MODÉLISATION MÉTABOLIQUE À L'ÉCHELLE DU GÉNOME DE LA BACTÉRIE QUASI-MINIMALE *MESOPLASMA FLORUM*

par

Jean-Christophe Lachance

Thèse présentée au Département de biologie en vue
de l'obtention du grade de docteur ès sciences (Ph.D.)
FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, janvier 2021

Le 26 janvier 2021

Le jury a accepté la thèse de Monsieur Jean-Christophe Lachance dans sa version finale.

Membres du jury

Professeur Sébastien Rodrigue
Directeur de recherche
Département de Biologie
Université de Sherbrooke

Professeur Pierre-Étienne Jacques
Codirecteur de recherche
Département de Biologie
Université de Sherbrooke

Professeure Ines Thiele
Évaluatrice externe
École de médecine
National University of Ireland, Galway

Chargé de cours Benoît Leblanc
Évaluateur interne
Département de Biologie
Université de Sherbrooke

Professeur Vincent Burrus
Président-rapporteur
Département de biologie
Université de Sherbrooke

REMERCIEMENTS

Je tiens d'abord à remercier chaleureusement mon directeur de recherche, Pr. Sébastien Rodrigue pour le support continu qui dépasse de loin les conversations scientifiques et les processus administratifs. Il a su être présent à chaque moment décisif pour me conseiller dans les choix cruciaux qui ont dicté l'évolution de mon parcours. La confiance et l'ouverture aux idées caractérise l'enseignement de Sébastien et j'espère que cet exemple me suivra pour le reste de ma carrière. Je souhaite également remercier mon co-directeur, le Pr. Pierre-Étienne Jacques, qui a toujours mis les bouchées doubles pour s'assurer que nous produisions un travail de qualité. Je retiens de son enseignement son souci du détail, son goût du travail bien fait et sa rigueur scientifique exemplaire.

Je dois un merci particulier au Pr. Bernhard O. Palsson de UCSD pour l'accueil en tant que visiteur pour une durée prolongée dans son laboratoire. La confiance qu'il a placée dans mes capacités s'est traduite en habiletés et en un meilleur esprit scientifique. Un merci spécial va à mes collègues avec lesquels j'ai échangé nombreuses conversations qui m'ont fait me questionner et grandir tant d'un point de vue scientifique qu'en tant qu'individu. Les gens avec qui j'ai partagé un café autant à San Diego qu'à Sherbrooke pourront donc se reconnaître. Je tiens à nommer spécifiquement Anand, Dominick, Patric, Laurence, David, Yara, Colton, Erol, CJ et Eddy. Enfin, l'importance d'un entourage solide et aimant est essentiel pour la réussite. En ce sens, je remercie mon père de m'avoir inculqué la curiosité scientifique en bas âge et pour son soutien constant à travers ce long parcours académique. Comme le disait Einstein, l'imagination est plus importante que le savoir, et pour moi cette créativité provient de ma mère que je remercie du fond du cœur. L'entourage d'amis en or est essentiel pour surmonter l'adversité que représente la complétion d'études graduées, ainsi je remercie fraternellement ces amis exceptionnels que sont Alexandre, Louis-Charles, Simon, Charles et Patrick. À nos prochaines aventures!

SOMMAIRE

Des avancées significatives au niveau de la synthèse et de l'assemblage de fragments d'acide désoxyribonucléique (ADN), le support physique des fonctions cellulaires encodées dans une cellule vivante, permettent maintenant la construction de génomes entiers. Ce progrès permet d'imaginer que la conception d'organismes synthétiques deviendra routinière au cours des prochaines années. Cette capacité promet de transformer radicalement le domaine de la biologie en formant une nouvelle discipline d'ingénierie biologique. Parmi les retombées anticipées, on note le remplacement de synthèses chimiques par des procédés biologiques renouvelables tels que la production de biocarburants, la synthèse de médicaments microbiens, ou des approches alternatives pour le traitement des maladies.

Dans ce contexte, il devient particulièrement important d'arriver à prédire correctement le phénotype résultant des génomes qui seront générés. Pour y arriver, il convient de réduire la complexité biologique en travaillant d'abord avec les cellules les plus simples possibles. Ce type d'organisme ayant subi un processus de réduction de génome et dont la majorité des gènes sont essentiels afin de survivre en conditions définies se nomme une cellule minimale. Le groupe phylogénétique des mollicutes, bactéries dépourvues de paroi cellulaire, contient les espèces vivant avec les plus petits génomes connus à ce jour. Membre de ce groupe, le pathogène humain *Mycoplasma genitalium* possède le plus petit génome capable de croissance autonome (560kbp codant pour 482 protéines. Cependant, sa pathogénicité et sa vitesse de croissance réduite (~24h) limitent l'applicabilité de *M. genitalium* en biologie synthétique.

Pour remédier à ce problème, notre laboratoire a choisi de travailler avec *Mesoplasma florum* dont le temps de doublement est très rapide (~32 min) et qui ne cause pas de maladies chez l'humain. Les travaux effectués chez *M. florum* permettent maintenant le clonage et la transplantation de son génome et des travaux récents ont permis de caractériser les propriétés

physico-chimiques de sa cellule ainsi que plusieurs paramètres biologiques. Afin de permettre la conception de génomes synthétiques basés sur *M. florum*, il convient d'intégrer un maximum de connaissances dans un cadre informatique structuré capable de générer des prédictions phénotypiques. Un modèle métabolique à l'échelle du génome (GEM) reposant sur la méthode d'analyse des flux à l'équilibre (FBA) représente un format particulièrement intéressant pour initier ces travaux de biologie des systèmes.

La qualité des prédictions générées par ce type de modèle est dépendante de la précision de l'objectif à atteindre. Pour simuler la croissance, les GEMs doivent satisfaire un objectif nommé "fonction objective de biomasse" (BOF) qui contient l'ensemble des métabolites nécessaires à la production d'une nouvelle cellule avec des coefficients stœchiométriques représentatifs de l'abondance de ces composantes dans la cellule. Pendant mon parcours de doctorat, j'ai développé le logiciel BOFdat qui permet la définition d'une BOF représentative de la composition cellulaire spécifique à une espèce avec les données expérimentales associées. Les deux premières des trois étapes de BOFdat déterminent les coefficients stoechiométriques de molécules connues pour faire partie de la composition cellulaire telles que les macromolécules principales (étape 1, ADN, ARN et protéines) et les coenzymes essentiels (étape 2). L'étape 3 de BOFdat propose une méthode non-biaisée pour déterminer les métabolites susceptibles d'améliorer la prédiction d'essentialité des gènes formulée par le modèle. Pour ce faire, un algorithme génétique maximise la composition de la biomasse en fonction des données d'essentialité expérimentales à l'échelle du génome. BOFdat a été validé en reconstruisant la BOF du modèle *i*ML1515 de la bactérie modèle *Escherichia coli*. L'utilisation de BOFdat a permis de récapituler le taux de croissance prédit avec la BOF originale tout en améliorant la qualité des prédictions d'essentialité de gènes de *i*ML1515. BOFdat est disponible en libre accès pour quiconque désire construire une BOF pour un modèle métabolique.

Ensuite, un GEM nommé *i*JL208 a été produit et contient 208 des 676 protéines représentant l'ensemble du métabolisme de *M. florum.* La qualité de l'annotation du génome a d'abord été

évaluée en intégrant l'information obtenue par trois approches bio-informatiques, révélant que la majorité des protéines (418/676) ont une qualité suffisante pour être incorporées dans le modèle. Ensuite, les réactions ont été identifiées et rigoureusement incorporées une à la fois afin de construire le réseau métabolique de cette bactérie quasi-minimale. L'étude de la carte métabolique reconstruite révèle une dépendance prononcée pour l'import de composantes à partir du milieu de culture ainsi que l'importance des mécanismes de recyclage des métabolites. Pour sa production d'énergie, *M. florum* est entièrement dépendante de la glycolyse et ne possède pas la machinerie nécessaire à la respiration cellulaire. L'élaboration d'un milieu de culture semi-défini a réduit la présence de sucres contaminants dans le milieu de culture initial et ainsi de distinguer la croissance avec ou sans supplémentation de sucrose. Cette avancée importante a permis de mesurer les taux d'assimilation de sucrose et de production des déchets métaboliques lactate et acétate. Ces paramètres ont été utilisés afin de contraindre le modèle et de mieux comprendre la sensibilité du modèle à une variété de paramètres. Aussi, la croissance de *M. florum* a pu être validée expérimentalement avec différents sucres. L'information contextuelle obtenue, combinée à une analyse de structures tridimensionnelles de protéines clés, a permis de suggérer des hypothèses crédibles supportant l'assimilation de ces sucres par *M. florum*.

Finalement, *i*JL208 a été utilisé afin de formuler une prédiction de génome minimal pour *M. florum* en simulant itérativement de larges délétions dans son génome. Combiner l'intégration de données expérimentales avec les prédictions du modèle constitue une voie d'avenir pour la conception de génomes synthétiques qui rejoint les capacités techniques d'assemblage de chromosomes en biologie synthétique. Globalement, les projets réalisés au cours de mon doctorat contribuent à l'avancement de la biologie des systèmes chez *M. florum* dans le but de prédire efficacement les phénotypes de la souche naturelle et de variants synthétiques qui pourront être produits au cours des prochaines années.

Mots clés : biologie des systèmes, biologie synthétique, bio-informatique, modélisation cellulaire, cellules minimales, *Mesoplasma florum*

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS ET DES SIGLES

ADN	Acide désoxyribonucléique
AFM	Analyse des flux métaboliques
ARN	Acide ribonucléique
ATP	Adenosine tri-phosphate
ATPM	*ATP maintenance* (Réaction de maintenance de l'ATP)
BLASTp	Outil de recherche d'alignement local de base pour protéines
BOF	*Biomass objective function* (Fonction objective de biomasse)
BOSS	*Biological Objective Solution Search* (
COBRA	*Constraint-based Reconstruction and Analysis* (Reconstruction et analyse basée sur les contraintes)
CS	Coefficients stoechiométriques
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise* (Regroupement spatial basé sur la densité des applications avec bruit)
EC	*Enzyme Commission* (Numéro de commission d'un enzyme)
FBA	*Flux Balance Analysis* (Analyse des flux à l'équilibre)
FBAwMC	*FBA with Molecular* Constraints (FBA avec encombrement moléculaire)
GAM	*Growth-Associated* Maintenance (Coûts de maintenance associé à la croissance)
GECKO	*Genome-scale model to account for Enzyme Constraints, using Kinetics and Omics* (GEM avec contraintes enzymatiques en utilisant des données cinétiques et omiques)
GEM	*Genome-scale Model* (Modèle métabolique à l'échelle du génome)
GPR	*Gene reaction rule* (Association gène-réaction)
HGP	*Human Genome Project* (Projet de séquençage du génome humain)

HPLC	*High-performance Liquid Chromatography* (Chromatographie en phase liquide à haute performance)
JCVI	*John Craig Venter Institute* (Institut John Craig Venter)
kb	kilobase
LC-MS	*Liquid Chromatography Mass Spectrometry* (Spectrométrie de masse couplé à la séparation par chromatographie en phase liquide)
MCC	*Matthews Correlation Coefficient* (Coefficient de corrélation de Matthews)
ME	Modèle métabolique et expression
MOMA	*Minimization Of Metabolic Adjustments* (Minimization des ajustements métaboliques)
NGAM	*Non-Growth Associated Maintenance* (Coûts de maintenance non-associé à la croissance)
NGS	*Next-Generation Sequencing* (Séquençage de nouvelle génération)
NOD	*Non-Orthologous gene Displacement*Déplacement de gènes non orthologues
ODE	*Ordinary Differential Equations* (Équations différentielles ordinaires)
oriC	Origine de réplication du chromosome
pb	Paire(s) de base(s)
pFBA	*Parsimonious Flux Balance Analysis* (Analyse des flux à l'équilibre parsimonieux)
RBC	*Reb Blood Cell* (Globule rouge)
SL	*Synthetic lethality* (Létalité synthétique)
TFA	*Thermodynamic Flux Analysis* (Analyse des flux métaboliques à l'équilibre basé sur la thermodynamique)
uFBA	*Unsteady-state Flux Balance Analysis* (Analyse des flux dans un état non équilibré)
WGS	*Whole-genome sequence* (Séquence de génome entier)

CHAPITRE 1
INTRODUCTION

1.1 CONCEVOIR LA VIE

"Scientists investigate that which already is; Engineers create that which has never been"
- Albert Einstein

Au cours des 200 dernières années, les biologistes ont fourni un large éventail de connaissances sur les fondements de la vie sur Terre. Les théories et les dogmes actuels ont émergé d'un labyrinthe de suppositions et d'hypothèses à travers une succession de découvertes clés et d'avancées progressives. Aujourd'hui, peu de fonctions moléculaires nécessaires au maintien de la vie restent inconnues. Bien que la biologie ait considérablement mûri en tant que discipline scientifique, je discuterai ici de la manière dont la caractérisation exhaustive des organismes, associée à des cadres de modélisation appropriés, devrait conduire à une nouvelle ère, dans laquelle l'ingénierie cellulaire est appelée à devenir une discipline indépendante. En raison de leur moindre complexité, les microorganismes - en particulier les bactéries minimales - devraient jouer un rôle très important dans cette entreprise.

Le contexte historique et les étapes clés qui ont conduit à la naissance du génie biologique sont abordés ici. Ce rappel historique devrait mettre en évidence l'importance des modèles cellulaires minimaux tout en offrant au lecteur une perspective sur l'ensemble du domaine de la biologie et est divisé ici en trois étapes : la biologie classique, la biologie moléculaire, la génomique et enfin la biologie synthétique (Figure 1.1). À noter, le texte des sections 1.1 à 1.6 est tiré de Lachance et al., 2019a.

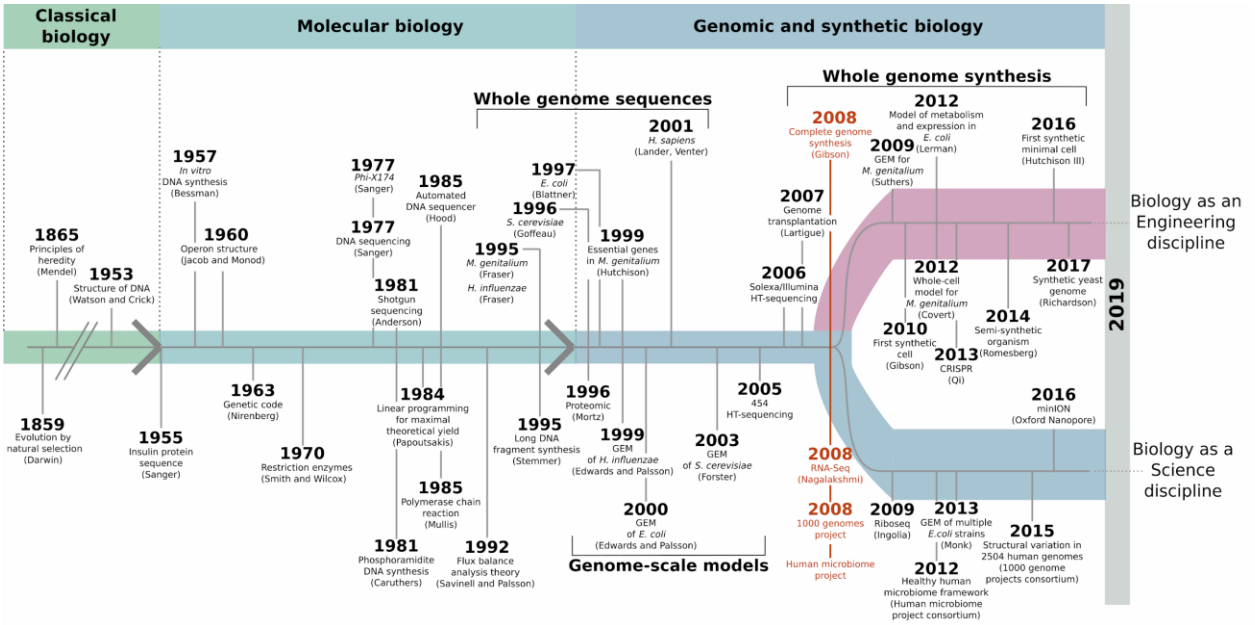

Figure 1.1 Biologie synthétique et cellules minimales : une perspective historique. L'élucidation de la structure de l'ADN a marqué le début de l'ère de la biologie moléculaire, en rendant possible l'étude des mécanismes moléculaires qui sous-tendent les phénotypes observables. Par la suite, le développement des méthodes de séquençage d'ADN a conduit au séquençage de génomes entiers à la fin des années 1990 et marque le début de l'ère génomique. Suivant l'établissement de méthodes de modélisation mathématique des cellules au cours des décennies 1980 et 1990, des modèles métaboliques à l'échelle du génome ont pu être reconstruits qui utilisent les séquences de génome entiers. La création du premier génome artificiel à partir de la séquence de M. genitalium marque le début de la biologie synthétique (rouge) (Reproduit de Lachance et al., 2019a).

1.1.1 La biologie classique

En 1859, Darwin publia son ouvrage intitulé "*On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life*". Moins d'une décennie plus tard, en 1865, Mendel proposa les mécanismes de l'hérédité. Les deux théories utilisaient des phénotypes observables au niveau de l'organisme pour déduire les mécanismes potentiels qui régissent leur évolution. Alors que les travaux de Darwin expliquaient les forces qui sous-tendent l'émergence des phénotypes et de la spéciation, ceux de Mendel se

concentraient sur une explication mécanistique des principes de l'hérédité. Bien qu'il ne soit pas spécifiquement décrit par Mendel, ses conclusions ont donné naissance au concept de gène. La compréhension de la base chimique du gène et de l'hérédité est alors devenue la principale entreprise de cette première ère de la biologie, définie ici comme l'ère de la biologie classique (Figure 1.1). Cet objectif est resté l'un des grands défis de la biologie jusqu'à ce que, en 1953, Watson et Crick publient la structure de la double hélice d'acide désoxyribonucléique (ADN) (Watson et al., 1953). Cette découverte historique a permis aux scientifiques de poser des questions plus complexes sur les fonctions moléculaires qui soutiennent la vie, marquant ainsi le début de l'ère de la biologie moléculaire (Waddington, 1961).

1.1.2 La biologie moléculaire

L'élucidation de la structure de l'ADN a rendu le concept de gène tangible, accélérant considérablement le rythme des découvertes. Parmi les découvertes emblématiques et révolutionnaires de l'ère de la biologie moléculaire, citons d'abord le déchiffrage du code génétique (Holley, 1965; Nirenberg et al., 1963, Nirenberg et al., 1965) et la définition de l'opéron par Jacob et Monod, qui, pour la première fois, a révélé les mécanismes moléculaires sous-jacents à l'expression des gènes (Jacob, et al., 1960). La découverte ultérieure d'une enzyme de restriction (enzyme capable de couper l'ADN selon une séquence spécifique) chez *Haemophilus influenzae* (Smith and Wilcox, 1970) et son application pour couper le génome du virus humain SV40 (Danna and Nathans, 1971) ont marqué le début des manipulations de l'ADN (Roberts, 2005). La réaffectation d'une enzyme de restriction a fourni le premier outil de génie génétique, permettant aux biologistes de commencer à déchiffrer les mécanismes moléculaires qui sous-tendent les phénotypes cellulaires.

