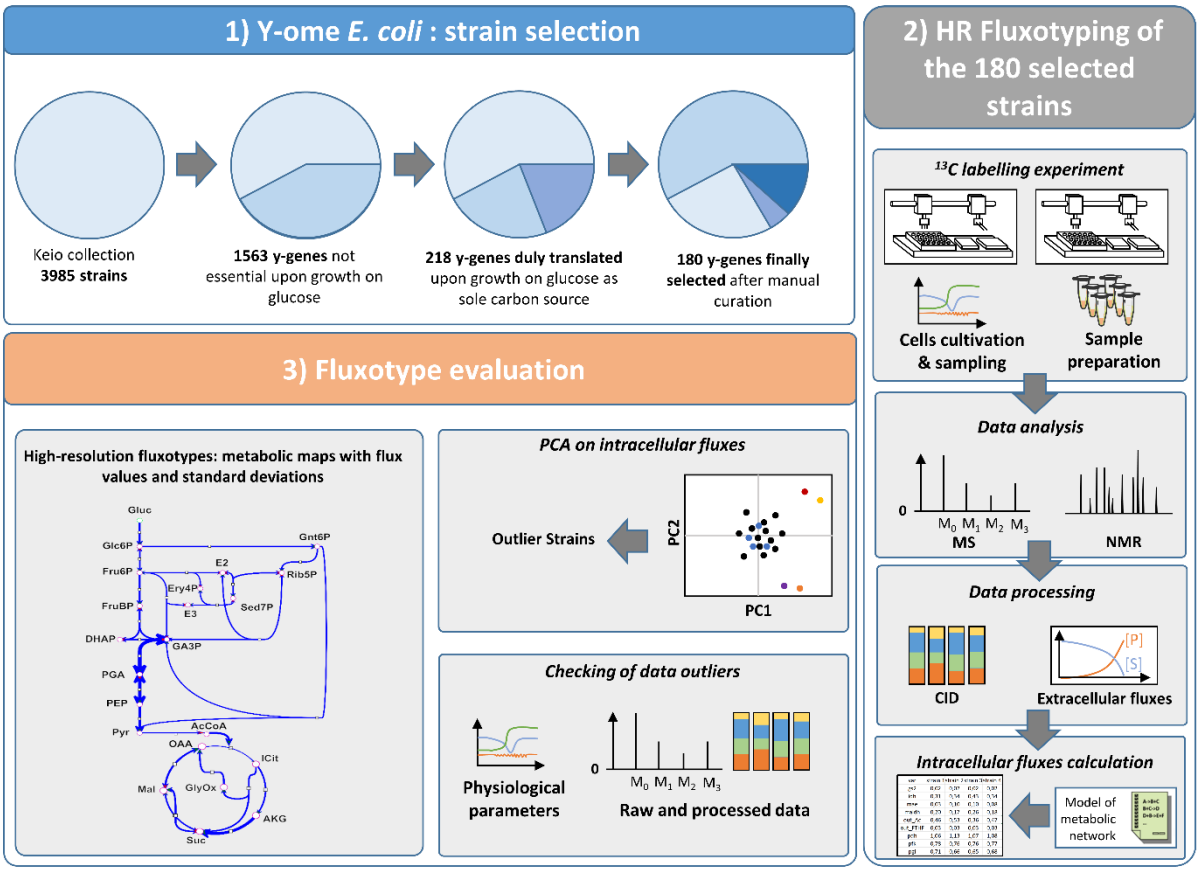


Figure 4.1 : Strategy for high-resolution fluxotyping of the E. coli *y-ome.* ***Selection of* E. coli *y-genes (1)*** *180 y-genes were selected based on their expression on glucose as sole carbon source.* ***High-resolution Fluxotyping (2)*** *of 180 mutants with single-deletion of a selected y-gene. Cell cultivation, metabolite harvesting and sample preparation are performed by robotic systems. Quantification of medium compounds and isotopic profiling of metabolites by NMR and MS. Fluxes are calculated using a model of* E. coli *central carbon metabolism.* ***Fluxotype evaluation (3)*** *Metabolic fluxes are calculated for all strains (high-resolution fluxotypes) and compared by statistical analysis to identify altered fluxotypes. Fluxotype outliers are checked against culture data to ascertain that outlying flux values are not due to a* ^{13}*C-labelling problem.*

2.2. Integrated workflow for high-throughput collection of high-resolution fluxotypes

To measure the fluxotypes of the 180 y-gene deletion mutants, we first built a HT platform enabling fully automated, parallelized measurements of metabolic fluxes under fully controlled conditions at a throughput consistent with the investigation of hundreds of experimental conditions. A typical ^{13}C-fluxomics workflow involves a combination of several complex experimental and computational steps, which were improved and optimized here to meet the needs of the y-ome investigation (Figure 4.1). The final setup included i) two automated robotic platforms, one for parallelized growth, ^{13}C-labelling experiments and sample collection, and the other for sample preparation, ii) optimized NMR and MS analytical methods for measuring metabolite concentrations and labeling patterns, iii) flux calculation, iv) statistical analysis, and v) a series of software programs to store, manage, and process the data and meta-data generated all through the process.

2.3.Design of fluxomics experiments

Very little is known about the phenotypic and metabolic behavior of y-gene–deleted strains, hence an important objective of the setup was to avoid presuppositions about the metabolic pathways potentially impacted by the lack of the y-gene product. This was achieved by using a generic model of *E. coli* metabolism for the flux investigations, containing all central carbon metabolism pathways (Supplementary data 3). The model contained 94 fluxes in total, representing central pathways, biosynthetic processes, and transport reactions, and 49 metabolites. The isotopic composition of the label input in the ^{13}C-labelling experiments was optimized using the software IsoDesign [26]. Based on i) the network topology, ii) the isotopic data collected in the study (*i.e.* isotopologue distributions of proteinogenic amino-acids), and iii) the objective of resolving the maximum number of fluxes across the whole metabolic network, the best label composition was determined to be a mixture of 80% [1-^{13}C]-glucose + 20% [U-^{13}C]-glucose. This result is consistent with those of previous investigations of the entire central carbon metabolism of *E. coli* [16]. All labeling experiments reported in this study were performed in minimal M9 medium with the above mixture as sole carbon source. Details of the conditions used to apply the workflow to all the investigations reported here are given in material and methods section. Briefly, cells were inoculated at 10^8 cell/mL and their growth was monitored by optical density under temperature, pH and pO_2 control. Medium samples

were collected throughout the growth process and analyzed by ^{1}H-NMR to measure the rates of substrate uptake and product release. When an OD_{600nm} of 1.2 – corresponding to mid-exponential growth under these conditions – was reached, the biomass was automatically sampled, hydrolyzed, and the labeling patterns of proteinogenic amino acids were analyzed by LC-HRMS. After correction for naturally occurring isotopes [27,28], the carbon isotopologue distributions (CIDs) of 16 amino acids were measured and used for flux calculations. Biosynthetic fluxes (*i.e.* precursor requirements for growth) were calculated based on the molecular composition of *E. coli* [29]. Intracellular fluxes were calculated by fitting the metabolic model described above with amino acid CIDs, biosynthetic fluxes, and extracellular fluxes, using the software influx_si [30]. The confidence intervals on calculated fluxes were estimated by Monte Carlo sensitivity analysis using the same software. All investigations were performed on the robotic system, which allowed 48 experiments to be run in parallel, yielding 48 flux maps per run each containing 94 fluxes, measured under fully controlled physiological conditions.

2.4. High-Resolution fluxotyping workflow validation

Before starting the investigation of the *E. coli* y-ome, we evaluated the workflow by performing several fluxomics experiments with known *E. coli* strains. Two WT strains (*K-12 MG1655* and *BW25113)* and one mutant strain (*BW25113 Δzwf*), with known and significant metabolic flux alterations, were considered to evaluate the biological relevance of the data collected with the integrated HT workflow. The data collected for the three strains were highly reproducible between replicates. The median relative standard deviation (RSD) of the central metabolic fluxes was 11%, 3.7% and 22% between respectively 5 (*K-12 MG1655),* 4 (*BW25113),* and 5 (*BW25113 Δzwf*) biological replicates. The macro-kinetic parameters (*i.e.* growth rate, glucose uptake, acetate production) and flux values collected for the two WT strains (Figure 4.2) were closely consistent with previous measurements on these strains, highlighting the consistency of the data collected with the integrated HT fluxomics workflow [22,31–33]. The deletion of *zwf* had little impact on the growth rate but led to a significant redistribution of metabolic fluxes to compensate for the impairment of the oxidative branch of the pentose phosphate pathway (PPP). The observed changes in metabolic fluxes in this strain were qualitatively and quantitatively consistent with previous reports [22,31,34,35]. The developed workflow's ability to reliably measure the growth and fluxotypes of these strains clearly demonstrates the consistency of the entire process.

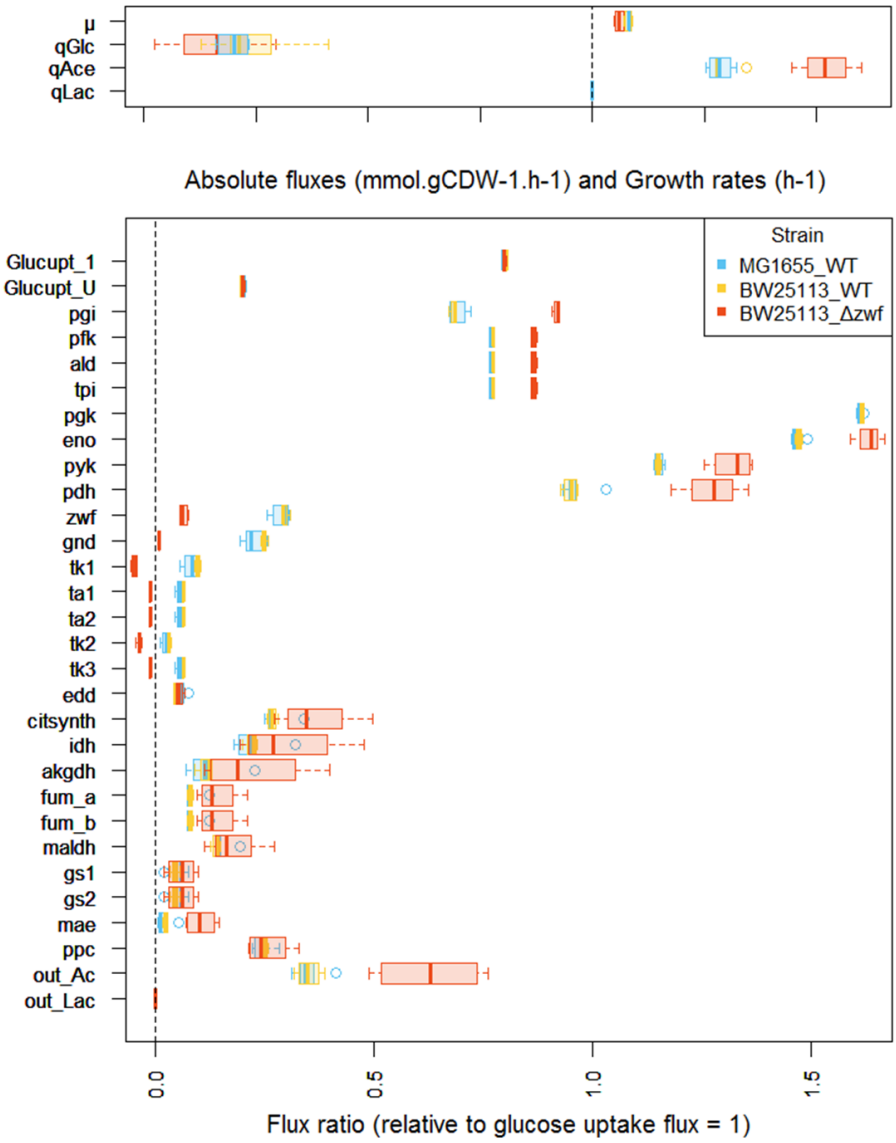


*Figure 4.2 : **Metabolic fluxes measured for E. coli WT strains and a Δzwf mutant strain**. Top: experimentally measured fluxes (expressed in mmol.gDW^{-1}.h^{-1}) and growth rate (expressed in h-^{1}). Physiological measurements: μ, growth rate; qGlc, glucose specific consumption rate; qAce, acetate specific production rate; qLac, lactate specific production rate. Bottom: Intracellular fluxes calculated from the ^{13}C-labelling data; fluxes are expressed relative to the glucose uptake rate set arbitrarily to 1 for each strain. Boxplot produced with the software R representing median fluxes (bold line) in boxes representing the interquartile range (quartile 1- quartile 3). The whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box. **Net fluxes :** Glucupt_1 and Glucupt_U: [1-^{13}C]-glucose and [1-^{13}C]-glucose importation (PTS system), pgi: glucose-6-phosphate isomerase, pfk: phosphofructokinase , ald: fructose-bisphosphate aldolase, tpi: triose-phosphate isomerase, pgk: phosphoglycerate kinase, eno: enolase, pyk: pyruvate kinase, pdh: pyruvate dehydrogenase, zwf: glucose 6-phosphate dehydrogenase, gnd: phosphogluconate dehydrogenase, tk1: half-reaction transketolase (1), ta1: half-reaction transaldolase (1), ta2: half-reaction transaldolase (2), tk2: half-reaction transketolase (2), tk3: tk2: half-reaction transketolase (3), edd: Entner-Doudoroff enzymes: equivalent to 6-phosphogluconate dehydratase and 2-keto-3deoxy-6phosphogluconate (KDPG) aldolase, citsynth: citrate synthase, idh: isocitrate dehydrogenase, akgdh: alpha-ketoglutarate dehydrogenase, fum_a and fum_b: fumarase, maldh: malate dehydrogenase, gs1: isocitrate lyase, gs2: malate synthase, mae: malic enzyme, ppc: equivalent to pep carboxylase (forward flux) and to pep carboxykinase (reverse flux), out_Ac : acetate output (equivalent to acetate k.inase)*

2.5. High-Resolution glucose fluxotyping of 180 selected y-gene mutant strains

The 180 strains deleted for the selected y-genes were isolated from the Keio strain collection, and their growth profiles and fluxotypes were measured in minimal medium with glucose as sole carbon source. The WT and *Δzwf* strains were included in each robotic run as reference strains to check the inter-batch consistency of growth parameters and flux data. In total, 198 cultures were performed from which 1074 OD_{600} points were measured and used for the calculation of growth rates and biomass yields; 1248 samples were collected (1050 filtrates and 198 cell pellets), from which 591 physiological data - *i.e.* rates of compound consumption or production - were extracted; and more than 20000 isotopic data measured (100 carbon isotopologues from 16 different metabolites by MS analyses and 6 positional labelling profiles for 2 metabolites by NMR for each cultivation). Finally, a cumulated total of 18612 fluxes were calculated (94 fluxes for each culture). The complete experimental process was completed within 786 hours, and required only 194 hours – *i.e.* about 1 h per measured fluxotype - of human intervention or supervision. These values emphasize the practical benefit of the established workflow.

The data collected for the reference strains – *i.e.* WT and *Δzwf* strains - in the various runs of the robotic platform did not differ significantly from the data collected in the validation stage. On average, the strains deleted for y-genes grew slower than WT strains but had similar biomass yields (Figure 4.3A). The rates of glucose consumption and acetate production, measured from exometabolome data, were consistent with these observations, *i.e.* no significant difference overall. Principal component analysis (PCA) was used to explore the distribution of intracellular flux data across the y-ome (Figure 4.3B, 4.3C). For a large majority of mutant strains (in blue), the flux data – expressed relative to the glucose uptake rate - clustered with those of the WT strains (in green). This result indicates that for most of the y-genes, deletion has a limited impact on the distribution of intracellular metabolic fluxes. The complete dataset is provided in Supplementary data 4.

A. Comparison of growth parameters between the *E. coli* Y-ome strains and the WT strain

	μ (h^{-1})	q_{glc} ($mmol.gDW^{-1}.h^{-1}$)	*Biomass yield* ($gDW^{-1}.mmol^{-1}$)	q_{ace} ($mmol.gDW^{-1}.h^{-1}$)
BW25113 *(n = 4)*	*0.66 ±0.02*	*-6.41 ±0.30*	*0.10 ±0.01*	*2.29 ±0.24*
E. coli Y_ome *(average of 180 strains)*	*0.57 ±0.06*	*-5.81 ±1.02*	*0.10 ±0.02*	*2.79 ±0.65*

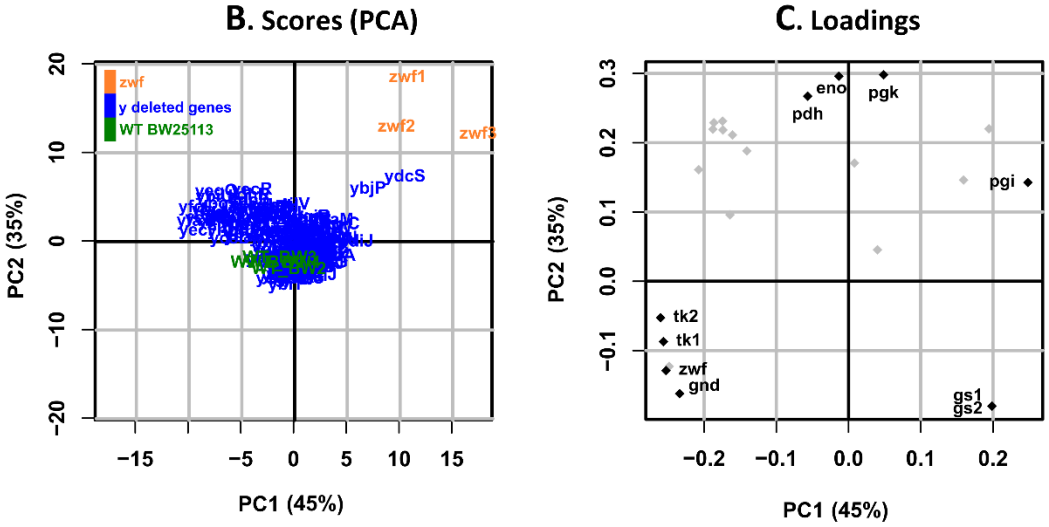


Figure 4.3 : ***Comparison of the growth parameters and fluxotypes of y-ome strains with those of their parental strain.*** *A: Comparison of growth parameters between y-ome strains and the BW25113 WT strain (μ: growth rate, q_{glc}: glucose consumption rate,* Biomass yield, q_{ace}*: acetate production rate). B: PCA plots showing the strains with the most different fluxotypes across the Y-ome. PC1 and PC2 represent Principal Component 1 & 2, explaining the largest variance between strains according to their flux values - the percentage of variance explained for each principal component is indicated into brackets [36], C: PCA loading plot with the most impacted fluxes indicated.*

Three mutant strains did not cluster with the WT strains (Figure 4.3B). One of these was the *Δzwf* strain, demonstrating the value of PCA to discriminate strains based on metabolic fluxes. Two other y-gene mutants, namely Δ*ybjP* and Δ*ydcS*, also had significantly different fluxotypes from those of the WT strains and other y-gene mutants. Loading plots (Figure 4.3C) indicated that the most discriminating fluxes were related to glycolysis (pgi, , pgk, eno, pdh), the PPP (zwf, gnd, tk1, tk2), and the glyoxylate shunt (gs1, gs2).

2.6. Glucose fluxotypes of ΔybjP and ΔydcS strains

The glucose fluxotypes of the *ΔybjP* and *ΔydcS* strains are shown in Figure 4.4). Consistent with the PCA data, significant differences were observed between the two strains and the wild type strain, specifically that both strains had higher glycolytic flux than the wild type. In the *ΔybjP* and *ΔydcS* strains, 82 and 76 % of glucose was metabolized through pgi, respectively, against 68% in the wild type strain. (*i.e.* pgi flux in the *ΔydcS* strain). As a result, the two strains had decreased fluxes through the oxidative branch of the PPP (*i.e.* zwf and gnd), while the flux through the ED pathway (i.e EDD) was stable. This trend was even stronger in the *ΔydcS* strain, for which the data also showed partial reversal of the non-oxidative branch of the PPP (tk1 and tk2 reactions), probably to compensate for the decreased production of pentose-5-P, a key precursor for major biomass components. These data suggest that the products of the *ybjP* and *ydcS* genes are – either directly or indirectly - involved in the control of flux in the PPP pathway.

Significant alterations of fluxes in the TCA cycle and related pathways were observed in the two mutant strains compared to the wild-type strain. Pyruvate dehydrogenase flux was significantly increased in the *ΔydcS* strain, and even more so in the *ΔybjP* strain, reaching 119% against 95% in the wild type. This was not accompanied by increased flux through the TCA cycle but instead by increased acetate excretion, as well as by increased flux through the glyoxylate shunt. Lactate production was observed furthermore in the *ΔybjP* strain, and this strain had no flux through malate dehydrogenase (*i.e.* maldh) and increased flux through malic enzyme (*i.e.* mae). Overall, these data highlight a rerouting of metabolic fluxes around the acetyl-CoA node mainly resulting in the diversion of glycolytic flux from oxidative metabolism to acetate excretion in both strains.