Bien qu'il soit utile de cliver l'ADN à des endroits spécifiques, un défi important à relever consistait à décoder la séquence des gènes. Le code génétique ayant été déterminé en 1965 (Holley, 1965; Nirenberg et al., 1963, Nirenberg et al., 1965), la capacité de séquencer l'ADN

permettrait d'obtenir la séquence d'acides aminés des protéines, qui à son tour assure la médiation de sa fonction. En 1977, Frederick Sanger a publié une méthode de séquençage de l'ADN par incorporation aléatoire de nucléotides radiomarqués dépourvus du groupe 3'-OH nécessaire à l'élongation des chaînes d'acides nucléiques (Sanger et al., 1977a). Cette méthode a permis le séquençage du génome complet du phage ϕX174 (5 375 pb) (Sanger et al., 1977b). Alors que la méthode basée sur l'intégration de didésoxyribonucléotides radiomarqués de Sanger était en mesure de générer des séquences allant de 15 à 200 nucléotides, une mise à l'échelle massive était toutefois nécessaire pour permettre des efforts de séquençage plus ambitieux.

Le séquenceur d'ADN automatisé (Smith et al., 1986) et l'avènement du séquençage aléatoire ("shotgun") (Anderson, 1981) ont considérablement augmenté la capacité de séquençage de l'ADN, permettant de séquencer des génomes entiers plus longs (Heather and Chain, 2016). Suivant la rencontre de Santa Cruz en 1985 (Sinsheimer, 1989), le projet du génome humain (HGP) a été lancé et s'est achevé en 2001 (Lander et al., 2001; Venter et al., 2001). Grâce aux technologies développées pour le HGP, des projets de séquençage de génomes complets (WGS) à plus petite échelle ont été achevés avant le nouveau millénaire (Figure 1.1). Dans une référence historique à la première enzyme de restriction de type II isolée, la première séquence complète d'un organisme vivant, *Haemophilus influenzae*, a été rapportée en 1995 (Fleischmann et al., 1995). Peu de temps après, le génome entier de *Mycoplasma genitalium*, le plus petit organisme vivant capable de survivre de manière autonome, a été publié (Fraser et al., 1995). Les organismes modèles plus complexes *Saccharomyces cerevisiae* et *Escherichia coli* ont suivi en 1996 et 1997, respectivement (Blattner et al., 1997; Goffeau et al., 1996).

1.1.3 Génomique

Le début de l'ère génomique est arbitrairement défini ici avec l'achèvement de la première séquence génomique d'un organisme vivant autonome (Fleischmann et al., 1995) (Figure 1.1). Le nombre de génomes générés à la suite de celle de *H. influenzae* a régulièrement augmenté, pour finalement inclure les 3,2 milliards de paires de bases (pb) du génome humain haploïde (Lander et al., 2001; Venter et al., 2001). L'amélioration des outils informatiques et d'automatisation ont permis d'accroître encore la capacité du séquençage aléatoire de type Sanger. Néanmoins, l'avènement des technologies de séquençage de nouvelle génération (NGS) développées par des entreprises privées à l'issue du projet de génome humain (HGP) a représenté une avancée majeure. Si le paradigme du séquençage par synthèse a été préservé entre le séquençage Sanger et les méthodes NGS, la capacité de paralléliser le séquençage au sein d'une même réaction a massivement augmenté le débit (Heather and Chain, 2016).

La NGS a permis l'élaboration de nouvelles initiatives telles que le projet de 1000 génomes (Spencer, 2008) et le projet de séquençage du microbiome humain (McGuire et al., 2008). Malgré leur échelle beaucoup plus grande, ces deux initiatives ont atteint leurs objectifs principaux en 4 ans (1000 Genomes Project Consortium et al., 2012; Human Microbiome Project Consortium, 2012), soit seulement le tiers du temps requis pour compléter le projet de séquençage du génome humain. Cet accomplissement démontre bien la puissance des technologies NGS. L'accessibilité du séquençage à haut-débit contribue maintenant à une expansion sans précédent des connaissances destinée à se poursuivre. Récemment, le développement par la compagnie Oxford Nanopore d'un séquenceur portable, pouvant être utilisé en laboratoire et capable de produire des séquences en temps réel (minION) (Lu et al., 2016) a permis d'élargir encore les applications des NGS pour la découverte fondamentale.

Obtenir la séquence complète d'un grand nombre de génomes d'espèces différentes est cruciale pour arriver à comprendre les relations phylogénétiques qui les unissent et l'ensemble des fonctions qu'elles encodent. Cependant, l'information génétique codée dans l'ADN d'une

cellule est essentiellement statique et ne révèle pas la nature dynamique des phénotypes moléculaires, une réalité devenue évidente peu après l'achèvement du séquençage du génome humain. Il s'est alors avéré que le nombre de gènes prédits chez l'humain avait été grossièrement surestimé (Brower, 2001). Heureusement, les efforts dédiés à l'interrogation à haut débit d'autres composantes cellulaires importantes ont commencé tôt avec le développement d'approches non ciblées pour le séquençage des protéines (Mørtz et al., 1996). Plus d'une décennie plus tard, l'élaboration d'un protocole de séquençage d'ARN à haut débit utilisant les technologies NGS a révélé le profil transcriptomique complet de la levure (Nagalakshmi et al., 2008). Dès lors, les trois macromolécules principales du dogme central de la biologie (Crick, 1970) ont pu être séquencées à l'échelle du génome de manière non ciblée.

Les autres composants de la cellule tels que les lipides, glycans et métabolites ne sont pas aussi ubiquitaires que l'ADN, l'ARN et les protéines, ce qui rend leur identification par des méthodes non ciblées à l'échelle de l'organisme plus complexe. L'identification de tous les composants solubles dans l'eau est appelée métabolomique, alors que le contenu hydrophobe est généralement appelé lipidomique (Riekeberg and Powers, 2017). La chromatographie liquide suivie de la spectrométrie de masse (LC-MS) permet une détermination à la fois métabolomique et lipidomique (Riekeberg and Powers, 2017; Yang and Han, 2016). Pour cette identification, la méthode d'extraction du matériel cellulaire varie en fonction de la polarité des composés. Ces méthodes, ainsi que d'autres (Ingolia et al., 2009; Lahner et al., 2003; Zamboni et al., 2009), permettent de caractériser un état dynamique de la cellule qui peut être exploité en biologie des systèmes (Haas et al., 2017).

1.1.4 Biologie synthétique

Le terme "biologie synthétique" est étroitement associé à l'application des principes d'ingénierie aux systèmes biologiques. La synthèse d'ADN a permis la génération et l'assemblage de fragments d'ADN complètement synthétiques. À leur tour, ces capacités ont

permis de créer des entités nouvelles, définissant ainsi la biologie synthétique comme un domaine s'apparentant à l'ingénierie (Andrianantoandro et al., 2006; Heinemann and Panke, 2006; Hughes and Ellington, 2017).

La première tentative de production d'ADN synthétique a eu lieu peu après l'élucidation de sa structure. En 1957, Bessman et ses collègues ont utilisé l'ADN polymérase de *E. coli* pour produire des fragments d'ADN. Ils ont constaté que la présence d'ADN polymérisé est nécessaire à la réaction. Ce concept a ensuite été réutilisé par Sanger pour le séquençage de l'ADN (Sanger et al., 1977a) et plus tard pour la fameuse réaction en chaîne de la polymérase (PCR) (Saiki et al., 1985). Les amorces d'oligonucléotides utilisées pour le développement de la PCR ont été produites par la méthode des phosphoramidites (Beaucage and Caruthers, 1981; Matteucci and Caruthers, 1981). Bien que cette chimie soit encore utilisée actuellement dans la plupart des plateformes modernes de synthèse d'ADN (LeProust, 2016), elle est toutefois limitée par rapport à la longueur des oligonucléotides qui peuvent être produits sans accumuler de mutations indésirables. Ce problème a été contourné par Stemmer en 1995, qui a été le premier à signaler une technique permettant de générer un long fragment d'ADN synthétique (>1000 pb) par assemblage d'oligonucléotides (Stemmer et al., 1995). Bien que le coût de la synthèse de l'ADN n'ait pas beaucoup diminué au cours des dix dernières années (Hughes and Ellington, 2017), les récents progrès vers des stratégies de synthèse d'ADN à haut débit utilisant des puces à ADN pourraient bientôt résoudre ce problème (LeProust, 2016), et promettent de faire de la synthèse de grands fragments d'ADN une solution abordable pour les expériences de routine en biologie moléculaire ou la conception de souches industrielles (Bassalo et al., 2016; Hughes and Ellington, 2017).

Un objectif important de la synthèse de l'ADN est la conception et l'assemblage de génomes entiers. Pour atteindre cet objectif, il était nécessaire de mettre au point des méthodes robustes pour assembler des fragments d'ADN en séquences plus grandes. Cet objectif a été atteint en 2008 lorsqu'une équipe du *John Craig Venter Institute* (JCVI) a réalisé la synthèse et l'assemblage complets du génome de *Mycoplasma genitalium* (Gibson et al., 2008). Cette

réalisation a été rendue possible par une stratégie hiérarchique reposant sur la recombinaison *in vitro* de cassettes d'ADN (Gibson et al., 2009). Cette méthode d'assemblage d'oligonucléotides se chevauchant permet de créer des fragments d'ADN plus grands et s'est ensuite révélée encore plus efficace *in vivo* en utilisant la levure (Gibson, 2009). Le développement de méthodes de synthèse et d'assemblage du génome entier, ainsi que celui de la transplantation du génome entier (Lartigue et al., 2007), a permis la création de la première cellule vivant avec un génome entièrement synthétique (Gibson et al., 2010).

Ces dernières années, des efforts spectaculaires ont été accomplis en matière de biologie synthétique et auront sans doute un impact sur l'avenir de cette discipline. En 2014, Romesberg et ses collègues ont créé une bactérie fonctionnant avec un ADN modifié contenant 6 bases différentes (Malyshev et al., 2014), offrant ainsi une combinaison d'appariement de bases supplémentaires. Aucun organisme vivant connu ne contient ces nucléobases synthétiques, cette réalisation a donc donné naissance à une nouvelle forme de vie sur Terre. Aussi, suivant le chemin tracé par la réalisation du premier organisme vivant autonome contenant un génome synthétique, l'équipe du JCVI a conçu et assemblé une cellule dont le contenu génétique est fortement réduit, permettant ainsi d'obtenir la première approximation fonctionnelle d'une cellule minimale (Hutchison et al., 2016). Enfin, le projet Sc-2.0 a été lancé et, en 2017, un consortium international a signalé la synthèse complète *de novo* de 5 chromosomes entiers de la levure *S. cerevisiae* (Richardson et al., 2017).

La combinaison du développement des multiples technologies mentionnées précédemment (technologies NGS, multiples méthodes omiques pour la caractérisation dynamique cellule, méthodes ciblées d'édition du génome (Qi et al., 2013) et méthodes de synthèse et d'assemblage d'ADN à haut débit), la biologie synthétique possède les outils nécessaires pour la production d'organismes partiellement ou complètement synthétiques. Ces nouvelles formes de vie révolutionneront de nombreux domaines de la recherche industrielle tels que la synthèse de médicaments microbiens, la production de biocarburants ou des approches alternatives pour le traitement des maladies (Smolke et al., 2018).

1.1.5 Le concept de cellule minimale

"The hydrogen atom of biology"
-Harold J. Morowitz

L'idée d'une cellule minimale a été abordée pour la première fois par le biophysicien Harold J. Morowitz lors d'une conférence en 1984 (Morowitz, 1984). Le raisonnement qu'il introduisit stipulait qu'un organisme vivant de manière autonome possède une limite inférieure quant au nombre d'atomes qui le compose. Sous ce nombre, les fonctions nécessaires au maintien de la vie ne seraient pas remplies. Cette déduction logique ressemble un peu à celle de Schrödinger dans son célèbre livre "Qu'est-ce que la vie?" (Schrodinger, 1967). Dans cet ouvrage paru avant l'élucidation de la structure d'ADN, le célèbre physicien questionnait le support matériel du gène et appliquait les contraintes connues imposées par la physique quantique pour prédire correctement qu'il s'agirait d'une molécule pouvant former un cristal. De son côté, Morowitz a proposé que les mollicutes, un groupe phylogénétique de bactéries dépourvues de paroi cellulaire, seraient les candidats se rapprochant le plus d'un nombre minimal d'atomes et seraient donc particulièrement adaptés pour générer ce qu'il définit alors comme une "cellule minimale". Ce choix d'organisme était basé sur l'observation de la taille des cellules, assumant qu'une cellule minimale, tout comme l'atome d'hydrogène en physique, serait parmi les plus petites cellules existantes et permettrait donc d'obtenir une compréhension fondamentale applicable à d'autres systèmes biologiques plus complexes. La prédiction était exacte puisque le mollicute *Mycoplasma genitalium*, deuxième organisme vivant de manière autonome à être entièrement séquencé (Figure 1.1) (Fraser et al., 1995), possède encore aujourd'hui le plus petit contenu génétique de tous les organismes naturels connus. L'objectif de l'étude des cellules minimales a alors été clairement énoncé : définir les principes de base de la vie (Glass et al., 2017).

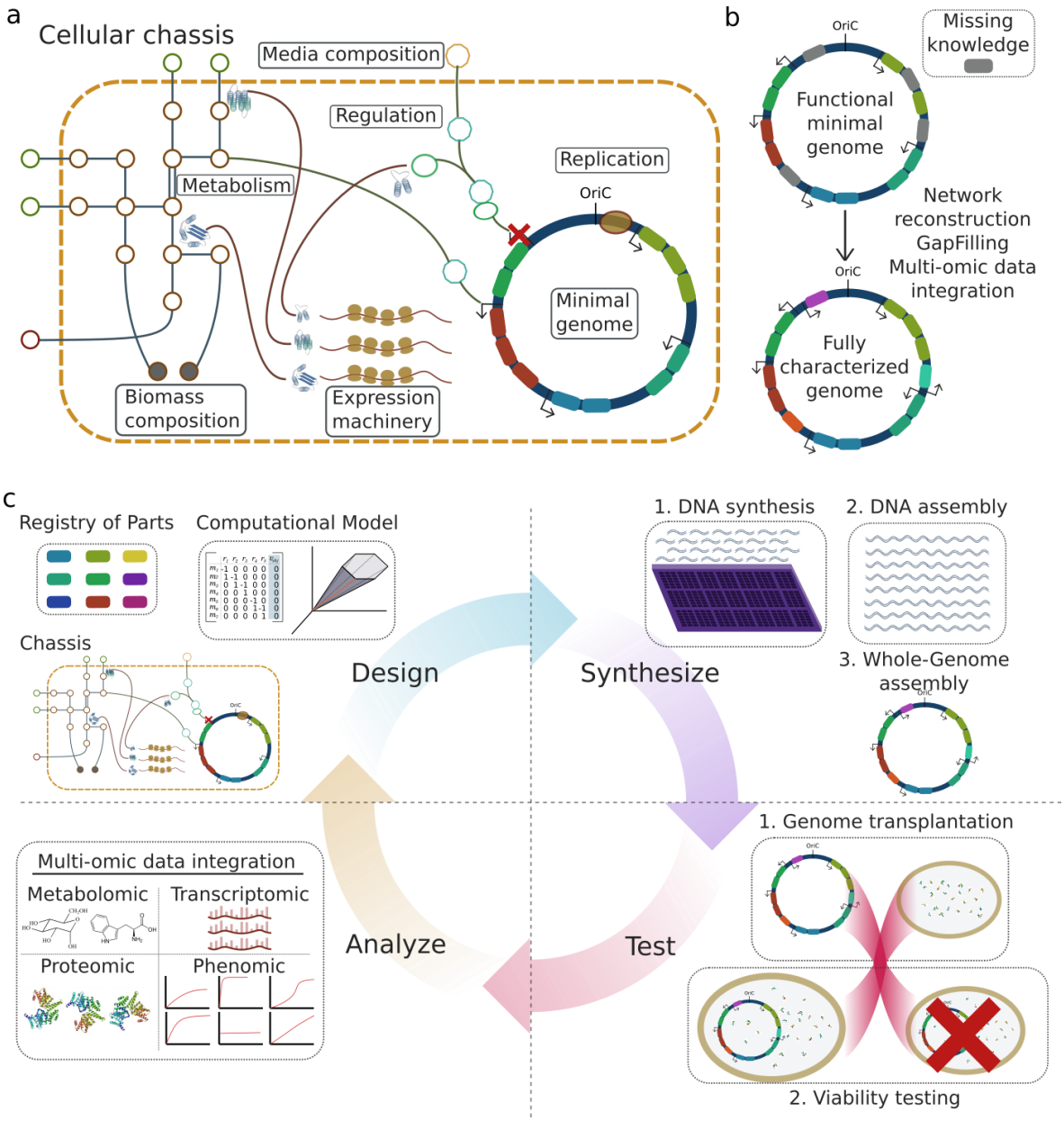

Figure 1.2. Conception de cellules à l'aide d'un modèle informatique. A. Représentation naïve d'un châssis cellulaire dans lequel toutes les fonctions cellulaires obligatoires et leurs interactions sont comprises et caractérisées. B. La génération de modèles informatiques pour les cellules minimales peut accélérer l'identification des connaissances manquantes et faciliter la génération d'hypothèses pour les fonctions cellulaires essentielles non caractérisées. C. Une boucle conception-construction-essai-analyse pour la génération de cellules minimales et leur amélioration vers des souches de production. Des modèles mathématiques sont utilisés pour prédire les génotypes fonctionnels et les technologies actuelles de synthèse d'ADN mentionnées dans le texte sont utilisées pour générer le génome proposé. Le clonage de génomes entiers dans des cellules vivantes permet de tester la viabilité et de multiples

ensembles de données omiques sont utilisés pour caractériser l'organisme synthétique (Reproduit de Lachance et al., 2020).