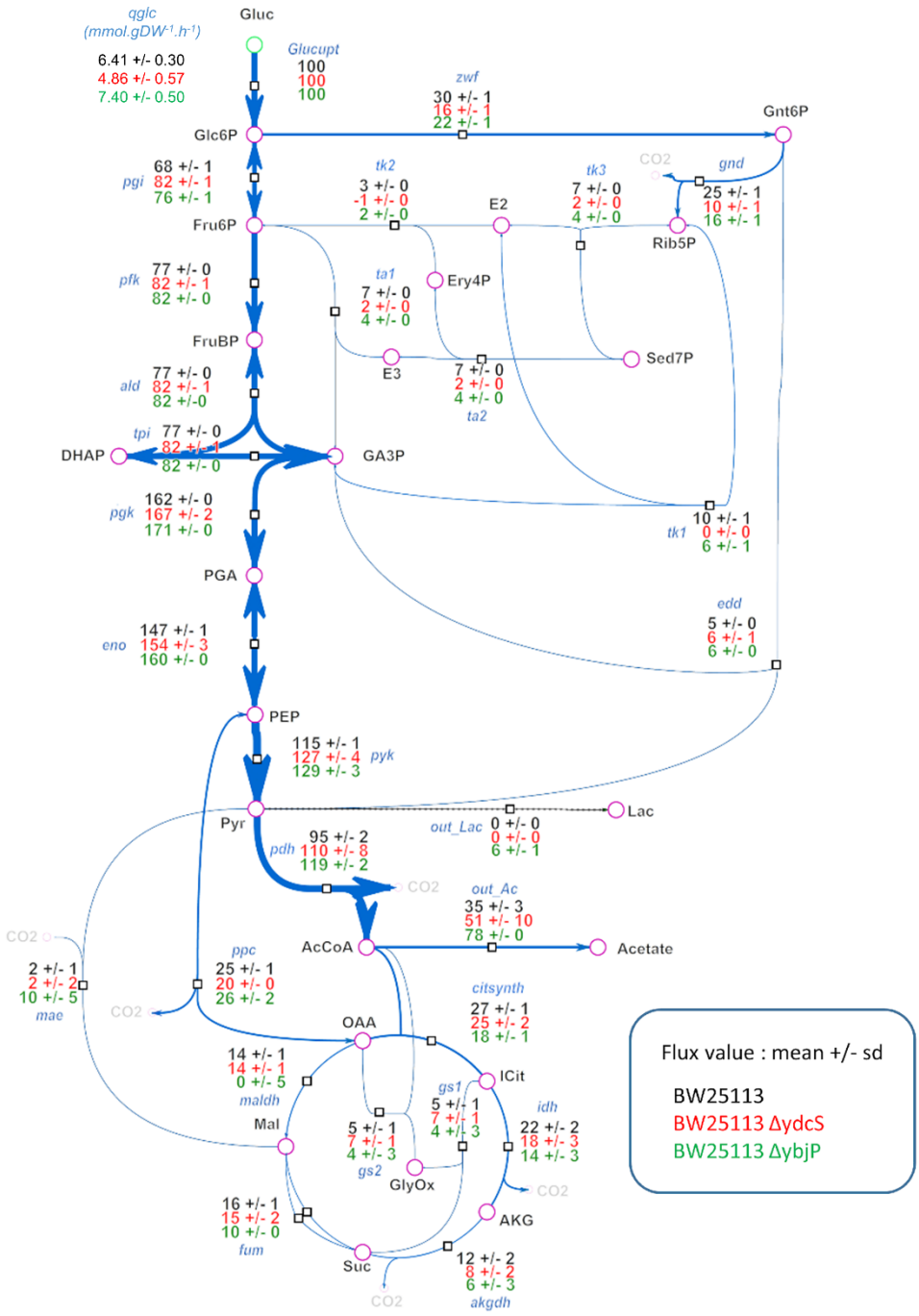


Figure 4.4 : ***Distribution of metabolic fluxes in ΔydcS and ΔybjP strains.*** *A: Intracellular flux distributions in the two mutant strains and comparison with the WT strain. All fluxes are expressed relative to the glucose uptake rate, set arbitrarily to 100 for each strain. The absolute glucose uptake rates (mmol.gDW^{-1}.h^{-1}) of the three strains are shown in the top left of the figure.*

2.7.Cofactor usage in the ΔybjP and ΔydcS strains

The changes in carbon fluxes were associated with changes in the supply of ATP and redox cofactors by the central carbon metabolism (Figure 4.5). The *ΔydcS* strain produced ATP at a slightly lower rate than the WT strain did. Its redox status was significantly impaired. In particular, NADPH production was drastically reduced, due to the much lower flux through the PPP and the absence of compensation via the TCA cycle or malic enzyme. Because this strain grows roughly as fast as the WT strain, hence has similar biosynthetic requirements, it is likely that the reduced production of NADPH via carbon metabolism (Figure 4.5) is compensated by increased transhydrogenase activity [37]. The *ΔybjP* strain had different ATP and redox profiles. First, this strain produces ATP faster than the WT strain does, because of significantly increased fluxes through glycolysis and acetate metabolism. Interestingly, the same ATP production profile was observed in the *Δzwf* strain (Figure 4.5), but the latter is significantly impaired for NADPH production whereas the *ΔybjP* strain produced NADPH at only a slightly lower rate than the WT strain did. This reflects a smaller reduction in PPP flux and an increase in malic enzyme flux, which accounted for 16% of CCM-derived NADPH production in this strain compared to 4% in the WT strain. The latter increase might be the result of a mechanism to compensate for decreased NADPH production via the oxidative PPP. In addition, NADH production was significantly higher in the *ΔybjP* mutant than in the WT strain, mainly due to a large increase in glycolytic production of NADH, and resulting in the production of lactate, suggesting a global imbalance in redox metabolism in this strain.

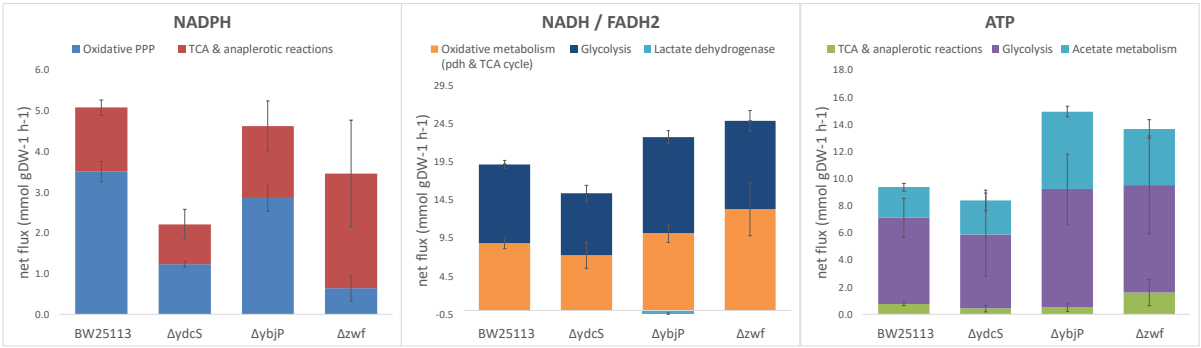


*Figure 4.5 : **Production of NADPH, NADH/FADH$_2$ and ATP in the central carbon metabolism**. The rates were calculated from the mean growth rates and flux distributions, and are expressed in absolute values (mmol.gDW^{-1}.h^{-1}) ± SD. Calculations are described in the materials and methods.*

2.8.Scope and quality of high-throughput fluxomics investigations

Because a specific objective of this work was to deploy high-resolution fluxomics in a HT manner, we evaluated the quality of the fluxome data obtained here in comparison with existing data in the literature. High-throughput investigations have to find the best compromise between throughput and information level (and quality), hence the volume and quality of the flux data have to be appreciated in terms of the actual size of the investigation, *i.e.* the total number of flux maps generated. To this aim, we used the four criteria defined below:

i) the *fluxome resolution*, which corresponds to the total number of fluxes that can be calculated from the experimental dataset, and is a measure of the coverage of the flux space in the investigated metabolic network. This resolution is generally low (e.g. 23) in high-throughput studies, and high (71-84) in low-throughput ones (Table 4.1). In this work, we could measure 94 fluxes per measured fluxotype, which is the highest fluxome resolution value among all ^{13}C-fluxomics studies reported in Table 4.1.

ii) the *isotopic resolutive power* of the flux map, defined as the ratio between the number of isotopic data and the number of fluxes calculated in the network. This index is assumed to reflect data redundancy in the flux estimates, hence higher precision in the calculated flux values. Low-throughput investigations of the *E. coli* fluxome have isotopic resolutive power in the range of 1 to 8 (Table 4.1), though very high isotopic resolutive power (up to 17.6), has been obtained by performing multiple (e.g. parallel) labelling experiments to calculate a single flux map [38–41]. In the present work, the isotopic resolutive power of the flux maps was 1.13 (106 isotopic data for 94 fluxes), which is comparable to the values obtained in low-throughput investigations (Table 4.1).

iii)*the total flux dimension*, which is equal to the number of fluxotypes multiplied by fluxome resolution, and is an indication of the scale of the fluxomics investigation. In the literature, this index ranges between 71 and 4370 (Table 4.1). In this study, the total flux dimension index was 18612 (94 fluxes x 198 fluxotypes), which reflects both the high-throughput and high-resolution character of the analysis. Interestingly, the number of fluxotypes generated in this study is similar to the number reported by Haverkorn van Rijsewijk et al. [33] (198 vs 190, respectively) but the total flux dimension of this work is higher (18612 vs 4370, respectively) because of the higher number of fluxes per fluxotype (94 vs 23, respectively).

iv)the *overall flux precision* index, which is the median RSD of all (free) fluxes across the entire flux dataset. Only free fluxes are considered because the others – e.g. most biosynthetic

fluxes – are determined or given as constraints to the model. The average value of this index was 32% in the present study, compared with 0.4–23% in low-throughput studies, and 14–253 % in HT studies (Table 4.1). The value obtained in this work appears to strike a good compromise between throughput and resolution. Moreover, when measured for specific pathways (*overall pathway-specific precision*), it appears that the level of precision achieved for most fluxes is drastically improved, notably for those in the glycolysis, PPP and ED pathways.

Table 4.1 : **Evaluation of data size and quality in HT fluxomics investigations**. The **overall flux precision** is the median RSD (relative standard deviation) calculated for all fluxes measured in each of the studied conditions, as given in the supplementary data of the cited studies. The **overall pathway precision** was calculated similarly by grouping the RSDs according to the metabolic pathways considered. The flux RSDs considered in these calculations were obtained by sensitivity analysis (S.A.) or by calculating classical standard deviations on the mean values of fluxes obtained when biological replicates were produced (as in this work).

	Leighty *et al.* 2013 [37]	Crown *et al.* 2015 [38]	Long & Antoniewicz 2019 [35]	Millard *et al.* 2014 [30]	Long *et al.* 2016 [39]	Heux *et al.* 2014 [22]	Haverkorn van Rijsewijk *et al.* 2011 [31]	This paper	
E. coli strains	*MG1655*	*MG1655*	*BW25113 DtpiA mutant*	*MG1655*	*BW25113*	*MG1655*	*BW25113 + Keio mutants*	*MG1655 + BW25113 + keio mutants*	
Number of measured fluxotypes	*1*	*1*	*1*	*1*	*1*	*20*	*190*	*198*	
Number of different label input(s) by strain or condition	*6*	*14*	*2*	*1*	*1*	*2*	*2*	*1*	
Fluxome resolution = number of (net) fluxes by fluxotype	***71***	***71***	***75***	***84***	***71***	***23***	***23***	***94***	
Isotopic resolutive power = number of isotopic data / number of calculated fluxes	**7.52**	**17.55**	**5.07**	**1.4**	**1.94**	**2.57**	**1.61**	**1.13**	
Total flux dimension = number of fluxes per fluxotype * number of fluxotypes	**71**	**71**	**75**	**84**	**142**	**460**	**4,370**	**18,612**	
Global flux precision = median RSD over the global dataset	12%	15%	7.80%	19%	23%	285%	14%	32%	
Global pathway precision = median RSD within specific pathways	S.A. (n = 1)	S.A. (n = 1)	S.A. (n = 1)	S.A. (n = 1)	S.A. (n = 1)	S.A. (n = 23)	S.A. (n=23)	S.A. (n=198)	biological replicates (n=20)
Glycolysis	1%	2%	4.20%	3%	3.50%	553%	7%	3%	1%
ppp+edp	14%	11%	17%	131%	10%	41%	25%	24%	7%
tca + gs	11%	19%	4.90%	20%	31%	795%	19%	40%	21%
anaplerosis	122%	144%	87%	13%	1822%	7%	15%	42%	20%
output fluxes	12%	25%	4.7%	15%	42.4%	426%	0%	10%	14%

3. Discussion

This work introduces significant improvements in the parallelization, automation and implementation of fluxomics in large-scale investigations. Altogether, these improvements offer significant gains in throughput, robustness, and data quality, while retaining a high level of biological information through the collection of high resolution fluxomics data. This optimized complete HT fluxomics workflow allowed 198 isotopically-resolved flux maps to be generated in a total (effective) experimental time of less than 33 days. This required a total human time of 194 h, or about 1 h per flux map. In the context of high-resolution fluxomics, the working volume of the bioreactors, 15 mL, seems an appropriate compromise between the low volumes required for HT investigations and culture volumes large enough for detailed biological or biochemical analysis. Indeed, this working volume allowed us to systematically collect all relevant physiological and biochemical data in a single bioreactor under fully controlled growth conditions, thereby decreasing biological variability and increasing data consistency. Further applications of this workflow to other experimental designs – e.g. parallel labeling experiments, other questions or other biological models (bacterial communities, eukaryotic cells, etc) should be facilitated by the flexibility of the setup, which is amenable to a variety of sample types (medium, cells, total broth), sampling methods, metabolite extraction processes (intra- or extracellular metabolites, biomass), and analytical techniques (basically any targeted/untargeted LC/GC-MS method or NMR sequence). This flexibility allows metabolic networks of various size and complexity, as well as other biological models, to be investigated. There is still scope for improvement in almost all stages of the workflow – from sampling to flux calculations and sensitivity analysis – to increase throughput and reduce experimental times and human involvement. A major challenge in the application of HT fluxomics is the level of precision on flux data that can be achieved at high or very high throughput. In this regard, the capability of the robotic platform to enable a high(er) number of biological replicate experiments to be carried out certainly contributes to increase the reliability of the data.

^{13}C-Fluxomics encompasses different approaches depending on the number of isotopic data collected and number of fluxes calculated [21, 23, 24]. These approaches include *isotopic profiling* (*i.e.* the statistical analysis of large isotopic datasets to discriminate variants), *targeted 13C-fluxomics* (in which a limited number of isotopic data are collected to measure fluxes in a few reactions or pathways), and *global ^{13}C-fluxomics* (the measurement of fluxes across complete metabolic networks from large isotopic datasets). The former two approaches are low-resolution methods because they provide no (isotopic profiling) or limited (targeted fluxomics)

flux information, though they are fully relevant for their respective purposes. High-resolution fluxomics refers to global ^{13}C-fluxomics, although the number of measured fluxes can vary significantly between studies (Table 1). HT fluxomics approaches have been discussed in terms of methodological challenges and biological value, but so far there has been no method proposed to evaluate the actual scale and quality of HT fluxomics investigations. This work introduces novel metrics for the actual scale and quality of HT fluxomics investigations. At the level of individual fluxotypes, key features of these studies are the total number of calculated fluxes (*fluxome resolution*), the number of isotopic data collected with respect to the number of fluxes to be measured (*isotopic resolutive power*), and the resulting precision in the calculated flux values (confidence intervals on flux values). At the level of the complete HT fluxomics investigation, three indices are introduced to measure the actual scale (*total flux dimension* and quality (*overall flux precision* and *overall pathway-specific precision*) of HT fluxomics investigations. Since HT studies aim to find the best compromise between throughput and information content (and quality), the number and quality of the resulting flux data have to be appreciated in terms of the actual size of the investigation, *i.e.* the total number of flux maps generated. HT fluxomics investigations are thereby characterized by high total flux *dimension* but low isotopic resolutive power and low flux precision [22]. On the contrary, high-resolution fluxomics is characterized by low total flux dimension but high isotopic resolutive power and high flux precision [39–41]. Finally, methods developed for the large-scale application of High-resolution fluxomics, such as the one described here, should aim to achieve good isotopic resolutive power and flux precision with high total flux dimension [33]. Indeed, we could apply High-resolution fluxomics (fluxome resolution of 94) at a high-throughput level (198 fluxotypes collected throughout the study) without loss of quality. It is important to note that these metrics are objective measures of the size and characteristics of fluxomics investigations; they do not quantify the biological value of the data.

The main objective of this work was to investigate whether the deletion of y-genes led to any changes in the distribution of metabolic fluxes in the central metabolism, *i.e.* in glucose metabolic fluxotypes. Deletion of the part of the y-ome considered in this study was associated with an overall tendency toward slower growth, but for the vast majority of the 180 strains investigated, the distribution of metabolic fluxes was remarkably similar to the one measured for the WT strain. Given that *E. coli* expressed the products of the all the investigated y-genes during growth on glucose, there are several possible explanations for these results. i) The protein is expressed but is not functional under the considered culture conditions, hence its absence has no impact in the deletion mutant. ii) The protein has a functional role under these

culture conditions but this function is not related to metabolism, hence its absence does not result in a metabolic phenotype. iii) The protein has a metabolic function –as an enzyme or as a metabolic regulator – in these culture conditions. In this case, the absence of a particular flux phenotype can be explained by the gene product having a quantitatively minor role, at least for the part of metabolism that was investigated, or by the onset of compensatory mechanisms that efficiently counterbalance the effects of the gene product's absence. Further investigations, including of fluxotypes under other experimental conditions, are required to answer these questions. However, regardless of the specific mechanisms involved for each y-gene, the data reported here highlight the robustness of *E. coli*'s central metabolism to the absence of all these gene products, whatever their roles, and to the potentially impairments caused by their absence.

In this work we were able to provide highly detailed functional information on the metabolic impact of y-gene deletion. Two strains, *ΔybjP* and *ΔydcS* showed interesting glucose fluxotypes, indicating that the products of the genes are involved in metabolism during growth on glucose as sole carbon source. Information on these genes is scarce and elusive. The gene *ydcS* was annotated as a putative putrescine ABC transporter in several databases and has also been proposed to encode poly-3-hydroxybutyrate synthase activity [43]. The gene *ybjP* has been predicted to be a lipoprotein [44]. Interestingly, the metabolic impact of both these genes is global since most of the pathways of the central carbon metabolism are affected. Indeed, both mutant strains show significant perturbations in the PPP and in the TCA cycle and related pathways. The deletions are also both associated with energy & redox status modifications, but with opposite consequences. The *ΔybjP* strain has significantly impaired ATP production and reduced cofactors within the central carbon metabolism while the *ΔybjP* has higher production levels than observed in the WT strain. The fact that the metabolic effects are so wide-ranging suggests that the products of these genes affect central metabolic parameters such as the redox status or that they are, either directly or indirectly, global regulators of metabolism. The dispensability of the two y-genes during growth on various carbon sources has recently been documented [12]. The *ΔydcS* strain shows reduced growth on certain sugars (mainly mannose, glucosamine, fructose) and two organic acids (alpha-ketoglutarate and acetate). The *ΔybjP* strain shows reduced growth on a few carbon sources, mainly including pyruvate, ribose, sorbitol and galacturonate. Deletion of these y-genes has no effect [12] or only a slight effect (this work) on growth with glucose as sole carbon source, in spite of significant metabolic changes in the two mutant strains. These observations emphasize that gene dispensability does not mean the gene lacks a metabolic function, but rather that the metabolism has successful adapted to the absence of the gene product [4]. This observation stresses the need to measure

the actual metabolic phenotype of mutants even if the gene deletion is silent in terms of growth. In this context, methods like HR fluxotyping are essential to achieve both the throughput and level of metabolic information required to investigate the role of y-genes in the metabolism of diverse carbon sources.

To conclude, this work shows that high-resolution fluxomics is amenable to high-throughput investigations and can provide detailed information on metabolic phenotypes. Our results also highlight how it can reveal the metabolic impact of gene deletion even if the gene has no known function or growth phenotype.

4. Materials and Methods

- ***Bacteria strains and cultivation conditions***

E. coli BW 25113 and the derived strains used in this study (listed in Supplementary data 1) were taken from the Keio collection (Baba et al. 2006). One hundred and eighty y-gene mutants were selected from the original glycerol stock of 3985 single-gene deletion mutants available in this collection (*E. coli* BW 25113 strains mutants; more details directly in paragraph 2.1 of the paper). The *E. coli* strain MG1655 was also used. All the selected strains were first cultivated overnight in LB medium (10 g/L tryptone, 5 g/L yeast extract and 10 g/L NaCl) with kanamycine (25 ug/ml) at 37°C and then stored in glycerol stocks.

To perform the experiments, the glycerol stocks of relevant strains were used to inoculate liquid LB medium in microplates. The LB cultures were used to inoculate precultures in minimal synthetic medium containing 17.4 $g.L^{-1}$ Na_2HPO_4,$12H_20$, 3.03 $g.L^{-1}$ KH_2PO_4, 0.51 $g.L^{-1}$ NaCl, 2.04 $g.L^{-1}$ NH_4Cl, 0.49 $g.L^{-1}$ $MgSO_4$, 4.38 $mg.L^{-1}$ $CaCl_2$, 15 $mg.L^{-1}$ Na_2EDTA $2H_2O$, 4.5 mg/L $ZnSO_4$ $7H_2O$, 0.3 $mg.L^{-1}$ $CoCl_2$ $6H_2O$, 1 $mg.L^{-1}$ $MnCl_2$ $4H_2O$, 1 $mg.L^{-1}$ H_3BO_3, 0.4 $mg.L^{-1}$ Na_2MoO_4 $2H_2O$, 3 $mg.L^{-1}$ $FeSO_4$ $7H_2O$, 0.3 $mg.L^{-1}$ $CuSO_4$ $5H_2O$, 0.1 $g.L^{-1}$ thiamine and 3 $g.L^{-1}$ glucose. For the ^{13}C-labeling experiments, unlabeled glucose was replaced by the same concentration of a mixture of 80% [1-$^{13}C_1$]-D-glucose + 20% [U-$^{13}C_6$]-D-glucose. To minimize sources of unlabeled carbon atoms from the first cultivation steps in subsequent experiments, the cells were inoculated at a starting OD between 0.08 and 0.12.

- ***Robotic platforms for culture, sampling and sample preparation***

Two robotic platforms were used to parallelize the cell cultures, ^{13}C-labelling experiments, sampling of labeled metabolites, and sample preparation for NMR and MS

analyses. The first system allowed 48 ^{13}C-labelling experiments to be performed automatically in parallel in 15 ml bioreactors under controlled growth conditions, with automated collection of labelled samples (of biomass or culture medium) at defined culture times or optical densities. The robot and its operation are described in detail by Heux et al. [22]. A second robotic workstation (Freedom EVO 200, Tecan) was designed and assembled to fully automate and parallelize the final preparation of labeled biological samples for the analysis of isotopic profiles by NMR or mass spectrometry. This device was used to handle the different containers used (NMR tubes, vials, multiwell-plates, etc.), to perform dilutions, add standards, take aliquots, and manage the samples.

The ^{13}C-labeling experiments were carried out in batches of 48 parallel cell cultures using the two robotic platforms described above. The input label was optimized for *E. coli*'s metabolic network using IsoDesign (1.2.1) [24], and consisted of a mixture of 80% [1-$^{13}C_1$]-D-glucose and 20% [U-$^{13}C_6$]-D-glucose. All cultures were performed in 15 mL reaction vessels, at 37 °C, pH = 7, a stirring speed of 2300 rpm and with 5L.min^{-1} of compressed air fed into the culture module. Four to eight supernatant samples were collected in each reactor throughout the culture to analyze the exometabolome and to calculate the rates of substrate consumption and end-product production. The biomass was automatically sampled once the OD_{600nm} had reached 1.2, indicating that the metabolic and isotopic steady-state had been achieved: 4 ml samples of culture were pelleted by automated centrifugation (5 min / 4410 g), manually hydrolyzed with 250 µL 6 N HCl for 15 h at 110 °C and washed twice in 1 mL D_2O by rotary evaporation (Büchi Labortechnik AG, CH) between each washing step. Aliquots (10 µL) of biomass hydrolysate supernatant were then collected and transferred into 96 well plates, diluted with 990 µL pure H_2O and transferred to vials for LC-MS analysis. Samples of biomass hydrolysate supernatant (150 µL) were also mixed with 50 µL of TSP d4 (4 mM in D_2O) and 150 µL of each mixture was then transferred into 3 mm NMR tubes.