Dès qu'une seconde séquence de génome entier fût générée, Mushegian et Koonin ont cherché à comparer les deux espèces phylogénétiquement éloignées dans l'espoir de trouver un ensemble de gènes orthologues représentant une approximation fonctionnelle d'un génome minimal (Koonin and Mushegian, 1996). La proposition initiale était que 256 gènes seraient suffisants pour entretenir la vie. Par la suite, cette proposition basée sur la génomique comparative s'est avérée être une estimation relativement basse. En effet, sonder l'essentialité des gènes par insertion aléatoire de transposons dans les bactéries à génome réduit a permis d'estimer que le nombre minimal de gènes suffisant pour soutenir la vie autonome serait compris entre 265 et 350 (Hutchison et al., 1999). Avec le nombre croissant de séquences de génome entier disponibles, la génomique comparative a permis d'approfondir la compréhension du concept d'ensemble minimal de gènes. En comparant l'eucaryote *S. cerevisiae* à sa proposition initiale, Koonin a réalisé que très peu de gènes étaient conservés (seulement 40%) (Koonin, 2000). L'explication suggérée était que le déplacement de gènes non orthologues (NOD) aurait une fréquence plus élevée que prévu initialement (Koonin et al., 1996). La définition du NOD stipule que les gènes ayant des fonctions similaires peuvent évoluer de manière indépendante. Cela a induit un changement de paradigme dans le concept d'ensemble minimal de gènes, où l'identité des gènes eux-mêmes a été reportée à un second niveau; l'activité fonctionnelle qu'ils fournissent devenant plus importante. D'un point de vue technique, l'ensemble minimal de fonctions est en effet plus intéressant que l'ensemble de gènes (Danchin and Fang, 2016). Dans ce contexte, les différents gènes deviennent des parties interchangeables qui permettent d'accomplir une fonction donnée.

Les nombreux progrès de la biologie synthétique réalisés par les scientifiques du JCVI ont permis de concevoir et de synthétiser la première approximation fonctionnelle d'une cellule minimale : JCVI-syn3.0 (Hutchison et al., 2016). Les 473 gènes encodés dans le chromosome de cette cellule sont moins nombreux que ceux de tout autre organisme connu vivant de manière autonome (Glass et al., 2017), mais sont nettement plus nombreux que les ensembles

de gènes minimaux déterminés par calcul et par expérience (Glass et al., 2006; Hutchison et al., 1999; Koonin, 2000; Mushegian and Koonin, 1996). Bien qu'essentielle pour la croissance cellulaire, une fraction importante (149/473, ~30%) de l'ensemble de gènes de JCVI-syn3.0 n'a pas de fonction proposée (Danchin and Fang, 2016; Glass et al., 2017; Hutchison et al., 2016). Danchin et Fang ont examiné ces gènes de manière approfondie à la recherche de mécanismes moléculaires devant être remplis (Danchin and Fang, 2016). Cette étude a réussi à fournir des fonctions potentielles basées sur les besoins connus ou prévus pour 32 de ces 84 gènes génériques et 65 "inconnus inconnus". La validité de ces hypothèses reste à déterminer et l'objectif de compréhension exhaustive d'un organisme vivant soulevé par Morowitz reste donc toujours d'actualité.

Alors que les cellules minimales dont le contenu génique se rapproche d'une limite inférieure absolue sont intéressantes pour informer sur les principes fondamentaux de la vie, les cellules fortement réduites présentent également un grand intérêt du point de vue de l'ingénierie biologique. Choe et ses collègues ont passé en revue certains avantages potentiels des bactéries réduites pour la conception de souches de production (Choe et al., 2016). Comme mentionné, la caractérisation à haut débit des phénotypes cellulaires obtenus par la génération de données omiques, combinée à l'augmentation du débit de la synthèse d'ADN, devraient permettre la construction de génomes conçus *in silico* (Figure 1.2). J'énumére ici certains des avantages de cette approche tel que soulignés par cette équipe.

La première bactérie vivant avec un chromosome synthétique, JCVI-syn1.0 (Gibson et al., 2010), aurait demandé un effort évalué à 40 millions de dollars (Sleator, 2010), un coût prohibitif pour la majorité des laboratoires. Synthétiser un génome contenant moins de paires de bases entraîne évidemment une réduction du coût de synthèse d'ADN. Aussi, une baisse du prix de synthèse par base pourrait réduire le coût total de production de génomes synthétiques. Bien qu'annoncée par des sociétés comme Twist Bioscience, une réduction du prix de la synthèse allant à l'encontre de la loi de Moore n'est pas prévue par d'autres équipes (Smolke et al., 2018). Ainsi, l'impact économique de la génération de plusieurs petits génomes resterait

important. Du point de vue de la conception de génome par biologie des systèmes, un nombre réduit de gènes se traduit par une réduction du nombre d'interactions imprévues qui pourraient affecter les résultats envisagés au moment de la conception. Le développement d'approches systématiques à haut débit à l'ère de la génomique a permis la caractérisation rapide des cellules, mais le résultat des modifications génétiques n'est pas encore entièrement prévisible. La modélisation à l'échelle du génome de cellules minimales pourrait conduire à des prédictions plus fiables des modèles. Par exemple, plusieurs efforts ont déjà été déployés pour réduire la complexité des modèles métaboliques afin de rendre les solutions générées plus faciles à visualiser (Ataman and Hatzimanikatis, 2017). La réduction et la minimisation du génome permettent également de concevoir des stratégies de confinement biologique. Celles-ci comprennent les auxotrophies ou la mort cellulaire programmée, qui seront très bénéfiques lorsque la biologie synthétique sera utilisée pour des applications commerciales. Enfin, pour les organismes plus complexes, la suppression de grandes sections du génome pourrait accélérer la réplication du génome tout en augmentant potentiellement la stabilité génomique par l'élimination des éléments dupliqués.

1.2 MODÉLISATION BASÉE SUR LES CONTRAINTES

Dans la dernière section, la transformation potentielle de la biologie de discipline de sciences pures à une discipline d'ingénierie en lien avec le développement de la biologie synthétique a été examiné. L'avènement de méthodes de caractérisation à haut débit des organismes ainsi que le réductionnisme biologique implique une description mécanistique des processus de maintien de la vie, qui ont conduit à la naissance de la biologie synthétique en tant que domaine. Dans ce contexte, l'idée qu'une cellule minimale constituerait un châssis fonctionnel pour la conception de souches de production et une plate-forme pour la compréhension fondamentale de la biologie (Danchin, 2012) a été revu. Tel que mentionné plus haut, les 149 gènes sans fonction associée dans l'organisme synthétique JCVI-syn3.0 (Hutchison et al., 2016) montrent l'état actuel de la recherche sur la cellule minimale où une caractérisation plus

poussée des fonctions moléculaires est nécessaire pour parvenir à une compréhension complète des fonctions moléculaires primordiales au maintien de la vie. Cette approche de réductionnisme biologique devrait s'inscrire dans un format informatique structuré permettant une analyse intégrative ainsi des simulations phénotypes en tirant parti des méthodes à haut débit pour valider les prédictions. Dans la prochaine section, l'analyse de l'équilibre des flux (FBA) (Orth et al., 2010) sera décrite. Au tournant du millénaire, cette approche mathématique a entre autres permis de générer des modèles à l'échelle du génome à partir des séquences de génome entier (Edwards and Palsson, 1999, 2000). Cette approche de modélisation constitue une base solide sur laquelle des cellules minimales peuvent être conçues *in silico*.

1.2.1 Concept de contraintes dans le métabolisme

Le FBA est né, dans les années 1980, d'une tentative de générer des modèles simples pour la fermentation des matières premières de l'industrie chimique par des bactéries (Papoutsakis, 1984). Un premier modèle proposé par Papoutsakis reposait sur l'hypothèse que le processus de fermentation pouvait être résumé dans une seule équation stoechiométrique où l'équilibre élémentaire serait conservé. Dans ce cas, ladite “équation de fermentation" utilisait la stoechiométrie connue de l’ensemble des réactions impliquées dans la fermentation de l'acide butyrique. D’une manière similaire, la stoechiométrie des réactions biochimiques contenue dans un réseau métabolique a ensuite été utilisée par Majewski et Domach pour tenter d'établir une compréhension théorique de la fermentation d'acétate dans les cellules d'*E. coli* cultivées en conditions aérobiques (Majewski and Domach, 1990). Le modèle présenté pour le phénomène d’excès d'acétate comportait de nombreux éléments clés du FBA. L'hypothèse proposée était qu'un réseau métabolique avec un objectif donné pouvait représenter et expliquer le changement d'état métabolique d'*E. coli* responsable de l'excrétion d'acétate.

Le problème a été résumé comme un problème d'optimisation linéaire sur lequel les contraintes du réseau peuvent s'appliquer. En fixant la production d'ATP comme objectif et en appliquant deux contraintes :

1) limiter la quantité d'équivalents réducteurs qui peuvent être produits par la chaîne de transport des électrons et

2) en supposant qu'une enzyme donnée du cycle de Krebs limite le flux à travers une réaction donnée; les auteurs ont démontré que la programmation linéaire pouvait correctement prédire un état métabolique bactérien.

L'utilisation d'un réseau de flux métabolique optimisé par la programmation linéaire a servi de base au développement du formalisme mathématique pour le FBA (Savinell and Palsson, 1992a, 1992b). Le concept a été étendu par la définition d'une matrice stoechiométrique (**S**). Dans cette matrice, chaque colonne représente une réaction du réseau métabolique et chaque ligne un métabolite différent (Figure 1.3). La formulation mathématique de la concentration des métabolites dans le temps utilisant la matrice **S** devient alors :

$$\frac{dX}{dt} = S \cdot v \qquad \text{(équation 1.1)}$$

Où **X** est le vecteur des métabolites et **v** est le vecteur de flux. Le FBA suppose que le réseau métabolique opère à un état d'équilibre. Dans ce cas, la concentration des métabolites au fil du temps devrait être en équilibre et les entrées sont égales aux sorties, de sorte que :

$$0 = S \cdot v \qquad \text{(équation 1.2)}$$

Le FBA présente l'avantage de ne nécessiter que la stoechiométrie des réactions pour fonctionner. Les détails de la thermodynamique pour chaque réaction ne sont pas nécessaires. Néanmoins, la direction des réactions peut être obtenue à partir de la thermodynamique, ajoutant ainsi un autre ensemble de contraintes sur le système. Un fonction objective (**Z**)

physiologiquement significative peut être défini afin de simuler le phénotype métabolique souhaité :

$$\begin{aligned} &maximize\ Z, \\ &0 = S \cdot v \\ &a_i < v_i < b_i \end{aligned} \qquad \text{(équation 1.3)}$$

Cette formulation mathématique, dans laquelle $\boldsymbol{a_i}$ et $\boldsymbol{b_i}$ représentent les bornes de flux inférieures et supérieures d'une réaction $\boldsymbol{v_i}$, peut être résolue par programmation linéaire et permet de trouver la solution optimale d'un réseau métabolique à l'état d'équilibre. Maintenant, la manière dont cette formulation permet de générer des modèles à l'échelle du génome ainsi que le rôle de la fonction objective et ses adaptations possibles pour représenter des états physiologiques spécifiques seront examinés.

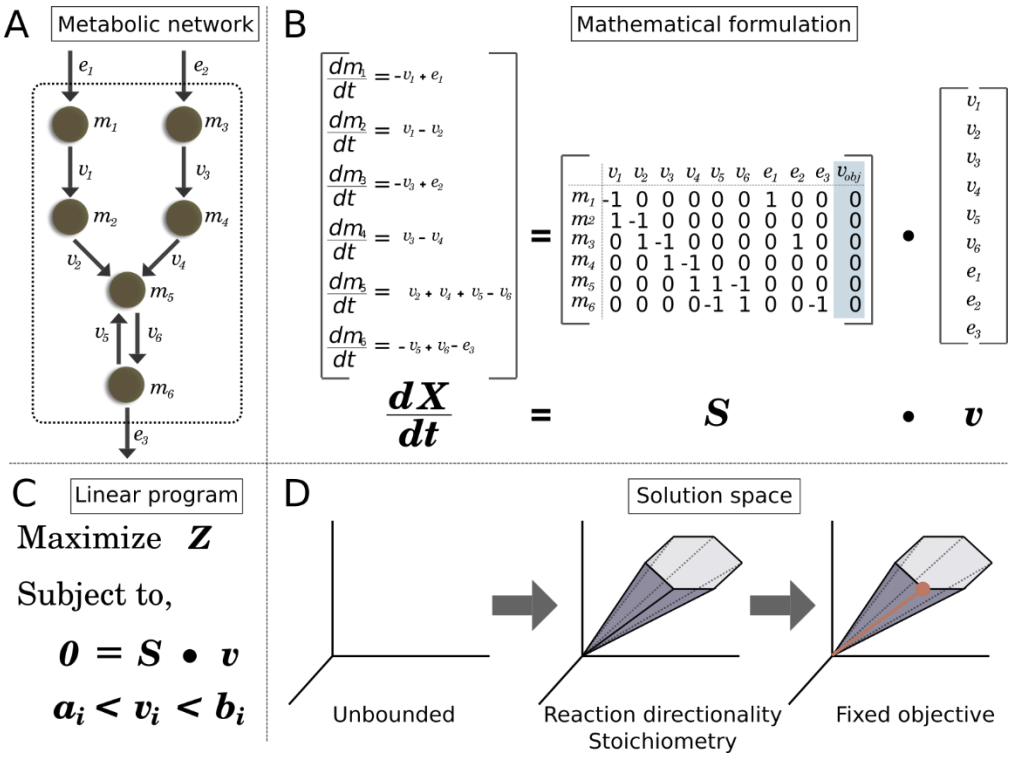

Figure 1.3 Modélisation basée sur les contraintes à l'aide de la programmation linéaire. A. Un réseau métabolique donné composé de métabolites (nœuds) et de réactions (liens). B. Le réseau en A peut être représenté sous la forme d'une matrice stoechiométrique **S**. Dans cette

matrice, chaque ligne représente un métabolite tandis que chaque colonne est associée à une réaction. La variation de la concentration du métabolite dans le temps (dX/dt) peut alors être représentée comme le produit matrice-vecteur de **S** par **v**, le vecteur des flux pour chaque réaction du réseau. C. En définissant un objectif physiologiquement significatif **Z**, la solution optimale pour le réseau métabolique peut être représentée comme un problème d'optimisation linéaire avec des contraintes de flux sur les réactions métaboliques. À l'état d'équilibre, la variation de la concentration de métabolites est égale à 0. D. L'application de contraintes sur le problème d'optimisation limite l'espace de solutions tandis que l'application d'un objectif approprié permet de trouver la ligne d'optimalité dans cet espace de solutions convexe limité (Reproduit de Lachance et al., 2020).

1.2.2 Reconstruction du réseau métabolique

L'achèvement de séquences de génome entiers à l'ère de la génomique (Figure 1.1) a permis de générer des modèles métaboliques à l'échelle du génome (GEM). Dans la plupart des cas, l'annotation du génome permet d'associer une fonction à des protéines codées par un organisme. Pour les enzymes métaboliques, l'annotation, associée à une recherche bibliographique approfondie, permet de lier une séquence d'ADN à une réaction biochimique dans le réseau métabolique. Le processus d'extraction d'un nombre maximum de réactions du génome est appelé reconstruction et a été examiné en détail (Thiele and Palsson, 2010). Ici, j'explique les étapes clés de la reconstruction d'une matrice stoechiométrique à l'échelle du génome (Figure 1.4).

Tout d'abord, une ébauche de reconstruction doit être générée. Le processus de construction de cette ébauche peut être effectué manuellement ou automatiquement. Les méthodes automatisées pour la reconstruction de l'ébauche des réseaux métaboliques sont examinées dans la section 3 de ce chapitre. Le processus de génération de cette esquisse initiale consiste à extraire les réactions biochimiques de l'annotation du génome. Grâce à ce processus, la stoechiométrie de chaque réaction du réseau métabolique est obtenue. Les réactions peuvent être extraites des numéros EC ou encore des noms de gènes annotés. Les gènes métaboliques candidats sont alors liés à une réaction de la matrice **S**. L'association entre un gène et sa réaction

est essentielle pour les prédictions qui seront générées par le modèle et doit donc être évaluée avec soin.

Ensuite, l'ébauche initiale est examinée de plus près grâce à une étape de raffinement. Les éléments clés de cette étape sont l'examen de la conservation du bilan massique pour chaque réaction; c'est-à-dire que le nombre d'atomes dans les réactifs soit égal au nombre d'atomes dans les produits. Le même raisonnement s'applique à la charge des réactions. Les équations équilibrées doivent avoir une charge neutre. Ces hypothèses sont liées aux principes fondamentaux de la chimie, ce qui garantit qu'aucune masse ou charge n'est créée dans une réaction du réseau métabolique. L'association gène-protéine-réaction (GPR) est ensuite vérifiée pour toutes les réactions et un score de confiance est attribué, ce qui facilite une évaluation plus approfondie des résultats lorsque les simulations du modèle sont comparées aux données expérimentales.

Les réactions non associées à un gène sont ensuite ajoutées. Parmi celles-ci, les réactions spontanées représentent l'occurrence naturelle d'une réaction thermodynamiquement favorable sans qu'il soit nécessaire de recourir à un catalyseur codé par un gène (enzyme). Les autres réactions non associées à un gène sont l'échange, les puits et les demandes. Les réactions d'échange représentent l'environnement ou encore le milieu de culture de la cellule. Elles ne sont pas équilibrées en masse ou en charge par défaut puisqu'elles représentent l'absorption ou le rejet de métabolites depuis ou vers le milieu. Elles sont néanmoins nécessaires pour la simulation des phénotypes de croissance dans un environnement donné. Les réactions de demandes représentent généralement une évacuation nécessaire de métabolite à partir du milieu alors que les puits sont des évacuations possibles de métabolites, généralement au niveau intracellulaire. Enfin, une réaction représentant la biomasse cellulaire produite et une de maintenance de l'ATP (ATPM) sont ajoutées. L'idée de l'équation de biomasse est de forcer le modèle à produire les métabolites nécessaires à la croissance de l'organisme et son impact dans la simulation de la croissance sera discuté plus tard. La réaction ATPM est une réaction d'hydrolyse de l'ATP qui permet à l'usager du modèle de fixer un certain taux de consommation

d'ATP pour une cellule en croissance. La connaissance des besoins énergétiques expérimentaux permet ainsi de prévoir le taux de croissance avec plus de précision.