- ***Isotopic profiling of proteinogenic amino-acids***

The incorporation of ^{13}C-label into proteinogenic amino acids (listed in Supplementary data 6) was analyzed by liquid chromatography–mass spectrometry, using an Ultimate 3000 HPLC system (Dionex, CA, USA) coupled to an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) equipped with a heated electrospray ionization probe described in detail by Heuillet et al. 2018 [45]. Full scan HRMS analyses were performed in positive FTMS mode at a resolution of 60 000 (at 400 m/z), using the following source

parameters: capillary temperature, 275 °C; source heater temperature, 250 °C; sheath gas flow rate, 45 a.u. (arbitrary unit); auxiliary gas flow rate, 20 a.u.; S-Lens RF level, 40 %; source voltage, 5 kV. Metabolites were identified by extracting the exact mass with a tolerance of 5 ppm. The raw MS isotopic profiles of proteinogenic amino acids were then quantified using Tracefinder (Thermo Fisher Scientific, Waltham, MA, USA). The isotopic profiles (Carbon Isotopologue Distributions) were obtained after correcting for natural isotopic abundances using IsoCor 1.2 [27,28] (https://github.com/MetaSys-LISBP/IsoCor). The raw MS data are available from Metabolights (accession number MTBLS2188).

● ***NMR analysis of extracellular medium***

Culture supernatants were analyzed by 1D-^{1}H NMR on a Bruker Avance III 800MHz spectrometer equipped with a 5 mm CQPI cryoprobe at 280 K. To precisely quantify the extracellular compounds, the 1D-^{1}H NMR data were recorded after a 30° presaturation pulse (zgpr30), with a relaxation delay of 7 s. NMR spectra were processed using Topspin 3.5pl6 (Bruker BioSpin, Rheinstetten, Germany). The raw NMR data are available from Metabolights (accession number MTBLS2188).

The rate of glucose consumption and the rates of acetate and lactate production were quantified by analyzing the 4–8 samples collected during the cell culture period. The ^{13}C labeling profiles of acetate and lactate, including positional information on label incorporation, were also measured for the flux calculations.

● ***Growth parameters***

The growth parameters calculated from the experimental datasets for each culture were the growth rate, glucose uptake rate, acetate and lactate production rates, and biomass yield. The growth rate was calculated from OD values measured during the culture period. The rates of glucose consumption and of acetate and lactate production were calculated from the exometabolome data. All calculations were performed using Physiofit (v0.9) [46], a software designed to determine growth parameters by fitting time-course data. The software includes a batch calculation mode which allows calculations on large series of datasets, and which was used to calculate the growth parameters for all considered experimental conditions. The biomass yields were calculated using the conversion factor 0.378 gDW/ OD600nm unit.

Extracellular fluxes were determined from the rates of disappearance (or appearance) of substrates and products, in the culture supernatants, as measured by NMR.

● ***Flux calculation and visualization***

Fluxes were calculated using the software influx_si 4.1 [30] (https://metasys.insa-toulouse.fr/software/influx/), including the mass balances and carbon atom transitions of the biochemical reaction network. The metabolic network contained the main pathways of *E. coli*'s central metabolism: glycolysis (EMP), the pentose phosphate pathway (PPP), the Entner-Doudoroff (ED) pathway, the tricarboxylic acid cycle (TCA), and anaplerotic reactions, the glyoxylate shunt, and the reactions for amino acid biosynthesis [32]. Intracellular fluxes were estimated from measurements of extracellular fluxes and from the ^{13}C-labelling patterns of metabolites using appropriate mathematical models of glucose metabolism in *E. coli* [32]. Labeling data were collected from intracellular metabolites by IC-MS/MS and from metabolic end-products by 1D ^{1}H NMR, as detailed above. The fluxes were normalized to the rate of substrate uptake, which was arbitrarily set at 100.

Metabolic fluxes were calculated for each of the 198 experimental conditions considered in the study. For each culture condition, the information required for the flux calculation (*i.e.* metabolic network, isotopic data, etc.) was written in a specific FTBL file [16], which was then submitted to influx_si.

The FTBL model was converted by an influx_si module (ftbl2metxml.py) into Systems Biology Markup Language (SBML, in xml format), readable by the online application Metexplore (http://metexplore.toulouse.inra.fr). Metexplore was used to visualize flux maps for strains of interest [47].

● ***Sensitivity analysis***

Confidence intervals for the calculated fluxes were determined using a Monte-Carlo approach in which 100 independent optimization runs were performed on datasets with noise added in proportion to standard measurement errors. The isotopic data and metabolic fluxes for each independent biological replicate are provided in the Supplemental data (dataset S1).

● ***Statistical analysis of flux datasets***

Flux maps were compared by PCA using the collaborative portal Workflow for Metabolomics (W4M) [36] . Results were visualized as boxplots and figures were generated using the software R.

- ***Meta-data management***

The data and meta-data generated throughout the workflow were stored, managed, and processed as described below.

Culture and sampling data related to the parallel cultures of *E. coli* strains with ^{13}C-labelled compounds and to the collection of samples: settings and data from the robotic culture platform (Tecan Evo 200) were automatically and separately recorded for each of the 48 bioreactors: pH values, O2 values, OD_{600nm} values, volumes, sampling times, and tube locations and associated barcodes. The data were linked and stored in an SQL database, making them traceable to each robotic run. The most important elements of this database are listed in Supplementary data 3 (mu, batch number, batch position, biomass (mg/L) at sampling time). A log file was also automatically generated for each robotic run, detailing errors and software information.

Sample preparation data: all the steps performed on the sample preparation platform (Tecan Evo 200) were automatically recorded in an SQL database: volumes, tube locations on the worktable, associated barcodes and sampling times. The data were visualized and previous runs reexamined using the robot's traceability software, and a log file was automatically generated for each run with errors and software information.

Analytical data: all mass spectrometry and NMR data were saved on a server to guarantee sample traceability. This included the raw MS data, the MS data processed with Tracefinder, the MS data processed with Isocor, the raw NMR data, the NMR data processed with TopSpin, and the NMR data processed with Physiofit. Corrected CID values, extracellular fluxes and the positional enrichment of extracellular metabolites are listed in Supplementary data 3.

Flux calculations: all flux calculation files were stored on a server. This included the input (FTBL) files for influx_si, and the output (KVH) files. The most important elements are included in Supplementary data 3 (chi-square tests of the fits, fit cost functions and net fluxes).

Statistical analysis and flux mapping: all the data used and generated for statistical analysis and flux mapping were stored on a server.

- ***Calculation of cofactor production and consumption rates***

The production rates ($mmol.gDW^{-1}.h^{-1}$) of NADPH, $NADH/FADH_2$, and ATP in the central carbon metabolism were calculated as the sum of the estimated fluxes of the reactions that are expected to produce (positive value) or consume (negative value) the cofactors. Values

were averaged over the biological replicates. NADPH production: Pentose-Phosphate Pathway (PPP) = glucose 6-phosphate dehydrogenase (zwf) + phosphogluconate dehydrogenase (gnd); TCA cycle = isocitrate dehydrogenase (idh); malic enzyme (mae). NADH production: glycolysis= glyceraldehyde-3-phosphate dehydrogenase (flux equal to pgk); Oxidative metabolism = pyruvate dehydrogenase (pdh) + TCA cycle enzymes, including alpha-ketoglutarate dehydrogenase (akgdh) + malate dehydrogenase (maldh); malic enzyme (mae) and biomass formation. NADH consumption: lactate dehydrogenase (out_Lac). $FADH_2$ production: TCA cycle = succinate dehydrogenase (flux equal to fum). ATP production: glycolysis = phosphoglycerate kinase (pgk) + pyruvate kinase (pyk); TCA cycle = succinyl-CoA synthetase (assimilated to akgdh); acetate metabolism = acetate kinase (ack, assimilated to out_Ac). ATP consumption: glycolysis = glucose PTS system (Glucupt_U + Glucupt_1) (assuming 1 consumed PEP is equivalent to 1 consumed) + 6-phosphofructokinase (pfk); ppc (equivalent to pep carboxylase (forward flux) and pep carboxykinase (reverse flux).

References

1. Hanson, A.D.; Pribat, A.; Waller, J.C.; de Crécy-lagard, V. 'Unknown' Proteins and 'Orphan' Enzymes: The Missing Half of the Engineering Parts List – and How to Find It. *Biochem. J.* **2009**, *425*, 1–11, doi:10.1042/BJ20091328.
2. Hu, P.; Janga, S.C.; Babu, M.; Díaz-Mejía, J.J.; Butland, G.; Yang, W.; Pogoutse, O.; Guo, X.; Phanse, S.; Wong, P.; et al. Global Functional Atlas of Escherichia Coli Encompassing Previously Uncharacterized Proteins. *PLoS Biol.* **2009**, *7*, e1000096, doi:10.1371/journal.pbio.1000096.
3. Ghatak, S.; King, Z.A.; Sastry, A.; Palsson, B.O. The Y-Ome Defines the Thirty-Four Percent of Escherichia Coli Genes That Lack Experimental Evidence of Function. *bioRxiv* **2018**, 328591, doi:10.1101/328591.
4. Raamsdonk, L.M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M.C.; Berden, J.A.; Brindle, K.M.; Kell, D.B.; Rowland, J.J.; et al. A Functional Genomics Strategy That Uses Metabolome Data to Reveal the Phenotype of Silent Mutations. *Nat. Biotechnol.* **2001**, *19*, 45–50, doi:10.1038/83496.
5. Fuhrer, T.; Zampieri, M.; Sévin, D.C.; Sauer, U.; Zamboni, N. Genomewide Landscape of Gene–Metabolome Associations in Escherichia Coli. *Mol. Syst. Biol.* **2017**, *13*, doi:10.15252/msb.20167150.
6. Sévin, D.C.; Fuhrer, T.; Zamboni, N.; Sauer, U. Nontargeted *in Vitro* Metabolomics for High-Throughput Identification of Novel Enzymes in *Escherichia Coli*. *Nat. Methods* **2017**, *14*, 187–194, doi:10.1038/nmeth.4103.
7. Blaby-Haas, C.E.; de Crécy-Lagard, V. Mining High-Throughput Experimental Data to Link Gene and Function. *Trends Biotechnol.* **2011**, *29*, 174–182, doi:10.1016/j.tibtech.2011.01.001.
8. Long, C.P.; Antoniewicz, M.R. Metabolic Flux Analysis of Escherichia Coli Knockouts: Lessons from the Keio Collection and Future Outlook. *Curr. Opin. Biotechnol.* **2014**, *28*, 127–133, doi:10.1016/j.copbio.2014.02.006.
9. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K.A.; Tomita, M.; Wanner, B.L.; Mori, H. Construction of Escherichia Coli K-12 In-frame, Single-gene Knockout Mutants: The Keio Collection. *Mol. Syst. Biol.* **2006**, *2*, 2006.0008, doi:10.1038/msb4100050.
10. Joyce, A.R.; Reed, J.L.; White, A.; Edwards, R.; Osterman, A.; Baba, T.; Mori, H.; Lesely, S.A.; Palsson, B.Ø.; Agarwalla, S. Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia Coli. *J. Bacteriol.* **2006**, *188*, 8259–8271, doi:10.1128/JB.00740-06.
11. Guzmán, G.I.; Olson, C.A.; Hefner, Y.; Phaneuf, P.V.; Catoiu, E.; Crepaldi, L.B.; Micas, L.G.; Palsson, B.O.; Feist, A.M. Reframing Gene Essentiality in Terms of Adaptive Flexibility. *BMC Syst. Biol.* **2018**, *12*, doi:10.1186/s12918-018-0653-z.
12. Tong, M.; French, S.; El Zahed, S.S.; Ong, W. kit; Karp, P.D.; Brown, E.D. Gene Dispensability in Escherichia Coli Grown in Thirty Different Carbon Environments. *mBio* **2020**, *11*, doi:10.1128/mBio.02259-20.
13. Warner, J.R.; Reeder, P.J.; Karimpour-Fard, A.; Woodruff, L.B.A.; Gill, R.T. Rapid Profiling of a Microbial Genome Using Mixtures of Barcoded Oligonucleotides. *Nat. Biotechnol.* **2010**, *28*, 856–862, doi:10.1038/nbt.1653.
14. Liu, X.; Gallay, C.; Kjos, M.; Domenech, A.; Slager, J.; van Kessel, S.P.; Knoops, K.; Sorg, R.A.; Zhang, J.-R.; Veening, J.-W. High-Throughput CRISPRi Phenotyping Identifies New Essential Genes in Streptococcus Pneumoniae. *Mol. Syst. Biol.* **2017**, *13*, 931, doi:10.15252/msb.20167449.