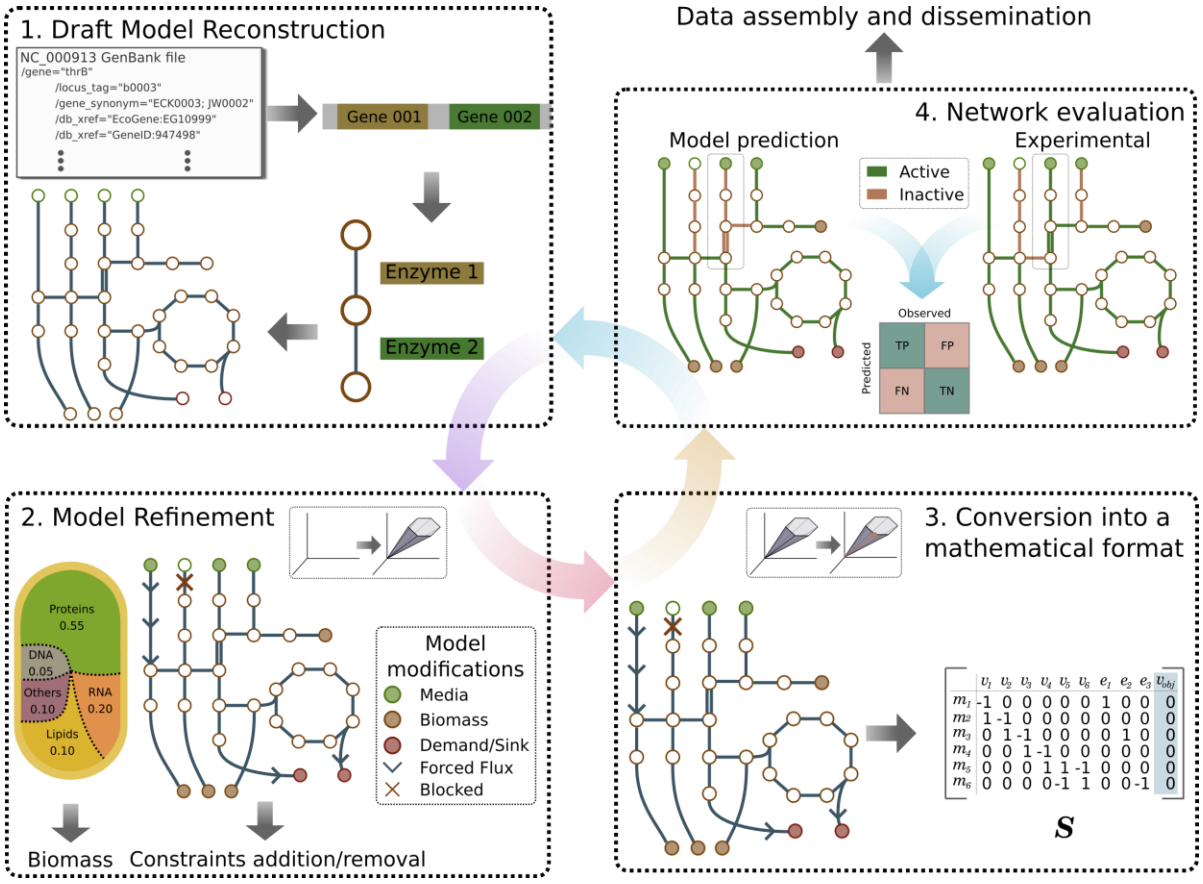

Figure 1.4 Les quatre principales étapes de la reconstruction et de la simulation du réseau métabolique. 1) Une ébauche de reconstruction est générée à partir de l'annotation du génome. 2) L'ébauche de reconstruction est affinée en générant une équation de biomasse et en appliquant des contraintes. 3) La reconstruction est convertie en un problème d'optimisation permettant la simulation du phénotype cellulaire. 4) Les prédictions du modèle sont comparées aux observations expérimentales. Les écarts entre les prédictions et les observations sont utilisés pour améliorer le modèle dans un processus itératif (Reproduit de Lachance et al., 2020).

Finalement, le réseau reconstruit est prêt pour la validation et la simulation. L'établissement de contraintes de simulation associées à un objectif défini permet de formuler des prédictions qui peuvent être testées. Le modèle peut ou non donner une solution réalisable. Dans ce dernier cas, des tests unitaires approfondis peuvent être nécessaires pour résoudre les problèmes liés à

la reconstruction formulée (par exemple, l'accumulation de métabolites). La résolution itérative de ces problèmes permet de générer un modèle fonctionnel qui peut être utilisé pour la simulation. La comparaison des prédictions formulées avec les données expérimentales disponibles peut soit confirmer la manière dont le système est censé fonctionner, soit révéler d'éventuelles lacunes dans les connaissances.

1.2.3 Fonction objective

Dans la programmation linéaire, la fonction objective est la valeur numérique à maximiser ou à minimiser. L'importance de la valeur à optimiser dépend de la situation que le modélisateur souhaite simuler. Par exemple, Papoutsakis (Papoutsakis, 1984) a généré un modèle pour la production de butyrate. Dans ce cas, la valeur numérique est la quantité de butyrate (une matière première de l'industrie chimique) qui peut être produite au fil du temps. Pour simuler et expliquer le surplus d'acétate, Majewski et Domach (Majewski and Domach, 1990) ont maximisé la production d'ATP par le réseau. Enfin, le modèle de globules rouges (RBC) (Bordbar et al., 2011) maximise le flux à travers la pompe à ATPase Na+/K+. Le choix de la fonction objective reflète donc la situation physiologique et détermine la qualité des prédictions générées par le modèle. Comme la RBC ne peut pas se reproduire, on suppose que l'objectif biologique réel de la cellule est de maintenir un gradient approprié de sodium et de potassium, une tâche qui nécessite la production d'énergie sous forme d'ATP. Cette définition appropriée de la fonction objective, ainsi que l'intégration de données expérimentales à haut débit, ont permis d'identifier des biomarqueurs de la dégradation des globules rouges lors du stockage (Yurkovich et al., 2017).

Un objectif commun des modélisateurs est de prédire un phénotype de croissance. Dans ce cas, une fonction objective de la biomasse (BOF) est définie et contient chaque métabolite nécessaire au doublement de la cellule (Feist and Palsson, 2010). La BOF est modélisée par l'ajout d'une réaction supplémentaire (colonne) dans la matrice stoechiométrique (**S**). Les

proportions de chaque élément de la cellule sont données sous forme de coefficients stoechiométriques dans la réaction. Afin de fournir une estimation du taux de croissance, une base est donnée (Varma and Palsson, 1993) telle que le produit du poids cellulaire par le temps est égal à 1 gramme de poids sec cellulaire par heure (gDW/hr). Bien que la composition en métabolites de la BOF puisse varier d'une espèce à l'autre, de nombreuses composantes nécessaires à la croissance sont partagées entre les procaryotes (Xavier et al., 2017). L'intégration correcte des composants de la biomasse effectivement présents ainsi que les coefficients stoechiométriques qui reflètent la composition expérimentale de la cellule (Beck et al., 2018) chez les espèces modifie la précision des prévisions du modèle (Lachance et al., 2019b). La définition de la BOF est donc cruciale pour générer des prédictions sur l'essentialité des gènes, une capacité très importante pour arriver à générer des cellules minimales *in silico* en utilisant la modélisation métabolique à l'échelle du génome.

1.2.4 Conversion en un format mathématique et évaluation

Tel que mentionné précédemment, les GEMs ont le pouvoir de simuler les capacités métaboliques d'un organisme. La conversion de la reconstruction en un format mathématique par l'établissement d'un objectif approprié (par exemple, la définition précise de la fonction objective de biomasse) et de contraintes (par exemple, la définition des milieux, les limites des flux internes, les taux d'absorption et de sécrétion). Le modèle peut ensuite être utilisé pour formuler des prédictions de l'état métabolique de l'organisme. Les prédictions formulées par le modèle et les ensembles de données utilisés pour les valider varient ainsi en fonction de l'objectif scientifique de la recherche menée.

Un objectif communément utilisé pour améliorer la qualité du modèle est d'optimiser la croissance (maximiser le flux à travers la réaction de biomasse). La prédiction directe est le taux de croissance, qui peut être mis en correspondance avec la valeur déterminée expérimentalement. L'obtention d'un temps de doublement correct dépend de la détermination

correcte des dépenses énergétiques cellulaires et des coefficients stoechiométriques des précurseurs de la biomasse inclus dans la BOF. L'optimisation pour la production de biomasse peut également être utilisée pour déterminer l'essentialité des gènes en éliminant itérativement des gènes individuels et en résolvant le modèle, une mesure commune de la qualité d'un modèle, qui sera traitée plus en détail ultérieurement (voir section 1.4). Enfin, la méthode FBA fournit un état de flux avec la solution donnée. Bien qu'une solution optimale unique soit trouvée pour la fonction objective choisie, de nombreux états de flux peuvent y conduire. Différentes méthodes ont été développées pour étudier la variabilité des états de flux (Monk et al., 2014) et seront couverts à la section 1.4. Les usagers du modèle peuvent alors échantillonner et étudier la variabilité de l'état de flux pour identifier les flux qui sont hors des plages biologiquement réalisables, ce qui permet d'appliquer des contraintes supplémentaires susceptibles d'améliorer la qualité du modèle.

La conformité aux données expérimentales peut alors être évaluée. Tel que mentionné, la prédiction d'essentialité des gènes du modèle est couramment utilisée comme référence pour évaluer la qualité générale d'un modèle puisqu'elle tient compte de la qualité des GPR attribués ainsi que de la topologie du réseau, de la biomasse et de la composition du milieu. Une matrice de Punnet est souvent utilisée pour visualiser certaines prédictions formulées par le modèle (ex. : essentialité) les quatre combinaisons de vrai/faux positif/négatif étant représentées. La précision ou le coefficient de corrélation de Matthews (MCC) peuvent alors être utilisés pour résumer la qualité de la prédiction du modèle en une seule valeur numérique.

1.3 MÉTHODES DISPONIBLES POUR LA RECONSTRUCTION MÉTABOLIQUE À L'ÉCHELLE DU GÉNOME

Au cours des deux dernières décennies, la disponibilité croissante de séquences de génome entier pour une diversité d'espèces a entraîné une augmentation constante du nombre de GEMs (Monk et al., 2014). Le nombre d'outils bio-informatiques adaptés à la reconstruction des

réseaux métaboliques ainsi qu'à l'analyse et l'intégration de données omiques dans ces modèles ont évolué en conséquence (Lewis et al., 2012). Dans cette section, je passe en revue les méthodes et les bases de données utilisées pour la reconstruction de la matrice stoechiométrique (**S**), les méthodes de remplissage des trous dans les réseaux métaboliques ("gapfilling") et de définition des objectifs.

1.3.1 Outils pour la reconstruction du réseau

Comme il a été mentionné, la reconstruction d'un GEM commence par la reconstruction de la matrice stoechiométrique contenant l'ensemble des réactions et métabolites présents dans un organisme donné (Figure 1.4). Une inspection minutieuse de l'annotation du génome permet de relier un gène et sa séquence à une fonction particulière dans le réseau. Afin de relier ces éléments entre eux, les usagers du modèle peuvent utiliser les nombreuses bases de données publiques contenant voies et réactions biochimiques qui sont spécifiquement conçues pour fournir l'association entre les gènes, les réactions biochimiques et/ou les voies métaboliques (Artimo et al., 2012; Aziz et al., 2008; Caspi et al., 2008; Devoid et al., 2013; Fabregat et al., 2017; Kanehisa et al., 2017; King et al., 2016; Placzek et al., 2017; Wattam et al., 2017).

L'identification des gènes candidats métaboliques dans le génome de référence est la première étape de la reconstruction métabolique à l'échelle du génome. Pour ce faire, les usagers du modèle peuvent soit obtenir les numéros de commission des enzymes (EC) à partir d'un logiciel spécialisé (Nursimulu et al., 2018), soit extraire les informations contenues dans les bases de données accessibles au public. Dans les deux cas, la standardisation de l'identificateur du métabolite et de la réaction est essentielle pour la cohérence et la lisibilité du modèle. Comme ces identificateurs varient considérablement d'une base de données à l'autre, les projets de reconstructions peuvent ne pas être lisibles dans un autre format. Ce type de problème a été traité et peut potentiellement être surmonté par l'utilisation de MetaNetX (Moretti et al., 2016) ou de BiGG (King et al., 2016). MetaNetX est une plateforme basée sur le web qui tente de

centraliser l'identification des métabolites et des réactions, tout en fournissant des méthodes de reconstructions automatisées à l'échelle du génome. L'objectif principal de la base de données BiGG est de répertorier les GEM formulés dans la nomenclature BiGG. Néanmoins, les réactions et les métabolites stockés sur BiGG sont liés à d'autres bases de données couramment utilisées telles que Reactome, KEGG, SEED, CHEBI, BioCyc et MetaNetX. Choisir un système d'identification et assurer la conversion d'un système d'annotation à un autre est donc essentiel pour l'établissement du projet de reconstruction du modèle.

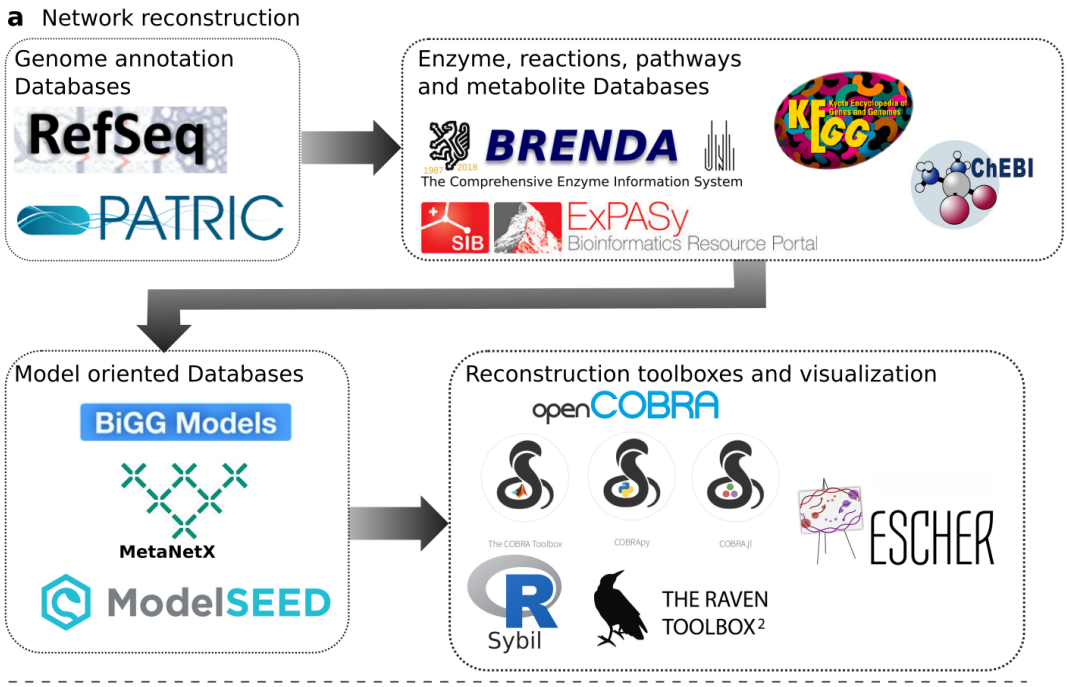

Figure 1.5 Outils pour la reconstruction et l'analyse à l'échelle du génome. A. Liste non exhaustive d'outils informatiques et de bases de données pour la reconstruction des réseaux métaboliques. L'interrogation des bases de données d'annotation et de réaction permet d'associer les réactions et métabolites à des identificateurs d'un format standard provenant de

bases de données dédiées à la modélisation métabolique. Les boîtes à outils de reconstruction sont conçues pour faciliter la création d'objets dans un langage de programmation qui sont utiles à la modélisation (réactions, métabolites, gènes et modèles). B. Liste non exhaustive d'outils informatiques pour faciliter l'identification des lacunes dans le réseau métabolique et les objectifs cellulaires (Reproduit de Lachance et al., 2020).

La reconstruction du réseau peut être exécutée dans différents cadres en fonction des préférences du modélisateur. Les logiciels SEED (Devoid et al., 2013) et Merlin (Dias et al., 2015) permettent tous deux la génération automatisée de GEM. Bien que ces modèles fonctionnels fournissent des prédictions, une recherche documentaire exhaustive et un réglage précis du modèle sont généralement nécessaires avant sa publication (Thiele and Palsson, 2010). La suite COBRA (Reconstruction et analyse basées sur les contraintes) est conçue pour inclure chaque étape du processus et est actuellement disponible dans trois langages de programmation différents : Python (COBRApy, (Ebrahim et al., 2013)), MatLab (COBRA Toolbox 3.0, (Heirendt et al., 2019)) et Julia (COBRAjl, (Heirendt et al., 2017)). Implémentée dans MatLab, la boîte à outils RAVEN (Agren et al., 2013) est une autre option de reconstruction qui inclut également la visualisation des réseaux métaboliques. La boîte à outils Sybil permet aux utilisateurs de R d'utiliser les méthodes FBA, MOMA (Segrè et al., 2002) et ROOM (Shlomi et al., 2005) dans leur langage de programmation préféré (Gelius-Dietrich et al., 2013). Bien que la suite COBRA n'inclue pas spécifiquement la visualisation, le réseau généré peut être visualisé en construisant une carte métabolique avec Escher (King et al., 2015).

1.3.2 Outils pour l'analyse des réseaux

Les principales fonctionnalités des boîtes à outils dédiées à la reconstruction métabolique mentionnées ci-dessus consistent à permettre la création “d'objets” (terme spécifique à la programmation orientée objet) fondamentaux des modèles (principalement sous-forme “d’objets” métabolites, réactions ou encore gènes) et à les stocker dans un “objet” modèle qui peut être sauvegardé ou importé dans le(s) format(s) souhaité(s). Ces boîtes à outils

comprennent également des fonctionnalités de base de simulation de modèles telles que la définition de l'objectif et un code enveloppe (“wrapper”) vers l'interface du solveur nécessaire pour optimiser le modèle. Ces fonctionnalités de base pour la simulation sont utiles pour convertir le modèle en un format mathématique qui peut ensuite être utilisé pour des processus de simulation plus avancés et l'évaluation des capacités métaboliques de l'organisme. Je couvre ici certains algorithmes qui ont été développés pour augmenter la qualité des modèles avant qu'ils ne soient utilisés pour la simulation (Figure 1.5).

1.3.2.1 Remplissage des trous dans le réseau

Afin de révéler les capacités biologiques d’un organisme, le réseau doit être le plus fonctionnel possible, c'est-à-dire que le flux soit en mesure de passer par autant de réactions que possible. Comme il a été montré, la formulation mathématique du FBA repose sur l'hypothèse d'équilibre qui ne permet pas l'accumulation de métabolites. Cela signifie que pour une voie métabolique linéaire donnée, une seule réaction manquante bloquerait le flux à travers toutes les réactions en amont et en aval. La voie entière serait alors considérée comme non fonctionnelle, une hypothèse dont la signification biologique reste discutable et qui devrait donc être traitée avec précaution par les usagers du modèle.

Plusieurs algorithmes ont été développés dans le but d'identifier, de résoudre les lacunes des réseaux biologiques et de proposer des gènes qui pourraient catalyser la ou les réactions suggérées (Orth and Palsson, 2010; Pan and Reed, 2018). Comme mentionné, le cadre général de ces algorithmes identifie d'abord les métabolites “cul-de-sac”, c'est-à-dire les métabolites qui ne peuvent être produits ou consommés dans le réseau métabolique. La résolution d'une lacune dans le réseau peut être réalisée en ajoutant une ou plusieurs réactions. Pour trouver des réactions candidates, ces algorithmes interrogent généralement des bases de données de réactions plus importantes telles que celles contenues dans KEGG (Kanehisa et al., 2017) ou MetaCyc (Caspi et al., 2008). La valeur d'ajouter spécifiquement une ou plusieurs réactions ne

peut être mesurée qu'en étudiant le lien entre le mécanisme proposé et le contexte dans lequel il s'inscrit dans l'espèce étudiée. Par conséquent, la troisième étape des algorithmes de remplissage des lacunes vise à identifier les meilleurs gènes possibles qui peuvent s'associer à ces réactions.