15. Price, M.N.; Wetmore, K.M.; Waters, R.J.; Callaghan, M.; Ray, J.; Liu, H.; Kuehl, J.V.; Melnyk, R.A.; Lamson, J.S.; Suh, Y.; et al. Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function. *Nature* **2018**, *557*, 503–509, doi:10.1038/s41586-018-0124-0.
16. Wiechert, W. 13C Metabolic Flux Analysis. *Metab. Eng.* **2001**, *3*, 195–206, doi:10.1006/mben.2001.0187.
17. Sauer, U. Metabolic Networks in Motion: 13C-based Flux Analysis. *Mol. Syst. Biol.* **2006**, *2*, 62, doi:10.1038/msb4100109.
18. Wittmann, C.; Portais, J.-C. Metabolic Flux Analysis. In *Metabolomics in Practice*; John Wiley & Sons, Ltd, 2013; pp. 285–312 ISBN 978-3-527-65586-1.
19. Ellis, D.I.; Goodacre, R. Metabolomics-Assisted Synthetic Biology. *Curr. Opin. Biotechnol.* **2012**, *23*, 22–28, doi:10.1016/j.copbio.2011.10.014.
20. Fan, J.; Ye, J.; Kamphorst, J.J.; Shlomi, T.; Thompson, C.B.; Rabinowitz, J.D. Quantitative Flux Analysis Reveals Folate-Dependent NADPH Production. *Nature* **2014**, *510*, 298–302, doi:10.1038/nature13236.
21. Heux, S.; Bergès, C.; Millard, P.; Portais, J.-C.; Létisse, F. Recent Advances in High-Throughput 13C-Fluxomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 104–109, doi:10.1016/j.copbio.2016.10.010.
22. Heux, S.; Poinot, J.; Massou, S.; Sokol, S.; Portais, J.-C. A Novel Platform for Automated High-Throughput Fluxome Profiling of Metabolic Variants. *Metab. Eng.* **2014**, *25*, 8–19, doi:10.1016/j.ymben.2014.06.001.
23. Sauer, U. High-Throughput Phenomics: Experimental Methods for Mapping Fluxomes. Current Opinion in Biotechnology 2004, 15, 58–63, doi:10.1016/j.copbio.2003.11.001.
24. Moxley, J.F.; Jewett, M.C.; Antoniewicz, M.R.; Villas-Boas, S.G.; Alper, H.; Wheeler, R.T.; Tong, L.; Hinnebusch, A.G.; Ideker, T.; Nielsen, J.; et al. Linking High-Resolution Metabolic Flux Phenotypes and Transcriptional Regulation in Yeast Modulated by the Global Regulator Gcn4p. PNAS 2009, 106, 6477–6482, doi:10.1073/pnas.0811091106.
25. Schmidt, A.; Kochanowski, K.; Vedelaar, S.; Ahrné, E.; Volkmer, B.; Callipo, L.; Knoops, K.; Bauer, M.; Aebersold, R.; Heinemann, M. The Quantitative and Condition-Dependent Escherichia Coli Proteome. *Nat. Biotechnol.* **2016**, *34*, 104–110, doi:10.1038/nbt.3418.
26. Millard, P.; Sokol, S.; Letisse, F.; Portais, J.-C. IsoDesign: A Software for Optimizing the Design of 13C-Metabolic Flux Analysis Experiments. *Biotechnol. Bioeng.* **2014**, *111*, 202–208, doi:10.1002/bit.24997.
27. Millard, P.; Letisse, F.; Sokol, S.; Portais, J.-C. IsoCor: Correcting MS Data in Isotope Labeling Experiments. *Bioinforma. Oxf. Engl.* **2012**, *28*, 1294–1296, doi:10.1093/bioinformatics/bts127.
28. Millard, P.; Delépine, B.; Guionnet, M.; Heuillet, M.; Bellvert, F.; Létisse, F. IsoCor: Isotope Correction for High-Resolution MS Labeling Experiments. *Bioinformatics* **2019**, *35*, 4484–4487, doi:10.1093/bioinformatics/btz209.
29. Neidhardt, F.C.; Curtiss, R. *Escherichia Coli and Salmonella: Cellular and Molecular Biology*; ASM Press: Washington, D.C., 1996; ISBN 978-1-55581-084-9.
30. Sokol, S.; Millard, P.; Portais, J.-C. Influx_s: Increasing Numerical Stability and Precision for Metabolic Flux Analysis in Isotope Labelling Experiments. *Bioinforma. Oxf. Engl.* **2012**, *28*, 687–693, doi:10.1093/bioinformatics/btr716.
31. Nicolas, C.; Kiefer, P.; Letisse, F.; Krömer, J.; Massou, S.; Soucaille, P.; Wittmann, C.; Lindley, N.D.; Portais, J.-C. Response of the Central Metabolism of Escherichia Coli to Modified Expression of the Gene Encoding the Glucose-6-Phosphate Dehydrogenase. *FEBS Lett.* **2007**, *581*, 3771–3776, doi:https://doi.org/10.1016/j.febslet.2007.06.066.

32. Millard, P.; Massou, S.; Wittmann, C.; Portais, J.-C.; Létisse, F. Sampling of Intracellular Metabolites for Stationary and Non-Stationary 13C Metabolic Flux Analysis in Escherichia Coli. *Anal. Biochem.* **2014**, *465*, 38–49, doi:10.1016/j.ab.2014.07.026.
33. Haverkorn van Rijsewijk, B.R.B.; Nanchen, A.; Nallet, S.; Kleijn, R.J.; Sauer, U. Large-Scale 13C-Flux Analysis Reveals Distinct Transcriptional Control of Respiratory and Fermentative Metabolism in Escherichia Coli. *Mol. Syst. Biol.* **2011**, *7*, 477, doi:10.1038/msb.2011.9.
34. Zhao, J.; Baba, T.; Mori, H.; Shimizu, K. Effect of Zwf Gene Knockout on the Metabolism of Escherichia Coli Grown on Glucose or Acetate. *Metab. Eng.* **2004**, *6*, 164–174, doi:10.1016/j.ymben.2004.02.004.
35. Hua, Q.; Yang, C.; Baba, T.; Mori, H.; Shimizu, K. Responses of the Central Metabolism in Escherichia Coli to Phosphoglucose Isomerase and Glucose-6-Phosphate Dehydrogenase Knockouts. *J. Bacteriol.* **2003**, *185*, 7053–7067, doi:10.1128/jb.185.24.7053-7067.2003.
36. Giacomoni, F.; Le Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; et al. Workflow4Metabolomics: A Collaborative Research Infrastructure for Computational Metabolomics. Bioinformatics 2015, 31, 1493–1495, doi:10.1093/bioinformatics/btu813.
37. Sauer, U.; Canonaco, F.; Heri, S.; Perrenoud, A.; Fischer, E. The Soluble and Membrane-Bound Transhydrogenases UdhA and PntAB Have Divergent Functions in NADPH Metabolism of Escherichia Coli. *J. Biol. Chem.* **2004**, *279*, 6613–6619, doi:10.1074/jbc.M311657200.
38. Long, C.P.; Antoniewicz, M.R. High-Resolution 13C Metabolic Flux Analysis. *Nat. Protoc.* **2019**, *14*, 2856–2877, doi:10.1038/s41596-019-0204-0.
39. Lange, A.; Becker, J.; Schulze, D.; Cahoreau, E.; Portais, J.-C.; Haefner, S.; Schröder, H.; Krawczyk, J.; Zelder, O.; Wittmann, C. Bio-Based Succinate from Sucrose: High-Resolution 13C Metabolic Flux Analysis and Metabolic Engineering of the Rumen Bacterium Basfia Succiniciproducens. *Metab. Eng.* **2017**, *44*, 198–212, doi:10.1016/j.ymben.2017.10.003.
40. Leighty, R.W.; Antoniewicz, M.R. COMPLETE-MFA: Complementary Parallel Labeling Experiments Technique for Metabolic Flux Analysis. *Metab. Eng.* **2013**, *20*, 49–55, doi:10.1016/j.ymben.2013.08.006.
41. Crown, S.B.; Long, C.P.; Antoniewicz, M.R. Integrated 13C-Metabolic Flux Analysis of 14 Parallel Labeling Experiments in Escherichia Coli. *Metab. Eng.* **2015**, *28*, 151–158, doi:10.1016/j.ymben.2015.01.001.
42. Long, C.P.; Au, J.; Gonzalez, J.E.; Antoniewicz, M.R. 13C Metabolic Flux Analysis of Microbial and Mammalian Systems Is Enhanced with GC-MS Measurements of Glycogen and RNA Labeling. *Metab. Eng.* **2016**, *38*, 65–72, doi:10.1016/j.ymben.2016.06.007.
43. Dai, D.; Reusch, R.N. Poly-3-Hydroxybutyrate Synthase from the Periplasm of Escherichia Coli. *Biochem. Biophys. Res. Commun.* **2008**, *374*, 485–489, doi:10.1016/j.bbrc.2008.07.043.
44. Juncker, A.S.; Willenbrock, H.; Von Heijne, G.; Brunak, S.; Nielsen, H.; Krogh, A. Prediction of Lipoprotein Signal Peptides in Gram-Negative Bacteria. *Protein Sci. Publ. Protein Soc.* **2003**, *12*, 1652–1662, doi:10.1110/ps.0303703.
45. Heuillet, M.; Bellvert, F.; Cahoreau, E.; Letisse, F.; Millard, P.; Portais, J.-C. Methodology for the Validation of Isotopic Analyses by Mass Spectrometry in Stable-Isotope Labeling Experiments. *Anal. Chem.* **2018**, *90*, 1852–1860, doi:10.1021/acs.analchem.7b03886.
46. Peiro, C.; Millard, P.; de Simone, A.; Cahoreau, E.; Peyriga, L.; Enjalbert, B.; Heux, S. Chemical and Metabolic Controls on Dihydroxyacetone Metabolism Lead to Suboptimal

Growth of Escherichia Coli. *Appl. Environ. Microbiol.* **2019**, *85*, doi:10.1128/AEM.00768-19.

47. Chazalviel, M.; Frainay, C.; Poupin, N.; Vinson, F.; Merlet, B.; Gloaguen, Y.; Cottret, L.; Jourdan, F. MetExploreViz: Web Component for Interactive Metabolic Network Visualization. *Bioinformatics* **2018**, *34*, 312–313, doi:10.1093/bioinformatics/btx588

Supplementary Data

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/metabo11050271/s1,

Supplementary data 1: List of the 180 E. coli BW 25113 strains selected for the fluxomics experiments

Supplementary data 2: Metabolic model considered in the flux calculations

Supplementary data 3: Complete dataset obtained for all the fluxomic experiments performed in the study

Supplementary data 4: Calculation of ATP and redox fluxes

Supplementary data 5: Isotopic data collected by LC HRMS and NMR.