Le premier algorithme du genre à avoir été développé se nommait GapFilling (Satish Kumar et al., 2007) et n'incluait pas cette troisième étape. Cependant, des versions ultérieures ont intégré différentes façons d'inclure les données expérimentales avec les réactions suggérées. Globalfit (Hartleb et al., 2016) et ProbannoPy (King et al., 2018) sont de bons exemples de méthodes de remplissage des lacunes visant à améliorer un modèle métabolique basé sur des données expérimentales. Pour une couverture plus approfondie des méthodes disponibles, les lecteurs intéressés peuvent consulter cette étude de Pan et Reed (Pan and Reed, 2018). Globalfit a été utilisé pour améliorer la qualité de deux GEMs, ceux de *Escherichia coli* (*i*JO1366) et de *Mycoplasma genitalium* (*i*PS189). Il utilise un problème d'optimisation à deux niveaux afin de minimiser l'écart entre l'essentialité prédite des gènes et les données expérimentales, en permettant l'incorporation de nouvelles réactions métaboliques au sein du modèle ou de nouvelles réactions d'échange (composants du milieu), ainsi que des métabolites de la BOF. Probanno (Web et Py) attribue une probabilité basée sur le nombre attendu de découvertes de qualité similaire ("e-value") de la recherche BLASTp pour classer les réactions utilisées afin de combler les lacunes du réseau.

De telles approches sont pertinentes dans le contexte actuel de recherche et de conception de cellules minimales. Même si une cellule minimale a déjà été générée expérimentalement, le nombre de gènes qu'elle contient et pour lesquels une fonction précise n'a pu être attribuée représente une partie importante du génome complet (149/473). Un châssis cellulaire idéal ne devrait pas avoir de propriétés inconnues (Danchin, 2012), car il doit servir de modèle pour les futures conceptions de génomes. Par conséquent, la reconstruction des réseaux métaboliques et l'utilisation d'algorithmes de remplissage des lacunes ("GapFilling") qui fournissent une

annotation fonctionnelle est un moyen systématique de combler les lacunes en matière de connaissances.

1.3.2.2 Fonctions objectives

Les objectifs métaboliques des cellules peuvent être résumés dans une réaction de la matrice stoechiométrique et fixés comme objectif : la fonction objective de biomasse (BOF). L'identification des composants clés nécessaires à la croissance d'une cellule est néanmoins une tâche ardue. Ce processus peut être accompli d'une manière biaisée, qui tente d'incorporer autant que possible les connaissances actuelles sur la composition de l'organisme, ou d'une manière non biaisée dans laquelle les données expérimentales sont utilisées pour déduire les objectifs cellulaires. Rocha et ses collègues ont fait un effort louable pour résumer les connaissances actuelles sur la composition de la biomasse procaryote (Xavier et al., 2017). Dans cette étude approfondie, la composition de la biomasse de 71 modèles préparés manuellement et disponibles dans la base de données BiGG (King et al., 2016) a été comparée avec la distance phylogénétique des espèces qu'ils représentent. L'échange de la BOF d'un modèle à l'autre a montré que la prédiction de l'essentialité de la réaction est sensible à la composition de la BOF. En étudiant davantage l'impact de la composition de la biomasse sur les prédictions d'essentialité des gènes de plusieurs espèces, les auteurs ont trouvé un ensemble de cofacteurs universellement essentiels chez les procaryotes. Ces connaissances fondamentales soulignent l'importance de la précision des BOFs pour la prédiction de l'essentialité des gènes par les GEMs et constituent une ressource importante pour les travaux de modélisations subséquents.

En utilisant des composants cellulaires essentiels préalablement établis, les usagers du modèle peuvent en partie définir la BOF de leur organisme d'intérêt. Néanmoins, la partie restante de la BOF est spécifique à l'espèce et peut être complétée en utilisant une approche non biaisée. Comme dans le cas du remplissage des lacunes, la recherche d’objectif cellulaire peut être

effectuée de manière algorithmique. Historiquement, la plupart des algorithmes développés à cette fin ont utilisé les données d'analyse du flux métabolique (AFM) ainsi que diverses méthodes d'optimisation (Burgard and Maranas, 2003; Gianchandani et al., 2008; Zhao et al., 2016). Bien que l'AFM soit un type de données particulièrement bien adapté aux modèles de flux, le nombre de flux générés par les méthodes à la fine pointe demeure grandement inférieur au nombre de réactions incluses dans les GEMs. Des algorithmes récemment développés tentent donc d'utiliser d'autres types de données pour trouver des objectifs cellulaires. BOFdat (Lachance et al., 2019b) utilise un algorithme génétique pour trouver les compositions de biomasse qui offrent la meilleure correspondance entre l'essentialité génétique prédite et expérimentale. Les métabolites identifiés par l'algorithme sont ensuite regroupés en fonction de leur distance relative dans le réseau métabolique pour former des groupes d'objectifs métaboliques qui peuvent être interprétés par les usagers du modèle. Cette méthode sera discutée plus en détail dans le chapitre 2. Une autre approche appelée BIG-BOSS intègre plusieurs types de données omiques pour formuler l'objectif cellulaire en utilisant un modèle contraint par le protéome, avec un problème d'optimisation à deux niveaux, similaire à celui de BOSS (Gianchandani et al., 2008). L'utilisation de cette méthode a montré qu'en combinant l'AFM pour un sous-ensemble de flux avec la protéomique, la composition de la biomasse a été récupérée de manière plus précise qu'en utilisant un seul type de données.

1.4 INTÉGRATION DES DONNÉES ET PRÉDICTIONS PHÉNOTYPIQUES

Une fois qu'un GEM est reconstruit, converti dans un format mathématique et validé avec des données expérimentales, il est possible de générer systématiquement des hypothèses à partir du modèle qui guideront la conception de la souche souhaitée. Tout comme la conception d'une souche de production, la réalisation d'une cellule minimale nécessite une connaissance approfondie de l'organisme, qui elle peut être acquise grâce à la génération de données à haut débit. L'intégration de ces données est rendue possible par les GEMs et une pléthore de logiciels a été écrite pour aider les usagers du modèle dans cette tâche. Je traite ici des méthodes

disponibles pour l'intégration de données à haut débit ainsi que des algorithmes de conception de contraintes qui peuvent être utilisés pour la conception de cellules minimales synthétiques (Figure 1.6).

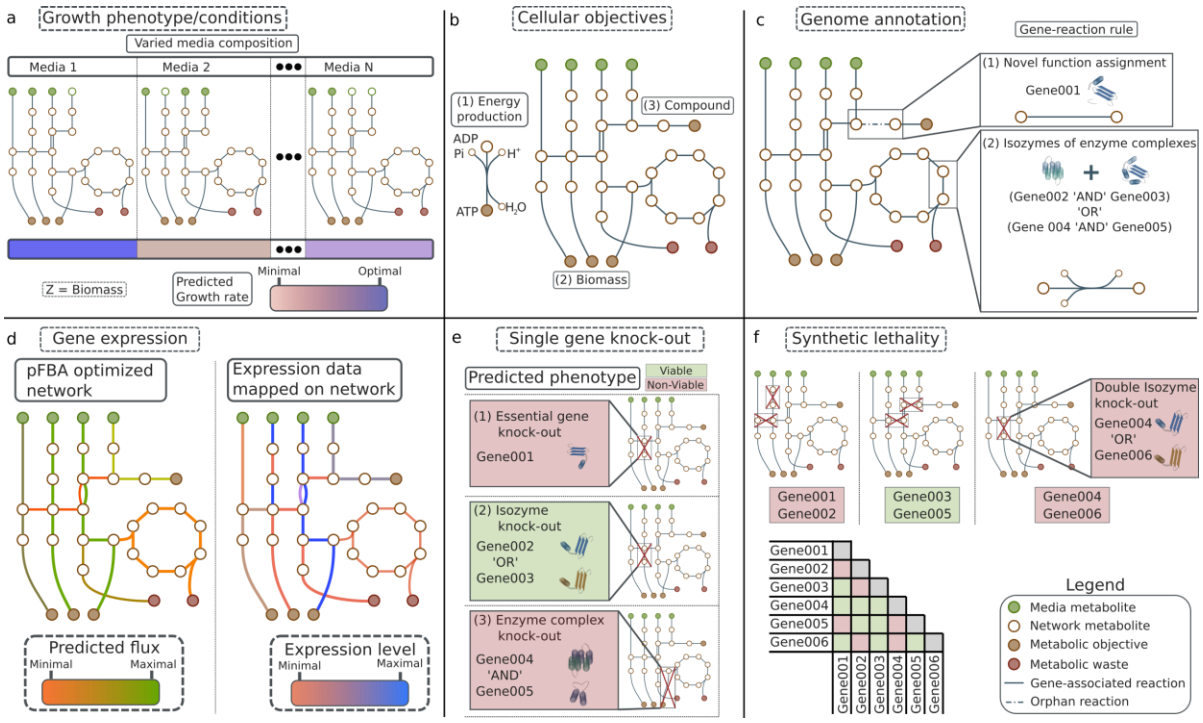

Figure 1.6 Les multiples utilisations des modèles à l'échelle du génome. A. Composition des milieux de croissance. L'ouverture des réactions d'échange détermine la composition du milieu *in silico*. La modification de la composition du milieu de simulation a un double impact. Premièrement, les métabolites nécessaires à la croissance sont identifiés, permettant de définir un milieu de culture minimal. Deuxièmement, les capacités de croissance de l'organisme sont explorées, c'est-à-dire l'identification des sources d'azote et de carbone permettant la croissance. B. Objectifs cellulaires. L'objectif cellulaire peut être modifié pour répondre aux besoins de modélisation. La production d'énergie, de biomasse ou d'un métabolite pertinent peut être étudiée. C. Annotation du génome. Les réactions dans le réseau sont associées à un gène via l'association entre les gènes et les réactions. Le modèle métabolique fournit alors un cadre dynamique dans lequel il est possible de contextualiser l'annotation du gène et d'en examiner la fonction. D. Expression des gènes. L'expression génique parcimonieuse (pFBA) est utilisée pour générer un état de flux optimal en supposant une utilisation optimale de l'ensemble des enzymes présentes dans la cellule. Le flux de chaque réaction peut ensuite être comparé au niveau d'expression des gènes. E. Essentialité des gènes. Les gènes et les réactions sont associés entre eux par la règle de la réaction génique (GPR). Un gène est considéré comme essentiel si sa délétion l'identifie comme le seul déterminant de la production d'un métabolite de la biomasse. F. Létalité synthétique. Les GEMs permettent de réaliser des knock-outs

simultanés, ce qui constitue un atout important pour la réduction du génome et la production de cellules minimales synthétiques (Reproduit de Lachance et al., 2020).

1.4.1 Objectifs cellulaires et prédiction de l'essentialité des gènes

Un concept clé pour la conception de cellules minimales est l'identification du contenu génique amovible. C'est-à-dire : "quels sont les gènes non essentiels dans les conditions de culture en laboratoire?". Pour formuler une telle prédiction, il faut d'abord déterminer les métabolites requis pour la croissance (Figure 1.6). Comme il a été mentionné, ceux-ci sont représentés par la BOF dans les GEM. La définition de la BOF est étroitement liée à la pression évolutive appliquée sur la souche, qui est à son tour fonction de son environnement de croissance. Par exemple, supposons qu'une souche d'*E. coli* passe soudainement de conditions aérobies à anaérobies. Les modifications rapides du phénotype sont le résultat de propriétés chimiques et physiques, c'est-à-dire : utilisation d'un nouveau substrat, changement d'état métabolique, modification de l'expression des gènes, etc. Cette adaptation rapide peut être qualifiée de causalité proximale (Palsson, 2015) car la cause de la modification de phénotype est très proche de l'observation. Sa contrepartie, appelée causalité distale, se produit au fil du temps et est le résultat d'une adaptation évolutive. La causalité distale est propre aux systèmes biologiques et implique une modification du génotype pour s'adapter aux contraintes imposées par l'environnement dans lequel l'espèce est cultivée. Comme la composition de la biomasse d'une cellule est le résultat de son évolution, chaque espèce a besoin de métabolites différents pour sa croissance, dont certains composants essentiels sont partagés par un large éventail d'organismes (Xavier et al., 2017) (Figure 1.6).

1.4.1.1 Prédiction de l'essentialité des gènes

Les GEMs peuvent être utilisés pour formuler des prédictions quant à l'essentialité de réactions ou de gènes (Figure 1.6). Pour formuler cette prédiction, chacune des réactions du modèle est

retirée individuellement et, à chaque fois, le modèle est optimisé pour la croissance (Joyce and Palsson, 2008; Suthers et al., 2009a). Un seuil de taux de croissance approprié est nécessaire afin de distinguer les phénotypes viables des phénotypes non viables pour l'ensemble des réactions. Cette approche revient à déterminer quelles réactions sont essentielles pour assurer le flux à travers la réaction de biomasse du modèle. La définition correcte de la BOF est donc essentielle pour la prédiction précise de l'essentialité des gènes. La définition qualitative de la BOF définit les besoins de croissance de l'organisme et les voies métaboliques qui conduisent à la production de ces précurseurs sont ensuite activées. D'autres contraintes telles que le milieu de croissance, les taux d'absorption des principales sources de carbone et/ou d'oxygène ont également un impact sur les prédictions de l'essentialité des gènes.

La valeur ajoutée des GEMs est que la plupart des réactions qui les composent sont associées à un ou plusieurs gènes. Cette association entre un gène et sa ou ses réactions est appelée GPR et rend compte des réactions catalysées par un seul gène ou plusieurs gènes dans un complexe, symbolisé par une règle "et", ainsi que des isozymes, symbolisées par une règle "ou" (Figure 1.6). L'essentialité des gènes à l'échelle du modèle entier peut être générée facilement à l'aide des boîtes à outils de reconstruction mentionnées précédemment, car elles comprennent une implémentation de cette fonction.

Il convient de noter que les GEMs sont très efficaces pour prédire l'essentialité des gènes. Des modèles très élaborés comme celui d'*E. coli* ont permis de prédire l'essentialité dans différentes conditions de croissance avec une précision allant jusqu'à 93,4% (Monk et al., 2017). La qualité de la prédiction repose à la fois sur le haut niveau d'informations biochimiques incluses dans la reconstruction d'*E. coli* et sur la connaissance précise des conditions de croissance. Ces limites seront discutées plus tard.

1.4.1.2 Au-delà de la délétion d'un seul gène

Un avantage des GEMs est la capacité à formuler des prédictions de létalité synthétique (SL) (Figure 1.6). Ce phénomène a été signalé au début de l'ère classique de la biologie pour tenter de décrire l'observation selon laquelle la combinaison de traits observables ne donnait pas de descendants viables (Bridges, 1922). Au niveau des gènes, la létalité synthétique est connue comme l'observation selon laquelle l'élimination simultanée de deux gènes produit un phénotype létal lorsque leur élimination individuelle indépendante produit un phénotype viable (Figure 1.6). L'étude expérimentale de la SL au niveau des systèmes est complexe car elle implique le criblage de plusieurs combinaisons d'inactivation de gènes. Pour un organisme contenant un nombre N de gènes individuellement non essentiels, le nombre de combinaisons est le coefficient binomial :

$$\frac{n!}{k!\,(n-k)!} \qquad \text{(équation 1.4)}$$

L'obtention de toutes les combinaisons possibles de SL pour un organisme implique la génération d'une banque d'inactivation de gènes sur une autre bibliothèque d'inactivation. Cette tâche a été accomplie pour des organismes très étudiés tels que *S. cerevisiae* pour lesquels les méthodes d'édition de gènes sont courantes (Deutscher et al., 2006; Goodson et al., 1996) mais pour la plupart des espèces, elles sont généralement trop exigeantes pour être générées.

La détermination informatique des gènes SL en utilisant les modèles FBA est coûteux en termes d'heures de calcul, mais est tout de même bien plus rapide que la génération expérimentale de ces données. L'utilisation de cette approche peut donc guider la conception de génomes minimaux puisqu'elle ajoute un niveau d'information qui ne pourrait pas être obtenu autrement à partir du génome ou des bibliothèques d'inactivation de simple gène. Les GEMs offrent également la possibilité d'étendre l'étude des SL à d'autres paires de gènes et d'inclure des triples ou quadruples knock-outs (Suthers et al., 2009a), un avantage indéniable

par rapport à l'approche strictement expérimentale. La précision de ces prédictions par les GEM est toutefois dépendante de la qualité des associations entre les gènes et les réactions et d'une définition adéquate de la fonction objective de biomasse.

Une utilisation intéressante de l'analyse des gènes SL est l'algorithme MinGenome écrit par Wang et Maranas (Wang and Maranas, 2018). Cet algorithme prend en entrée la séquence du génome de l'organisme d'intérêt, un GEM, des données d'essentialité *in vivo* à l'échelle du génome, des sites d'opéron et de promoteur et des informations sur les facteurs de transcription. En utilisant ces informations, MinGenome trouve de manière itérative la plus grande section d'ADN pouvant être retirée sans être létal pour la cellule. La structure de l'opéron ainsi que les promoteurs et les informations sur les facteurs de transcription sont dès lors utilisés pour maintenir les éléments de régulation en place, ce qui devrait augmenter la probabilité que le génome minimal suggéré soit fonctionnel *in vivo*.

1.4.2 Intégration de plusieurs ensembles de données « omiques »

Comme mentionné précédemment, l'ère génomique a permis la génération de données à haut débit ("omique") pour de nombreux types de molécules. L'intégration de ces ensembles de données est importante pour augmenter le niveau de connaissance biologique mais nécessite une méthode appropriée et bien structurée. Il a été démontré que les modèles métaboliques fournissent un moyen systématique d'intégrer de multiples ensembles de données omiques pour une compréhension mécanistique (Bordbar et al., 2014; Monk et al., 2014). Ralser et ses collègues ont discuté de l'intégration de 7 types d'ensembles de données omiques : génomique, transcriptomique, protéomique, lipidomique, métabolomique, ionomique et phénomique (Haas et al., 2017). L'approche utilisée pour incorporer ces informations multi-omiques dans les GEMs sera discutée ci-dessous.

Les GEMs utilisent l'information génomique pour extraire les fonctions biologiques des gènes métaboliques. Bien que la régulation de l'expression des gènes ne soit pas prise en compte dans les modèles métaboliques, les ensembles de données transcriptomiques et protéomiques peuvent être utilisés pour appliquer des contraintes supplémentaires sur le modèle. Le flux au travers d'une réaction peut être limité en fonction du niveau d'expression du gène qui la catalyse ou simplement arrêté lorsque les gènes ne sont pas exprimés (Figure 1.6). Le concept de cellule minimale suppose une cellule très spécialisée avec des capacités métaboliques réduites. L'intégration d'ensembles de données d'expression des gènes dans les modèles permettrait donc de générer des modèles spécifiques au contexte particulier des cellules minimales.

D'autres ensembles de données caractérisent les molécules en dehors du dogme central de la biologie (Crick, 1970). Les concentrations de métabolites elles-mêmes ne sont pas incluses dans la méthode FBA standard, mais une variante appelée uFBA (Bordbar et al., 2017) permet l'incorporation de données métabolomiques variant en fonction du temps dans le GEM, ce qui permet de prédire avec plus de précision l'état métabolique de la cellule. Les résultats de lipidomiques et d'ionomiques sont quant à elles utiles pour déterminer la composition de la cellule, une information précieuse pour la définition de la BOF.

L'intégration de plusieurs ensembles de données omiques avec des modèles à l'échelle du génome permet de fournir une explication mécanistique du phénotype de l'organisme dans différents environnements (Lewis et al., 2010). En utilisant de multiples ensembles de données omiques, Lewis et al. ont montré que les souches d'*E. coli* évoluant dans des conditions différentes modifient leur patron d'expression génique d'une manière cohérente avec une variante du FBA appelée FBA parcimonieuse (pFBA). Le pFBA utilise une approche de programmation linéaire à deux niveaux pour minimiser le flux associé à l'ensemble des enzymes tout en maximisant la production de biomasse. La solution de flux générée à l'aide du pFBA a été montré comme étant cohérente avec l'expression différentielle des gènes dans différentes conditions et confirme la pertinence biologique du pFBA. En utilisant cette

méthode, il serait donc possible de prédire l'état optimal d'une cellule avant même sa réalisation expérimentale.

1.5 BIOLOGIE DES SYSTÈMES DES CELLULES MINIMALES

Depuis la proposition de Morowitz selon laquelle les cellules minimales permettraient de comprendre les principes de base de la vie (Morowitz, 1984), de nombreux efforts ont été déployés pour identifier des ensembles de gènes minimaux théoriques par le biais de la génomique comparative (Mushegian and Koonin, 1996), de l'analyse de l'essentialité des gènes (Glass et al., 2017) et d'une combinaison de ces approches (Baby et al., 2018a). La réduction du génome de bactéries complexes a également été tentée expérimentalement à plusieurs reprises (Choe et al., 2016) et finalement, près de dix ans d'efforts novateurs ont permis de réaliser une approximation fonctionnelle d'une cellule minimale *in vitro* (JCVI-syn3.0) (Hutchison et al., 2016; Sleator, 2010).

Il a été montré que l'utilisation des GEMs, des banques de connaissances du métabolisme structurées mathématiquement, permet de générer des prédictions phénotypiques à partir d'informations génomiques et peuvent donc être utilisées pour la conception rationnelle de cellules minimales (Wang and Maranas, 2018). Je vais maintenant passer en revue les GEMs pour certaines bactéries quasi-minimales naturelles de la classe des mollicutes, puis j'aborderai l'extension des méthodes de modélisation au-delà du métabolisme.

1.5.1 GEMs disponibles pour les organismes minimaux naturels

Les mollicutes ont fait l'objet de nombreuses recherches depuis qu'ils ont été proposés comme les plus petits organismes vivant de manière autonome (Morowitz and Tourtellotte, 1962). Une connaissance approfondie du métabolisme particulier de ces espèces (Miles, 1992) a permis de

générer des GEM pour les espèces les plus étudiées de ce groupe. Le premier GEM pour un mollicute a été reconstitué pour l'agent pathogène urogénital humain *Mycoplasma genitalium (Suthers et al., 2009b)*. Ce modèle comprend 189 gènes, 168 réactions associées à un gène et 274 métabolites. En utilisant les données expérimentales d'essentialité (Glass et al., 2017), le modèle était cohérent avec 87% des gènes essentiels et 89% des gènes non essentiels. Bien que la précision des prédictions de ce modèle soit élevée, plusieurs approximations ont été utilisées pour sa reconstruction. La composition de la biomasse ainsi que les coûts de maintenance associés à la croissance et à la non-croissance qui peuvent être calculés à partir du taux d'absorption du substrat et des taux de sécrétion ont été estimés à partir d'*E. coli*. Comme il n'existe pas de milieu défini pour *M. genitalium*, le milieu de croissance a également été estimé.

Anciennement connu sous le nom d'agent Eaton, *Mycoplasma pneumoniae* est associé à une pneumonie atypique chez l'homme (Dajani et al., 1965; Lind, 1966). De multiples efforts de caractérisation de *M. pneumoniae* ont été entrepris, permettant une ré-annotation du génome (Dandekar et al., 2000) et le transcriptome (Güell et al., 2009), le protéome (Kühner et al., 2009) et le métabolisme (Yus et al., 2009) ont été étudiés en profondeur. Cela a permis de générer un modèle quantitatif pour *M. pneumoniae* (Wodke et al., 2013). La quantité de données expérimentales disponibles a permis aux modélisateurs de comparer l'utilisation prévue du sucre et d'obtenir l'utilisation de l'énergie tout au long des phases de croissance. Le fait de contraindre le modèle avec ces données a permis de disséquer les utilisations des différentes voies métaboliques au cours des stades de croissance.

Les prédictions formulées par le modèle de *M. pneumoniae* ont révélé qu'une quantité substantielle d'ATP n'est pas dirigée vers la production de biomasse mais plutôt vers les fonctions de maintien des cellules telles que le repliement des protéines assisté par un chaperonne, le maintien de l'ADN et les modifications post-traductionnelles. Il est frappant de constater que l'ATPase est responsable de la plus grande partie de l'utilisation de l'énergie (57-80%) afin de maintenir le pH intracellulaire et un gradient de protons favorable à travers la membrane. Les auteurs suggèrent que quatre facteurs peuvent avoir un impact sur la

consommation globale d'énergie : la topologie du réseau métabolique, le taux de croissance, les conditions environnementales et la taille des cellules. Ces résultats sont particulièrement intéressants car ils montrent que l'utilisation d'une approche de biologie des systèmes telle que les GEMs pour la conception des bactéries peut aller au-delà de la prédiction de l'essentialité des gènes et révéler des propriétés intrinsèques affectant l'énergie cellulaire. Ces facteurs pourraient difficilement être prédits sans l'intégration de données expérimentales dans une banque de connaissances structurée mathématiquement.

1.5.2 Modélisation à l'échelle du génome des organismes minimaux synthétiques

Récemment, des efforts de modélisation ont été consacrés à JCVI-syn3.0, une approximation fonctionnelle d'une cellule minimale (Breuer et al., 2019). La reconstruction métabolique a été générée en utilisant l'annotation génique de la souche parentale JCVI-syn1.0 (*Mycoplasma mycoides*) pour laquelle de nombreuses informations sont disponibles. La collecte de l'ensemble des connaissances dans un format de calcul unique est une avancée significative afin de définir les besoins métaboliques d'une cellule minimale fonctionnelle. Comme il a été démontré, les GEMs peuvent être utilisés pour formuler des prédictions phénotypiques telles que l'essentialité des gènes et intégrer des données à haut débit comme l'expression des gènes (voir section 4). Dans le cadre de leur étude de modélisation, Breuer *et al.* ont aussi fourni un ensemble de données sur la mutagenèse par transposons à haute densité opérée sur JCVI-syn3.0 ainsi qu'un ensemble de données de protéomique quantitative. Les données d'essentialité des gènes ont permis d'identifier les divergences entre les prédictions et les observations du modèle. Parallèlement au processus de reconstruction, les auteurs ont pu formuler plusieurs hypothèses sur les fonctions restantes des gènes qui ne pouvaient pas être supprimées mais néanmoins inconnues.

L'utilisation de données de protéomique a permis de contextualiser l'activité des protéines exprimées dans JCVI-syn3.0, mais l'analyse est néanmoins limitée. En effet, alors que le GEM

résultant pour cet organisme est la première représentation *in silico* d'une cellule minimale synthétique, des prédictions plus précises du modèle auraient nécessité la composition détaillée de la biomasse de cette bactérie ainsi qu'un milieu complètement défini. L'inclusion de ces paramètres dans le modèle devrait élargir ses capacités de prédiction.

1.6 PERSPECTIVES SUR L'UTILISATION DE MODÈLES POUR LA CONCEPTION DE CELLULES MINIMALES

"What I cannot create, I do not understand"
-Richard Feynman

Un des principaux objectifs de la recherche sur les cellules minimales est de recueillir une compréhension exhaustive de la cellule. Le FBA permet de générer de multiples prédictions sur l'état métabolique de la cellule, mais sa portée est limitée au métabolisme, généralement ~30% du contenu génique bactérien. D'autres approches ont été développées qui permettent d'inclure les contraintes de diverses fonctions cellulaires telles que la machinerie d'expression, le réseau de régulation, la cinétique enzymatique et la thermodynamique. Suivant l'intégration de ces multiples contraintes dans un modèle unique, il sera possible d'étendre ou de remplacer la proposition initiale de Morowitz qui stipulait que l'objectif ultime de la biologie moléculaire serait d'atteindre une compréhension exhaustive en visant désormais la création de nouvelles entités.

1.6.1 Élargir le champ d'application des modèles au-delà du métabolisme

1.6.1.1 Modélisation de l'expression des gènes

L'utilisation de la contrainte imposée par la stoechiométrie des réactions a été essentielle pour le développement du FBA (Kauffman et al., 2003) et, ultérieurement, pour le développement des modèles métaboliques à l'échelle du génome. Dans une tentative d'étendre la portée des modèles au-delà du métabolisme, Thiele *et al.* ont reconstruit la matrice d'expression pour *E. coli* (Thiele et al., 2009). La reconstruction de cette matrice, appelée matrice E pour expression par opposition à la matrice M pour le métabolisme, a été réalisée en utilisant le même protocole que celui mentionné ci-dessus (Thiele and Palsson, 2010). Toutes les réactions nécessaires à la transcription de l'ARN et à la traduction des protéines sont incluses dans la matrice E. Il est intéressant de noter que chaque élément nécessaire à la synthèse des protéines est considéré comme un métabolite dans le réseau. Par exemple, l'ARN de transfert (ARNt) et l'ARN ribosomique (ARNr) sont tous deux des métabolites qui peuvent être produits à partir des réactions de transcription. Les ARNt sont ensuite chargés et utilisés dans une autre réaction qui synthétise les protéines. Alors que le nombre de gènes inclus dans la matrice E (423 gènes) était inférieur à celui de la matrice M (1515 gènes (Monk et al., 2017)), le nombre de réactions est nettement supérieur (13 694 réactions dans la matrice E contre 2719 réactions dans la matrice M). La grande taille de la matrice E est due au nombre élevé de réactions similaires catalysées par les mécanismes d'expressions des gènes.

Tout comme la matrice M, la stœchiométrie imposée par la matrice E peut être utilisée comme une contrainte et la reconstruction peut être convertie en un format mathématique en appliquant des limites au flux des réactions et en fixant un objectif. Dans ce cas, les taux d'absorption des acides aminés et des nucléotides doivent être fixés car ce sont les métabolites nécessaires à la production de tous les métabolites en aval. La production de ribosomes par le modèle peut alors être optimisée pour différents taux de croissance puisque la production de ribosomes est

essentielle à la croissance des cellules. Dans cette étude, le raffinement des contraintes a permis au modèle de générer un nombre de ribosomes correspondant aux données expérimentales. Ces travaux ont démontré l'applicabilité de l'analyse de la structure des ribosomes à des systèmes autres que le métabolisme.

Afin de coupler la machinerie de l'expression des gènes au métabolisme de la cellule et de générer un modèle unifié pour la croissance cellulaire, des contraintes supplémentaires ont été nécessaires. Appelées "contraintes de couplage", ces équations sont fonction du temps de doublement de l'organisme et tiennent compte de la dilution des métabolites dans les cellules qui se divisent et fournissent aussi des limites supérieures à l'expression enzymatique (Lerman et al., 2012; Lloyd et al., 2018; O'brien et al., 2013; Thiele et al., 2012). Ces nouvelles contraintes sont à la fois entières et linéaires et définissent donc un problème de programmation linéaire avec nombres entiers (MILP). Ce type de problème est plus intense en termes de calcul que le problème de programmation linéaire ordinaire résolu en FBA et nécessite également des solveurs plus spécifiques (Yang et al., 2016).

1.6.1.2 Simulation avec des modèles ME

Un modèle ME relie le métabolisme à l'expression des gènes et peut être utilisé pour générer des prévisions testables expérimentalement telles que : le taux de croissance, les taux d'absorption et de sécrétion des substrats, les flux métaboliques et les niveaux d'expression des produits géniques (O'brien et al., 2013). Cette dernière propriété est importante car elle simplifie la comparaison avec les niveaux expérimentaux d'expression des gènes, qui peuvent maintenant être générés de façon routinière dans plusieurs conditions différentes. La facilité d'intégration de multiples données omiques dans les modèles ME a permis d'identifier des régularités biologiques clés (Ebrahim et al., 2016). Les données protéomiques expérimentales peuvent fournir le nombre absolu de protéines dans une cellule, qui peuvent être utilisés pour limiter la quantité de protéines dans le modèle ME. Les données de fluxomique peuvent également être utilisées comme une contrainte puisqu'elles fournissent le flux à travers un

certain nombre de réactions. La combinaison de ces deux types de données dans les simulations du modèle ME a permis de générer des taux de rotation (k_{eff}) pour les enzymes du modèle, un exemple de génération de connaissances guidée par le modèle.

En simulant des modèles ME dans 333 conditions environnementales différentes, Yang *et al.* ont identifié des gènes systématiquement essentiels pour une croissance optimale de *E. coli* (Yang et al., 2015b). Dans cette étude, le protéome commun exprimé dans toutes les conditions a été prédit par le modèle. Ce noyau commun du protéome était consistant avec l'expression différentielle de gènes pour l'ensemble de ces conditions. Aussi, il est possible de regrouper les gènes d'un ensemble minimal déterminé par le modèle ME en catégories fonctionnelles afin d'obtenir une représentation plus globale des fonctions conservées. À cette fin, les auteurs ont utilisé les catégories COG (Koonin et al., 2004). Ce regroupement en catégories fonctionnelles a mis en lumière le fait que les modèles ME n'incluent pas les mécanismes de réparation et de réplication de l'ADN. Leur absence dans le noyau commun justifie de poursuivre l'extension des modèles afin d'inclure un plus grand nombre de systèmes cellulaires, et permettant aussi d'augmenter leur potentiel prédictif. Ainsi, en incluant des systèmes tels que celui de la régulation, tout en conservant une approche de modélisation basée sur les contraintes, il serait possible d'obtenir une approximation fonctionnelle d'un modèele de cellule entièere nécessitant moins de paramètres expérimentaux que celui qui a été généré précédemment pour *M. genitalium* (Karr et al., 2012).

Finalement, en raison de la taille de la matrice E, la reconstruction de modèles ME entiers n'est présentement limitée qu'à deux espèces, soit *Thermotoga maritima* et *E. coli* (Lerman et al., 2012; O'brien et al., 2013). Tout comme la génération de modèles M est facilitée par l'existence de boîtes à outils, la reconstruction de modèles ME pourrait être généralisée grâce à la récente publication de COBRAme, un cadre Python pour la reconstruction de modèles ME (Lloyd et al., 2018).

1.6.2 Perspectives sur l'utilisation de modèles pour concevoir des cellules minimales

Jusqu'à maintenant, l'évolution historique de la biologie a été abordée et la possibilité qu'une partie de cette discipline se transforme progressivement en un domaine d'ingénierie, dans lequel le concept de cellule minimale jouerait un rôle central a été souligné. L'idée principale qui entoure ce concept de cellule minimale est celle du réductionnisme biologique (Glass et al., 2017), qui implique la description complète de toutes les fonctions moléculaires hébergées par une cellule vivante libre (Lachance et al., 2019a). Atteindre ce niveau de connaissance est d'une importance capitale pour l'établissement de règles de conception pour la synthèse d'organismes entiers. Avec l'avènement des techniques de synthèse de l'ADN et de l'assemblage du génome entier, la création d'organismes entièrement nouveaux est à portée de main. Un tel exemple a été obtenu avec JCVI-syn3.0 (Hutchison et al., 2016), la première approximation fonctionnelle d'une cellule minimale *in vitro*.

JCVI-syn3.0 révèle l'état de l'art pour la conception des cellules minimales. Des méthodes de pointe ainsi que des travaux approfondis sur de nombreuses années ont été mis en place afin de produire ce châssis cellulaire. La quantité de travail nécessaire pour arriver à un tel résultat est rendue possible par les méthodes de synthèse, de clonage ou d'assemblage d'ADN à haut débit développées au cours des 20 dernières années. Malgré ces avancées significatives, la capacité de prédire si une conception sera ou non viable demeure un facteur limitant en biologie synthétique. Ce défi de taille, qui concerne à la fois les chercheurs universitaires et industriels, est l'un des grands enjeux de la biologie synthétique et il est entendu que les laboratoires qui posséderont le meilleur pouvoir de prédiction pourront dépasser ceux qui disposent de capacités de production et d'analyse à haut débit.

Dans ce contexte, le développement de modèles pour les cellules minimales est d'une importance capitale. J'ai passé en revue l'approche standard FBA pour la modélisation du métabolisme à l'échelle du génome (Figures 1.3, 1.4 et 1.5) et ses applications pour l'intégration de données à haut débit et la formulation de prédictions phénotypiques telles que le flux au

travers des réactions métaboliques et l'essentialité des gènes (Figure 1.6) (Suthers et al., 2009a; Zomorrodi and Maranas, 2010). L'intégration de ces connaissances dans un cadre unique est importante pour offrir un moyen systématique de combler les lacunes en matière de connaissances (Orth and Palsson, 2010; Pan and Reed, 2018), comme l'ont démontré Breuer *et al*. dans leur GEM de JCVI-syn3.0 (Breuer et al., 2019).

Il y a lieu de débattre de ce qui nous attend. La poursuite du développement de modèles pour les mollicutes nécessitera une définition plus exhaustive de la biomasse et des milieux de croissance afin d'imposer des contraintes pertinentes au système. Compte tenu de la petite taille de leurs génomes, le nombre d'études biochimiques nécessaires pour parvenir à une caractérisation exhaustive est réduit et, à l'aide de modèles, pourrait être traité assez rapidement (Danchin and Fang, 2016). La méthode algorithmique récemment développée par Wang et Maranas permet de générer une séquence minimale du génome à partir de l'architecture transcriptionnelle et d'un modèle M (Wang and Maranas, 2018) qui pourrait aider à réduire les génomes d'organismes plus élaborés qui sont déjà utilisés comme souches de production tel que *E. coli* et *S. cerevisiae*. Comme il vient d'être présenté, les approches basées sur les contraintes peuvent être étendues au-delà du métabolisme, permettant la génération de modèles du métabolisme et de l'expression, les modèles ME. Ces modèles ont déjà été utilisés pour générer une prédiction *in silico* du protéome commun dans une simulation d'un large éventail d'environnements (Yang et al., 2015b). L'une des principales conclusions de cette étude étant que l'inclusion d'un plus grand nombre de systèmes cellulaires est importante pour des prédictions précises d'un ensemble minimal de gènes, il est intéressant de considérer que l'extension des méthodes de modélisation au-delà du métabolisme et de l'expression peut être la clé de la conception rationnelle de cellules minimales.