Conclusions et perspectives

Les travaux présentés dans cette thèse s'inscrivent dans le domaine de la biologie des systèmes. Ils sont basés sur le développement de méthodologies et d'outils informatiques innovants permettant d'étudier le métabolisme d'organismes vivants à l'échelle du système.

Nous avons dans un premier temps mis en place une méthodologie permettant d'évaluer la qualité de données isotopiques, basée sur l'utilisation d'un standard marqué présentant l'ensemble du profil isotopique des métabolites, Cette méthodologie permet d'assurer le traitement optimal des données de masse complexes dans les études de traçage isotopique non ciblé du métabolisme. Avec les progrès de l'instrumentation MS et des méthodes analytiques, qui permettent d'étendre la couverture du métabolome, l'application de la méthodologie proposée maximise la valeur biologique des études de traçage isotopique en révélant l'ensemble des informations métaboliques encodées dans les profils de marquage des métabolites.

Dans la deuxième partie de cette thèse, nous avons développé une approche de fluxomique qui vise à valider des données expérimentales à partir de topologies de sous-réseaux métaboliques proposées. Il s'agit d'une approche « data-driven » basée sur des données de marquage isotopique en dynamique acquise par spectrométrie de masse haute résolution. Ce travail a mis en avant deux stratégies innovantes. La première consiste à construire des sous-réseaux minimum à partir de données expérimentales HRMS et d'un ensemble de réactions biochimiques pouvant avoir lieu dans un organisme. La deuxième consiste au développement de stratégies de simulation des cinétiques de marquage isotopique permettant de valider la consistance des sous-réseaux sur la base des cinétiques de marquage de leur substrat. Ces deux approches ont pu être mises en place grâce au développement d'un outil bioinformatique robuste et puissant. Elles ont été validées sur la base du modèle simplifié d'*E. coli* core. L'objectif final de cette approche de fluxomique est d'accéder directement au réseau métabolique actif d'un organisme dans un contexte donné, sur la seule base de données expérimentales de marquage isotopique non-ciblées, de manière orthogonale et complémentaire aux approches de reconstruction *in silico.* Cette approche est générique et peut-être appliquée à d'autres organismes complexes, notamment des organismes compartimentés (cellules mammifères). A partir des travaux initiés au cours de ce travail de thèse, elle sera appliquée dans un premier temps pour l'étude du métabolisme d'*E. coli* à plus large échelle.

Une perspective d'application biologique plus complexe est l'étude du métabolisme des adipocytes. Des travaux pionniers réalisés au sein du laboratoire RESTORE ont montré que des métabolites associés au métabolisme redox, comme le lactate et les corps cétoniques, peuvent induire la différenciation adipocytaire et le brunissement du tissu adipeux blanc en induisant fortement la protéine découplante UCP1[1]. Bien que cette différenciation redox-dépendante

puisse jouer un rôle important dans l'homéostasie énergétique de l'organisme, elle n'a été pour l'instant que très peu étudiée. Cette approche pourra être appliquée dans l'objectif d'identifier les modifications métaboliques induites par les composés redox, ainsi que pour caractériser le devenir métabolique de ces composés eux-mêmes. L'intérêt étant notamment de permettre une identification sans *à priori* des voies métaboliques impliquées dans l'utilisation des nutriments au cours de cette reprogrammation adipocytaire.

En parallèle de ces développements, une approche de fluxomique à haut-débit a été développée, pour permettre l'optimisation et l'automatisation de l'ensemble du processus de fluxomique à hau-débit et l'intégration d'outils pertinents. Comme nous l'avons vu dans ce travail de thèse, la fluxomique basée sur les expériences de marquage est de plus en plus utilisée dans les sciences biologiques pour identifier les voies métaboliques, aider à la découverte de nouvelles interactions de régulation ou quantifier les réponses des flux aux perturbations environnementales ou génétiques. Il s'agit d'un outil essentiel pour répondre à des questions fondamentales appliquées dans le domaine de la santé et de la biotechnologie. Cependant il s'agit d'un processus long et complexe qui nécessite un grand nombre d'étapes expérimentales et informatiques et inclut une grande variété d'approches (ex : ^{13}C-MFA stationnaire, instationnaire, dynamique …). Ces différents facteurs limitent globalement le débit et la reproductibilité des études de flux C13, notamment au niveau informatique (conception expérimentale, traitement des données et calcul des flux). Une autre perspective est le développement d'une plateforme informatique automatisée comprenant l'ensemble du protocole de calcul de flux. L'objectif de ce projet est d'améliorer l'accessibilité et la robustesse des approches de fluxomique à haut débit.

Référence bibliographique

[1]Carrière A, Jeanson Y, Berger-Müller S, André M, Chenouard V, Arnaud E, Barreau C, Walther R, Galinier A, Wdziekonski B, Villageois P, Louche K, Collas P, Moro C, Dani C, Villarroya F, Casteilla L. Browning of white adipose cells by intermediate metabolites: an adaptive mechanism to alleviate redox pressure. Diabetes. 2014 Oct;63(10):3253-65

Annexes

Annexe 1 : Illustrations des différents analyseurs de masse

1. Analyseurs à basse résolution

- ***L'analyseur quadripolaire (Q)***

L'analyseur quadripolaire est le plus simple des analyseurs de spectrométrie de masse. Il est composé de quatre électrodes parallèles auxquelles sont appliqués deux potentiels déphasés de 180°, l'un continu, l'autre alternatif (figure S1). Ces potentiels créent un champ quadripolaire qui guide les ions de la source d'ions vers l'analyseur quadripolaire. Au sein de ce champ, seuls les ions présentant une trajectoire stable sont conduits jusqu'à l'extrémité de l'analyseur. Les ions ayant une trajectoire instable sont arrêtés lorsqu'ils entrent en collision avec les électrodes. Le domaine de stabilité des ions au sein du quadripôle dépend de leur rapport m/z et des tensions appliquées aux électrodes permettant ainsi de sélectionner uniquement les ions appartenant à une gamme de masse définie.

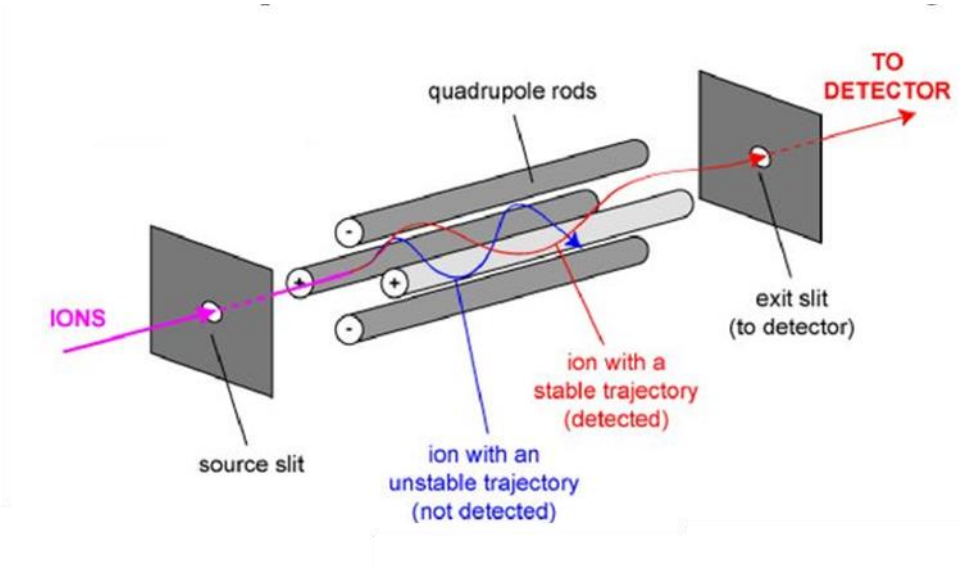


Figure S1 : Principe de fonctionnement d'un analyseur quadripolaire

- ***Le piège à ions linéaire (ion trap)***

Un piège à ion linéaire est un analyseur composé de quatre éléctrodes entre lesquelles les ions sont piégés (figure S2). Les ions sont piégés radialement en appliquant un champ de radiofréquence et axialement en utilisant un potentiel statique sur les électrodes d'extrémité. Ils sont ensuite éjectés du piège par modulation des radiofréquences appliquées. Une gamme de radiofréquence est appliquée au piège afin de balayer toute la gamme de masse étudiée.

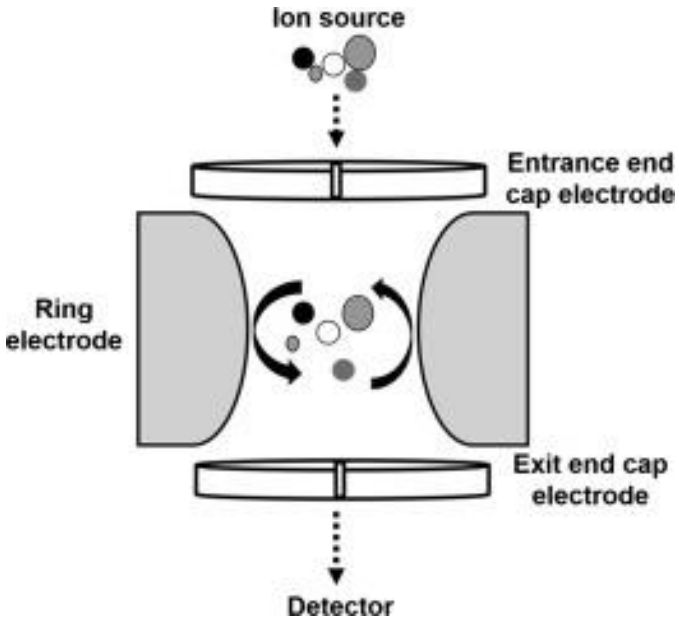


Figure S2 : Principe de fonctionnement d'un piège à ions linéaire. (Stephanie N. Thomas, 2019)

2. Analyseurs à haute résolution

- ***L'analyseur à temps de vol (TOF)***

Dans les analyseurs à temps de vol, les ions sont introduits orthogonalement au tube de vol et sont accélérés par l'application d'une différence de potentiel entre une électrode et une grille d'extraction. Les ions de même charge reçoivent alors la même énergie cinétique et se déplacent donc avec une vitesse inversement proportionnelle au carré de leur rapport m/z. Après un parcours en vol libre dans le tube de vol, les ions entrent dans un réflectron qui, par application d'un champ électrique graduel, va inverser la direction des ions (figure S3). La mesure du temps de vol de chaque ion permet de déterminer leur rapport masse/charge.