Enfin, l'écriture *in silico* de génomes fonctionnels devrait être l'étape suivante. L'intégration d'outils logiciels pour la conception de génomes est en cours avec les publications récentes d'un "Autocad" pour génome (Bates et al., 2017) ainsi qu'un compilateur de circuits génétiques (Waites et al., 2018). Ces outils s'inspirent de l'expérience acquise dans le domaine de

l'ingénierie, et l'intérêt suscité par la communauté suggère une application généralisée pour l'avenir de la biologie. Pour l'instant, aucun organisme n'est entièrement caractérisé et, par conséquent, l'exhaustivité proposée de la biologie (Morowitz, 1984) n'est pas encore atteinte. L'utilisation de modèles à l'échelle du génome et d'outils d'écriture du génome pourrait accélérer ce processus, et une fois qu'un châssis cellulaire minimal bien compris sera décrit, un changement de paradigme au niveau de la conception des souches s'opérera.

1.7 HYPOTHÈSES ET OBJECTIFS DU PROJET DE RECHERCHE

Jusqu'à maintenant, une révision de l'histoire de la biologie a permis de comprendre le changement de paradigme que la biologie synthétique propose par rapport à la biologie classique. Cette discipline vise en effet à créer de nouvelles entités biologiques de toute pièce. Pour arriver à ses fins, la biologie synthétique requiert un haut niveau de connaissance biologique, ce qu'on appelle une compréhension exhaustive de la cellule. Cependant,j'ai aussi énoncé que, même si elle était entièrement connue, la grandeur de la complexité biologique serait écrasante. En effet, le nombre de molécules différentes présentes dans une cellule est très élevé et le nombre d'interactions entre chacune d'elles est d'autant plus grand.

Pour surmonter ce défi de conception substantiel, il convient d'utiliser des approches simplifiant le système. D'abord, pour réduire cette complexité inhibitrice, il a été proposé de travailler avec des cellules contenant le moins de gènes possible (Morowitz, 1984), c'est-à-dire des cellules minimales (Glass et al., 2017). Comme je l'ai mentionné, une cellule minimale peut être obtenue par une réduction de la taille de son génome qui peut être naturelle ou artificielle.

Une réduction artificielle est produite en laboratoire et peut être obtenue en enlevant itérativement des portions du génome bactérien (Hirokawa et al., 2013; Reuß et al., 2017), ou encore, en assemblant un génome de toutes pièces par synthèse d'ADN (Gibson et al., 2010). D'un autre côté, les phénomènes de sélection naturelle qui sous-tendent la réduction de génome

sont complexes et moins bien compris (Batut et al., 2014). Parmi les génomes séquencés jusqu'à maintenant, les organismes vivant de manière autonome possédant les plus petits génomes appartiennent au groupe phylogénétique des mollicutes (Morowitz and Tourtellotte, 1962).

1.7.1 *Mesoplasma florum*, un candidat idéal

Parmi les mollicutes, on distingue la bactérie *Mesoplasma florum*. La première mention de cet organisme dans la littérature est survenue suite à son isolation d'une feuille de citronnier (McCoy et al., 1984). Les expériences de croissance originales ont montré qu'il était possible de cultiver cet organisme dans un milieu sans sérum ou sans cholestérol en supplémentant avec une faible concentration de détergent (Tween 80 à 0.04%) (Tully, 1983) et cette capacité lui a valu l'attribution du nom *Acholeplasma florum*. Bien que l'isolation ait été faite sur une plante, plusieurs études subséquentes ont montré que cet organisme et des parents proches (*Acholeplasma entomophilum* et *Acholeplasma seiffertii*) se retrouvaient fréquemment dans le tube digestif ou encore l'hémolymphe d'insectes (Clark et al., 1986). Des études phylogénétiques ont toutefois démontré plusieurs différences significatives entre *Acholeplasma florum* et d'autres bactéries du groupe *Acholeplasma* (Figure 1.7). Des exemples de différence incluent: l'utilisation du codon UGA comme codon tryptophane chez *A. florum* alors qu'il code pour un arrêt chez les *Acholeplasma*, ou encore la taille du génome variant entre 800 et 1000 kbp chez *A. florum* alors qu'elle est supérieure à 1000 kbp chez les *Acheloplasma*. Ces différences ont valu à *Acholeplasma florum* d'être renommée *Mesoplasma florum* (Tully et al., 1993).

Bien qu'isolée au début des années 1980, la souche *M. florum* L1 n'a été complètement séquencée qu'en 2004 (https://www.patricbrc.org/view/Genome/265311.5). L'assemblage et l'annotation du génome entier révèle une séquence de 793 224 paires de bases, contenant entre 682 (RefSeq) et 685 (PATRIC) séquences codant pour des protéines, ainsi que 2 loci encodant chacun les ARN ribosomaux. Une étude récente menée par notre laboratoire a analysé les

séquences de 13 différents isolats de *Mesoplasma florum* (Baby et al., 2018a). Cette approche a permis de déterminer que le noyau de gènes conservés entre ces souches représente environ 80% de la somme des gènes identifiés dans ces séquences. En combinant l'information obtenue par génomique comparative avec l'insertion de transposons à l'échelle du génome (essentialité), il a été possible de formuler différents scénarios de réduction de génome pour *M. florum*. Chez la souche L1, 290 gènes ont été catégorisés essentiels, 585 sont des gènes du noyau partagés avec d'autres souches et 409 gènes ont un homologue chez JCVI-syn3.0 (Hutchison et al., 2016). Fait intéressant, la presque totalité des gènes partagés avec JCVI-syn3.0 font aussi partie noyau de gènes partagés entre les souches (404/409). Cette grande similarité entre JCVI-syn3.0 et *M. florum* devrait permettre une comparaison efficace entre les prédictions de génomes minimaux formulées à l'aide d'un modèle *in silico* et la réalité expérimentale.

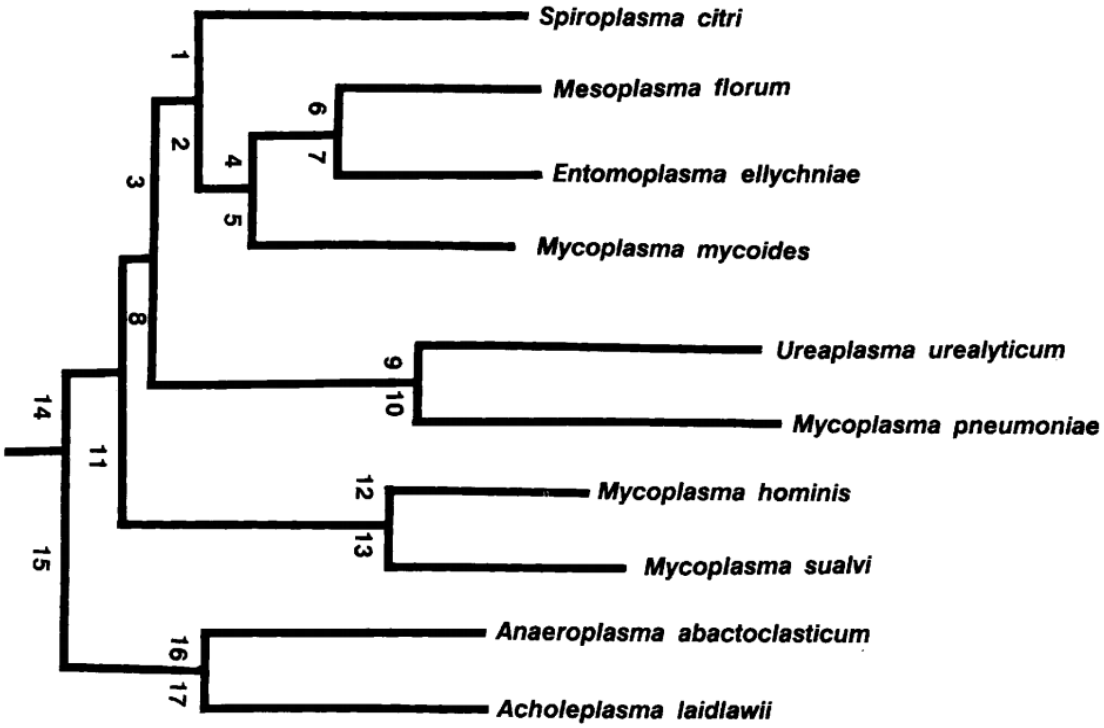

Figure 1.7. Arbre phylogénétique d'espèces clés des mollicutes. Suite à l'évaluation de ses capacités métaboliques, le nom de *Acholeplasma florum* a été changé pour *Mesoplasma*

florum. Le séquençage d'ARN ribosomal de quelque 47 espèces de mollicutes a permis de tracer l'arbre phylogénétique de ce groupe de bactéries (Reproduit de Tully et al., 1993).

Un organisme pour lequel des outils de génie génétique efficaces sont disponibles et qui comporte aussi des outils de biologie des systèmes résumant l'ensemble de la connaissance disponible se nomme un châssis cellulaire pour la biologie synthétique (Danchin, 2012). Idéalement, un tel châssis permettrait de concevoir des modifications au génome et la prédiction du phénotype résultant des modifications seraient systématiquement exactes. Pour atteindre ce niveau avec *M. florum*, il faut d'abord développer certains outils de modification génétiques. Chez les mollicutes, la combinaison d'une faible taille de génome qui contiendrait moins de mécanismes de réparation de l'ADN et d'un moins grand nombre d'études dédiés à ce sujet résulte en des modifications génétiques plus ardues que chez des organismes modèles bien étudiés comme *E. coli* et *S. cerevisiae*. Cependant, de récents efforts en provenance de notre laboratoire ont permis de distinguer *M. florum* des autres mollicutes au niveau de la capacité de modifier son génome.

Le plasmide, un fragment d'ADN circulaire auto-réplicatif, est un outil important de modification génétique dont l'intégration dans une cellule est généralement sélectionnée par une résistance aux antibiotiques. Des plasmides basés sur l'origine de réplication du chromosome de *M. florum* (*oriC*) ont été développés en testant l'efficacité de différentes boîtes de liaison du gène *dnaA* (Matteau et al., 2017). La sélection de ces plasmides a été assurée en déterminant la sensibilité de *M. florum* à divers antibiotiques et une résistance à la spectinomycine ou à la streptomycine a été ajoutée pour garantir une sélection efficace des transformants. Cet outil moléculaire important ouvre la voie à l'expression de protéines hétérologues chez *M. florum*.

Une avancée importante de la biologie synthétique et un outil de manipulation génétique disponible uniquement chez les mollicutes est la transplantation de génomes entiers (Lartigue et al., 2007). Au cours de cette manipulation, un génome bactérien est isolé d'une espèce bactérienne, puis transplanté dans le cytoplasme d'une espèce réceptrice. Suivant l'activation

de l'expression des gènes du génome transplanté, la bactérie réceptrice adopte les phénotypes du nouveau génome. Un marqueur de sélection permet de s'assurer qu'il ne reste que le génome synthétique dans les bactéries réceptrices qui ont alors littéralement changé d'espèce. Ce type de transfert peut aussi s'effectuer d'une espèce de mollicutes vers la levure *S. cerevisiae (Gibson et al., 2008).* Une telle approche a été développée pour *M. florum* de sorte à ce qu'il soit désormais possible d'effectuer des modifications sur son génome en utilisant la pléthore d'outils de modifications génétiques disponible chez *S. cerevisiae* (Baby et al., 2018b). Le retour de ce génome dans un mollicute est aussi possible (Labroussaa et al., 2016), permettant ainsi de créer une souche modifiée de *M. florum*.

Récemment, une caractérisation intégrative très détaillée de la cellule de *M. florum* a été publiée par notre laboratoire (Matteau et al., 2020). Cette étude a permis de déterminer avec précision plusieurs paramètres physiques, chimiques et biologiques de la cellule incluant son volume, son poids, ses paramètres de croissance ainsi que sa composition en biomasse. Des données omiques à haut-débit ont aussi été produites permettant de saisir quels ARN et quelles protéines sont exprimés dans un milieu riche.

1.7.2 Hypothèses et objectifs

Considérant l'avancement du développement d'outils moléculaires pour les modifications génétiques ainsi que la quantité d'information disponible pour M. florum provenant de sa caractérisation intégrative, je pose l'hypothèse que le développement d'outils de biologie des systèmes pour cet organisme est une étape essentielle afin qu'il devienne un châssis cellulaire pour la biologie synthétique. Comme la fonction objective de biomasse (BOF) d'un modèle métabolique est critique afin de déterminer correctement l'essentialité des gènes, un objectif préalable à celui de la reconstruction d'un modèle est de:

1. ***générer un logiciel permettant la définition de la fonction objective de biomasse dans les modèles métaboliques à l'échelle du génome.***

Ce qui précède l'objectif principal de ma thèse de doctorat, soit de :

2. ***réaliser un modèle métabolique à l'échelle du génome pour M. florum***

3. ***et de l'utiliser pour formuler une prédiction de génome minimal.***

CHAPITRE 2
BOFdat: GENERATING BIOMASS OBJECTIVE FUNCTIONS FOR GENOME-SCALE METABOLIC MODELS FROM EXPERIMENTAL DATA

2.1 CONTEXTE

Les modèles métaboliques à l'échelle du génome (GEM) sont largement utilisés pour générer des prédictions phénotypiques à partir du génome et possèdent des applications très variées, allant de la découverte fondamentale au génie métabolique (Monk et al., 2014). Pour calculer les phénotypes de croissance avec un GEM, il faut d'abord définir une fonction objective de biomasse (BOF) qui englobe tous les principaux composants de la cellule. Cette procédure peut se faire en utilisant des données tirées de mesures expérimentales et de la littérature (Feist and Palsson, 2010; Thiele and Palsson, 2010). La BOF comprend les macromolécules majeures telles que l'ADN, l'ARN, les protéines et les lipides, ainsi que les coûts de maintenance énergétique des cellules. Paramétrer la formulation d'une BOF avec ces éléments permet de calculer plusieurs phénotypes de croissance sur différents milieux de culture tels que le taux de croissance, les besoins en nutriments et le potentiel biosynthétique. L'ajout de coenzymes, d'ions inorganiques ainsi que de métabolites spécifiques à l'espèce (c'est-à-dire les composants de la paroi cellulaire) augmente la qualité de la prédiction de l'essentialité des gènes à partir d'un modèle. Xavier *et. al*. (Xavier et al., 2017) ont récemment tenté d'organiser le contenu de la BOF de plusieurs modèles métaboliques existants pour comprendre le lien entre la relation phylogénétique des espèces et la composition de la BOF. Si leurs travaux ont permis d'obtenir des informations sur la composition qualitative de la BOF existante, ils ont également constaté qu'il n'existe aucune méthode informatique structurée pour sa définition. Dikicioglu *et. al.*

(Dikicioglu et al., 2015) ont étudié l'importance de la définition quantitative de la BOF et ont montré son impact sur le comportement et les prédictions du modèle. Ensemble, ces études soulignent l'importance du développement d'une plateforme informatique pour la définition de fonctions objectives basées sur des données spécifiques à l'espèce.

Nous proposons ici un logiciel Python pour la définition complète de la BOF spécifique à un organisme à partir de données expérimentales : BOFdat. Ce logiciel intègre une procédure systématique en trois étapes qui permet d'effectuer la définition à la fois qualitative et quantitative de la BOF. Les outils développés précédemment fournissent aux utilisateurs une manière biaisée ou non biaisée de définir la BOF. La plateforme SEED (Devoid et al., 2013) automatise le processus de reconstruction et permet à l'utilisateur de décider de la fonction objective à partir d'une liste de BOF définie selon les groupes phylogénétiques (par exemple, une BOF pour toutes les bactéries gram-négatives). Des approches impartiales comme ObjFind (Burgard and Maranas, 2003), BOSS (Gianchandani et al., 2008) et invFBA (Zhao et al., 2016) ont utilisé la fluxomique pour prédire les objectifs métaboliques finaux de la cellule. Cette dernière approche peut être limitée par la disponibilité des données de fluxomique qui sont rarement générées et le petit ensemble de flux *in vitro* qui peuvent être obtenus au cours d'une expérience. BOFdat tire plutôt profit des multiples ensembles de données omiques (génomique, transcriptomique, protéomique et lipidomique) générés en routine ainsi que des données sur l'essentialité des gènes pour définir pleinement la BOF.

2.2 CONTRIBUTION DES AUTEURS

Ce chapitre présente un article publié grâce à une collaboration entre les laboratoires des Pr Pierre-Étienne Jacques et Pr Sébastien Rodrigue à l'Université de Sherbrooke et des Pr Bernhard O. Palsson et Adam M. Feist à l'Université de Californie à San Diego où j'ai été accueilli en tant que visiteur pendant une bonne partie de mon doctorat. Au cours de ma visite, j'ai joint le sous-groupe du Dr Jonathan M. Monk pour des rencontres hebdomadaires qui ont

permis d'identifier une lacune dans l'écosystème des logiciels utilisés pour faire la reconstruction métabolique. J'ai donc entrepris de concevoir un logiciel permettant de reconstruire la BOF à partir de données expérimentales. La supervision du Dr Zachary King a permis de guider la rédaction du code informatique afin d'être conforme aux bonnes pratiques d'écriture et de distribution des logiciels. Afin de valider l'utilisation du logiciel, le manuscrit présente une reconstruction de la BOF pour le modèle métabolique de la bactérie *Escherichia coli* (*i*ML1515). L'expertise des Dr Colton J. Lloyd et Jonathan M. Monk, co-premiers auteurs du modèle *i*ML1515 utilisé pour la validation, a été fort utile afin de bien comprendre les éléments importants de la biomasse de *E. coli*. Les conversations avec le Dr Anand D. Sastry ont permis de mieux formuler l'analyse des métabolites suggérés par l'algorithme génétique de la troisième étape de BOFdat, notamment en incluant un algorithme de regroupement (DBSCAN). L'expertise du Pr Laurence Yang a permis de valider le choix des métabolites suggérés par BOFdat contre les choix formulés par l'algorithme BOSS. Lorsque le paquet a atteint un stade assez avancé, la Dr Yara Seif a pu le tester dans la reconstruction de son modèle métabolique pour *Staphylococcus Aureus*. En plus d'avoir supervisé l'ensemble du projet, la contribution du Pr Pierre-Étienne Jacques a été monumentale au moment de la rédaction, de la correction et de la soumission du manuscrit. Le Pr Sébastien Rodrigue a aussi contribué à la révision et l'édition du manuscrit.

Référence bibliographique: Lachance, J.-C., Lloyd, C.J., Monk, J.M., Yang, L., Sastry, A.V., Seif, Y., Palsson, B.O., Rodrigue, S., Feist, A.M., King, Z.A., Jacques, P.-E. BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. PLOS Computational Biology vol. 15 e1006971 (2019).