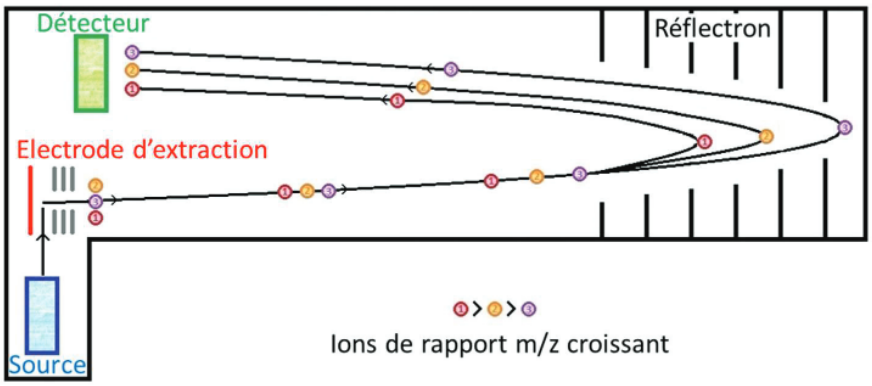


Figure S3 : Principe de fonctionnement d'un analyseur à temps de vol (Barbier Saint Hilaire, 2019)

- ***L'analyseur Orbitrap***

L'analyseur orbitrap est basé sur le piégeage d'ions et sur l'enregistrement d'un courant induit converti en gamme de fréquences par transformée de Fourier. Le dispositif de piégeage est constitué d'un câble chargé placé le long de l'axe d'une électrode centrale en forme de fuseau, entourée d'une électrode cylindrique externe (figure S4). Les ions formés par l'application d'une décharge à l'intérieur du cylindre sont attirés par le câble central qui établit un champ électrostatique. Les ions qui possèdent une vitesse tangentielle suffisamment élevée tournent autour du câble plutôt que d'entrer directement en collision avec lui. La fréquence d'oscillation axiale de ces ions est variable en fonction de leur rapport m/z. Les électrodes externes sont utilisées comme plaques réceptrices pour la détection du courant induit par ces oscillations axiales. Un spectre de masse est généré en utilisant la transformée de Fourier pour transformer les fréquences en m/z et les amplitudes en intensités.

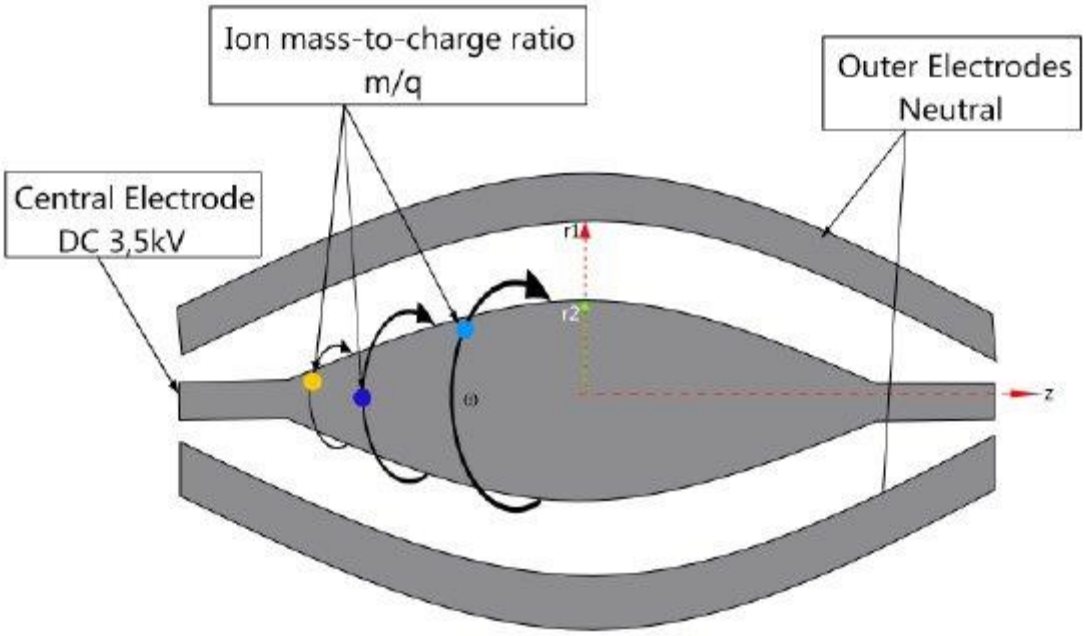


Figure 58 : Principe de fonctionnement d'un analyseur orbitrap (Ronaldo Ferreira do Nascimento, 2017)

- ***L'analyseur à résonance cyclonique ionique (FT-ICR)***

Dans les analyseurs à résonance cyclonique ionique, le rapport masse/charge (m/z) d'un ion peut être déterminé expérimentalement en mesurant la fréquence à laquelle l'ion se transforme dans un champ magnétique. Ce champ magnétique piège les ions au sein d'une cellule constituée de deux paires de plaques (plaques émettrices et plaques réceptrices). Ces plaques sont situées dans un plan parallèle au champ magnétique (figure S5). Soumis à ce champ magnétique, les ions vont alors adopter une trajectoire circulaire (principe du cyclotron) dans un plan perpendiculaire à l'axe du champ magnétique. Les ions de même rapport m/z ont une même fréquence cyclotronique mais vont se mouvoir indépendamment dans le centre du piège. Pour augmenter l'efficacité et la résolution de la détection, les ions sont excités par une impulsion radiofréquence jusqu'à obtenir un rayon cyclotronique plus grand. La fréquence cyclotronique est alors mesurée en détectant un "courant d'image" qui est induit lorsque les ions passent devant les plaques réceptrices. La transformée de Fourier de la décroissance de l'induction libre convertit le signal en spectre de masse.

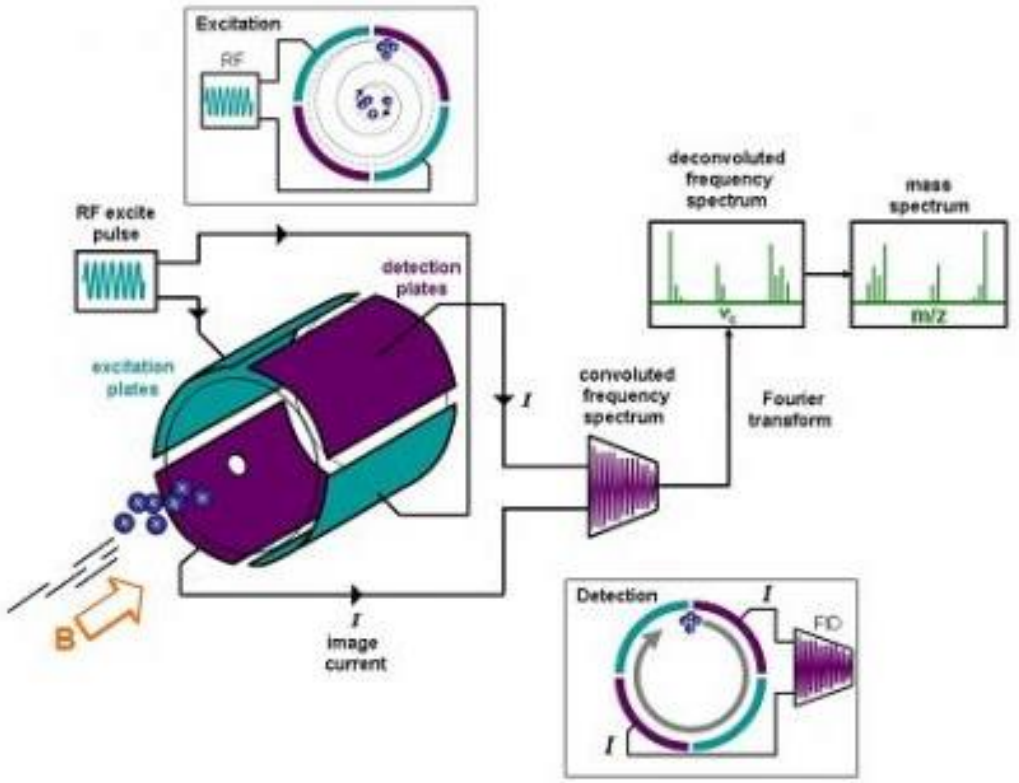


Figure S5 : Principe de fonctionnement d'un analyseur à résonance cyclonique ionique (Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," 2019)

Annexe 2 : Analyse comparative des logiciels de profilage isotopique non ciblé

		X13CMS	geoRge	MetExtract II	mzMatchIso	DynaMet	HiResTec	NFTD	MIA
Objective	annotation, quantification, isotopic profiling, comparative labeling*, fluxomics, pathway analysis	Isotopic profiling / Comparative labeling	Isotopic profiling / Annotation / Comparative labeling	Annotation / Relative quantification	Annotation / Relative quantification / isotopic profiling	Isotopic profiling / Comparative labeling / Annotation	Isotopic profiling / Comparative labeling	Quantification / isotopic profiling	Isotopic profiling / Quantification / Comparative labeling / Visualization
Strategies	Targeted/Untargeted	Untargeted	Untargeted	Untargeted	Targeted & Untargeted	Untargeted	Untargeted	Untargeted	Untargeted
Data acquisition	Sample requirements	2 biological equivalent samples Native & Labeled	2 biological equivalent samples Native & Labeled	mixtures of native and uniformly labeled tracer in a single sample	labeled biological samples + mixtures of authentic standards	Uniformly labeled tracer, Time Course Labeling Experiments / a mixture (50/50) of natural labeled and uniformly 13C labeled for identification	Uniformly or partially labeled tracer	mixture of unlabeled and labeled samples	unlabeled and labeled biological samples
	Type of isotopic tracer	^{13}C, ^{15}N, ^{2}H, ^{18}O	^{13}C, (^{15}N, ^{33}S, ^{18}O supposed)	^{13}C, ^{15}N, ^{34}S	^{13}C	^{13}C	^{13}C	^{13}C, ^{15}N, ^{33}S, ^{18}O	^{13}C, 15N
	Analytical platform	LC/MS	LC/MS	LC/HRMS	LC/MS	LC/HRMS	LC & GC/HRMS	GC/MS (LC)	GC/MS
	MSmode	MS	MS	MS, MS/MS	MS	MS	MS	MS	MS
	Data file format	".mzXML"	".mzXML"	".mzXML; .mzML"	".mzXML; .mzML ; .mzData"	".mzXML; .mzML"	".mzXML; .mzML"	".netCDF"	".netCDF"
Software	Dynamic labeling experiments	Not considered	Not considered	Not considered	option	required	option or required ?	Not considered	Not considered
	Dependancies	R	R	python, R	R	python	R	C++, Qt4	C++, Qt5, NTFD, MetaboliteDetector
	Peakpicking	XCMS	XCMS	3 modules : AllExtract, MetExtract,FragExtract	XCMS	OpenMS	XCMS	NFTD algorithms	MetaboliteDetector / NTFD library and GraphViz
	Isotopic grouping	X13CMS functions	geoRge functions		mzMatchIso	emzed (isotope_regrouper)	algorithms in HiResTec package	NFTD algorithms	
	Pre-processing (filtration)	No	No	No	Yes	Yes	Yes	No	No
	Identification	No	HMDB, METLIN, personal database	Number of labeled atoms / heteroatoms / Correlation / Fragmentation	HMDB, KEGG or common metabolic transformations	databases (eg: KEGG)	No (export for further annotation)	reference libraries	reference libraries
	Naturally abundance correction	No	No	No	No	Yes (if mixture of natural labeled and uniformly ^{13}C labeled)	No (export for correction after annotation)	Yes	Yes
Output	Files	Text & Plot	Text	Text & Plot	Text & Plot	Text & Plot	Text & Plot	Text	Text & Plot
	Isotopic data	MID, Enrichment	MID, Enrichment	MID	MID	MID	MID	MID	MID, visualization
	Flux calculation	No	No	No	No	Input for ^{13}C-MFA	export of MIDs for flux modelling software	relative fluxes	No
REFS		[Huang, X. *et al,* 2014]	[Capellades, J. *et al,* 2016]	[Bueschl, C. *et al,* 2017]	[Chokkathukalam, A. *et al, 2013]*	[Kiefer, P. *et al, 2015]*	[Hoffmann, F. et al, 2018]	[Hiller, K., et al, 2010]	[Weindl D, et al, 2016]

** Label comparison between two biological condition (called relative quantification in some software)*