2.3 TITLE PAGE

BOFdat: generating biomass objective functions for genome-scale metabolic models from experimental data

Jean-Christophe Lachance[1], Colton J. Lloyd[2], Jonathan M. Monk[2], Laurence Yang[2], Anand V Sastry[2], Yara Seif[2], Bernhard O. Palsson[2,3,4,5], Sébastien Rodrigue[1], Adam M. Feist[2,5], Zachary A. King[2*], Pierre-Étienne Jacques[1*]

[1] Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada

[2] Department of Bioengineering, University of California, San Diego, La Jolla, USA

[3] Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, USA

[4] Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA

[5] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Lyngby, Denmark

* Corresponding authors

E-mail: pierre-etienne.jacques@usherbrooke.ca (PEJ)

E-mail: zaking@ucsd.edu (ZAK)

2.4 ABSTRACT

Genome-scale metabolic models (GEMs) are mathematically structured knowledge bases of metabolism that provide phenotypic predictions from genomic information. GEM-guided predictions of growth phenotypes rely on the accurate definition of a biomass objective function (BOF) that is designed to include key cellular biomass components such as the major macromolecules (DNA, RNA, proteins), lipids, coenzymes, inorganic ions and species-specific components. Despite its importance, no standardized computational platform is currently available to generate species-specific biomass objective functions in a data-driven, unbiased fashion. To fill this gap in the metabolic modeling software ecosystem, we implemented BOFdat, a Python package for the definition of a Biomass Objective Function from experimental data. BOFdat has a modular implementation that divides the BOF definition process into three independent modules defined here as steps: 1) the coefficients for major macromolecules are calculated, 2) coenzymes and inorganic ions are identified and their stoichiometric coefficients estimated, 3) the remaining species-specific metabolic biomass precursors are algorithmically extracted in an unbiased way from experimental data. We used BOFdat to reconstruct the BOF of the *Escherichia coli* model *i*ML1515, a gold standard in the field. The BOF generated by BOFdat resulted in the most concordant biomass composition, growth rate, and gene essentiality prediction accuracy when compared to other methods. Installation instructions for BOFdat are available in the documentation and the source code is available on GitHub (**https://github.com/jclachance/BOFdat**).

2.5 AUTHOR SUMMARY

The formulation of phenotypic predictions by genome-scale models (GEMs) is dependent on the specified objective. The idea of a biomass objective function (BOF) is to represent all metabolites necessary for cells to double so that optimizing the BOF is equivalent to optimizing growth. Knowledge of the qualitative and quantitative organism's composition (i.e. which

metabolites are necessary for growth and in what proportion) is critical for accurate predictions. We implemented BOFdat with the idea that experimental data should drive the definition of the biomass composition. As omic datasets become more available, the possibility of integrating them to obtain a condition-specific biomass composition is in reach and therefore one of the main features of BOFdat. While major macromolecules, coenzymes, and inorganic ions are ubiquitous components across species, several species-specific components exist in the cell that should be accounted for in the BOF. To identify these we implemented an approach that minimizes the error between experimental essentiality data and GEM-driven prediction. Hence BOFdat provides an unbiased, data-driven approach to defining BOF that has the potential to improve the quality of new genome-scale models and greatly decrease the time required to generate a new reconstruction.

2.6 INTRODUCTION

Genome-scale metabolic models (GEMs) are widely used to generate phenotypic predictions from genomic information, with wide-ranging applications from discovery to metabolic engineering [1]. To compute growth phenotypes, one must first define a biomass objective function (BOF) that encompasses all the major components of the cell, using data drawn from experimental measurements and literature [2,3]. The typical BOF primarily includes the cell's major macromolecules (DNA, RNA, protein, lipids, carbohydrates), crucial coenzymes and inorganic ions, species-specific metabolites such as cell wall components and finally growth and non-growth associated maintenance costs which represent basal energy requirements to sustain the cell and divide. The formulation of a BOF parameterized with these elements allows the computation of phenotypes on different media, such as growth rate, nutrient requirements, and biosynthetic potential. Dikicioglu *et al.* studied the importance of the quantitative definition of the BOF and showed its impact on model behavior and predictions [4]. Xavier *et al.* attempted to organize the content of the BOF of existing genome-scale models to understand the phylogenetic relationship of cellular compositions in prokaryotes [5]. While

their work yielded information on the qualitative composition of existing BOF, they also noted that no systematic computational framework exists for its definition.

Experimental measurements of biomass composition have been developed, and it is already known that the cellular composition varies upon different growth conditions. Ratios between DNA, RNA and proteins have been shown to vary depending on growth rate and nutrient availability [6]. The growth rate also affects cellular volume [7], which in turn impacts the total cell weight and the proportion of its components. Finally, Beck *et al.* recently reviewed and provided a state of the art method of determining experimental macromolecular composition, and showed that the macromolecular composition varies considerably from one species to another [8]. While these studies clearly showed diversity across species and conditions in terms of biomass composition and the impact of the BOF on model predictions, modelers often default to copy the BOF of a quality GEM rather than generating their own. The lack of a defined computational workflow facilitating the inclusion of experimental data to the BOF of GEMs may account for this behavior within the modeling community.

Previously developed tools provided modelers with either a biased or unbiased approach to define the BOF for a metabolic reconstruction. Biased approaches have the benefit of being based on current biological knowledge but are hampered by potential fluctuations in metabolic objectives over specific conditions or across species and strains. For example, the SEED platform [9] automates the reconstruction process and lets users decide on the objective function from a list of BOF defined according to phylogenetic groups, i.e. one BOF for all gram-negative bacteria. However, the unbiased computational extraction of metabolic objectives from experimental data is expected to better represent the content and objectives of the cell. Such approaches have been developed by other groups and include ObjFind [10], BOSS [11] and more recently invFBA [12]. These methods use fluxomics data to predict the metabolic end goals of the cell [13], an approach that is currently limited by the availability of this type of data, and by the small number of fluxes generated by fluxomics experiments. Here

we present **BOFdat**, a Python software package for the complete definition of organism-specific BOF from experimental data. This package embeds a systematic 3-step procedure, where each step can be performed independently, that generates both qualitative and quantitative definitions of the BOF. BOFdat takes advantage of nowadays routinely generated multiple omic datasets (genomic, transcriptomic, proteomic and lipidomic) as well as the increasingly available gene essentiality data [14] to generate a BOF specific to the organism of interest.

2.7 METHODS

2.7.1 A computational workflow for biomass definition from experimental data

In a GEM, the BOF is represented as a reaction in the stoichiometric matrix, where metabolites are consumed or produced in a proportion given by each stoichiometric coefficient to represent the net requirements for generation of cell biomass at steady state [15]. Feist and Palsson [2] previously divided the definition of the BOF for a given organism into three different levels, based on the amount of knowledge available for the organism. A basic level BOF includes major macromolecules of the cell (DNA, RNA, proteins, and lipids), an intermediate BOF provides the polymerization and maintenance costs of macromolecules and the cell in general, whereas an advanced BOF definition includes metabolites that are specific to the organism of interest (coenzymes, inorganic ions, cell wall components, etc.) [2]. BOFdat closely follows this logic by dividing the biomass definition process into 3 different steps illustrated in Fig 2.1 The modular implementation of BOFdat allows users to perform each step independently or in order, following their needs. **Step 1** aims at calculating the stoichiometric coefficients of the major macromolecules mentioned above, thereby providing a computational tool compatible with the experimental methods for the measurement of macromolecular cell composition reviewed by Beck *et al.* [8]. As mentioned below, an important input is the macromolecular weight fraction (MWF) of each category. Also included in the first step is the computation of

the growth and non-growth associated maintenance costs from growth data. **Step 2** aims at refining the BOF by adding coenzymes and inorganic ions. The stoichiometric coefficients for suitable metabolites are calculated using the MWF of the soluble pool. This MWF may vary from an organism to another and can be generated experimentally or obtained from the literature. **Step 3** aims at finding condition and species-specific metabolic end goals. To do so, an unbiased approach based on experimental gene essentiality data uses the power of the metaheuristic genetic algorithm (GA) and spatial clustering to identify groups of metabolites that represent cellular objectives under the condition of interest.

Overall, BOFdat is designed to enable modelers to operate each step independently rather than being restrained to a single "do-it-all" function. This design is based on the reality of metabolic modeling where the availability of experimental data is variable and may increase over time. We provide detailed data requirements of each step of the workflow in the "General Usage" section of the documentation (https://bofdat.readthedocs.io/). The characterization of the MWF of the major macromolecules is the most important factor to determine the stoichiometric coefficients and should be prioritized in experimental design. The use of genomic, transcriptomic, proteomic and lipidomic data refines the stoichiometric coefficient under a given condition (Fig. S2.7), with lipidomic data also informing on the lipid species that should be added to the BOF. To generate accurate growth rate predictions modelers should obtain experimental uptake and secretion rates of major carbon sources and metabolic waste. Lastly, if the modeler's goal is to maximize the gene essentiality prediction, BOFdat Step 3 can be used along with experimental genome-wide essentiality data to find the metabolites that will optimize this phenotypic prediction.

While strictly following the workflow is not mandatory, this facilitates the mass-balancing of the objective function. Indeed, each addition of a macromolecular category requires the input of its MWF. A fundamental concept for the prediction of growth rate using flux balance analysis is the establishment of a basis such that the product of cell weight by time is equal to

1 gDW/hr [16]. A careful use of BOFdat would ensure that the sum of all MWF is equal to 1, hereby ensuring that the basis for the prediction of growth rate is respected. A tool for the verification of this assumption was previously published and is therefore not included in BOFdat [17].

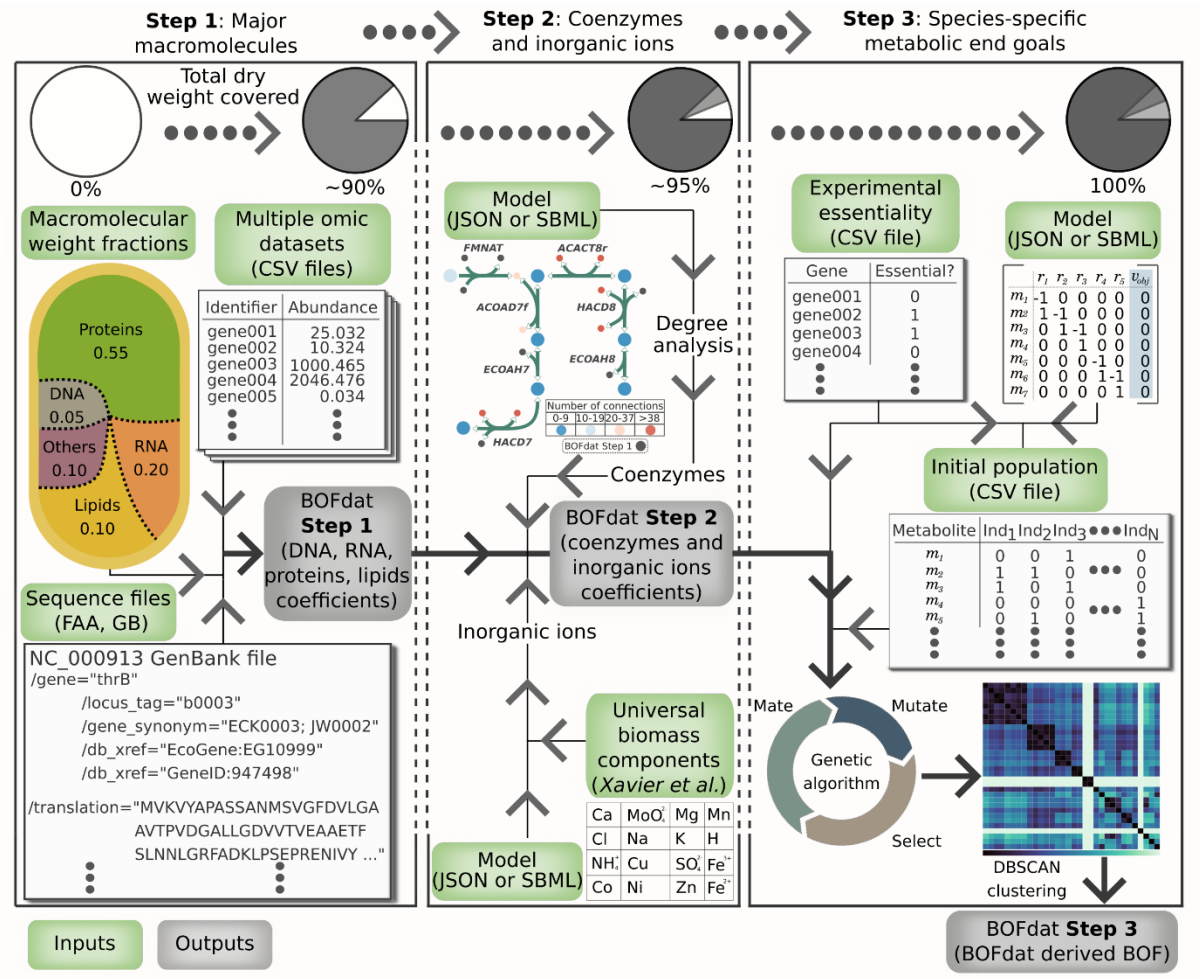

Figure 2.1. The three-step workflow for generating biomass objective functions from experimental data with BOFdat. Each step is presented in a rectangular frame in which input, and output files are shown using green or grey boxes, respectively. The modular implementation of BOFdat allows performing each step sequentially or independently, i.e. Step 3 can be used by itself to improve the gene essentiality prediction of an existing BOF. When the sequence of the workflow is observed, the output biomass function from Step 1 and 2 is the input for the subsequent step. Following the light arrows leads from the input to the output

of each Step. The thicker arrows present the normal workflow for BOFdat leading to the final output of Step 3.

Finally, while the first step of BOFdat does not require a complete metabolic network, results from the following Step 2 and 3 may be affected by the completeness of the network. We suggest using available tools to fill gaps in metabolic networks [18–22] prior to performing these steps of BOFdat.

2.7.2 Step 1: Determining macromolecular composition and maintenance costs

In most cells, major macromolecules (DNA, RNA, proteins and lipids) occupy a significant fraction of the total mass [8] (Fig. 2.2A). Quantifying these molecules is thus crucial for a GEM to generate accurate phenotypic predictions. The first step of BOFdat specifically aims at determining the stoichiometric coefficients for each metabolite building block composing these macromolecules. For each molecular category, a BOFdat function requires the input from experimental data (Fig 2.2). An exhaustive description of the required files and file formats can be found in the docstring of the package and in the documentation (https://bofdat.readthedocs.io/). Implementation and calculation details are provided in S1 Text.

The DNA coefficients are calculated using the genome sequence, which is used to determine the relative abundance of each nucleotide (dATP, dTTP, dCTP, dGTP). The MWF of DNA is provided by the user as a fraction between 0 and 1. The molar weight of each nucleotide is extracted from the model formula weight. Together this information is sufficient to calculate the stoichiometric coefficient (expressed in mmol/gDW/hr) of each metabolite [3] (Fig 2.2C).

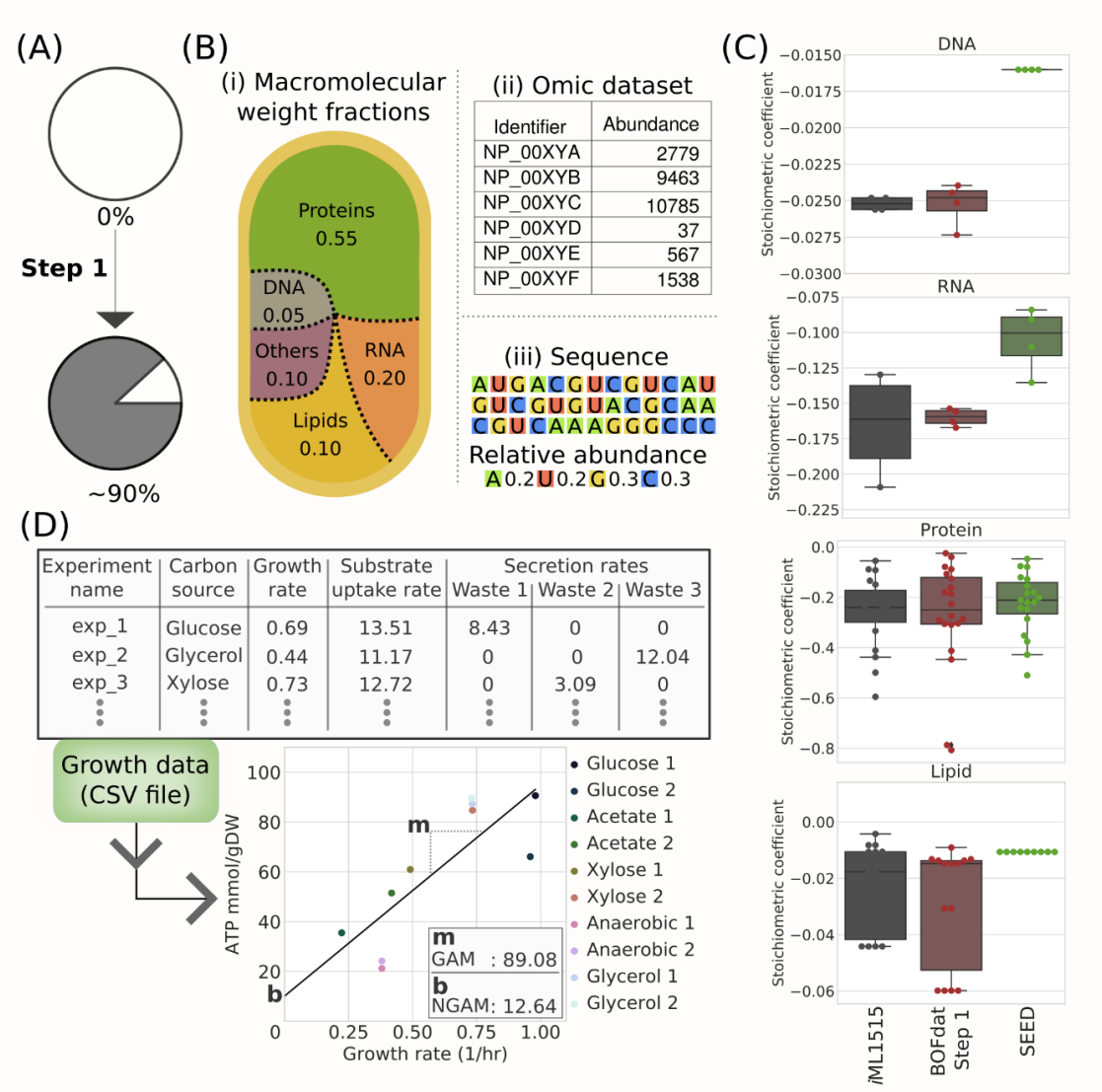

Figure 2.2 BOFdat Step 1: Calculating the biomass objective function stoichiometric coefficients (BOFsc) for the 4 principal macromolecular categories of the cell. (A) Impact of BOFdat Step 1 on the description of the cellular dry weight. (B) BOFsc are calculated using three data types: i) Macromolecular weight fractions, ii) Omic datasets, and iii) the genome sequence. (C) Comparison of the stoichiometric coefficients used in *i*ML1515 (grey) with those generated by BOFdat (red) and SEED (green). (D) Experimentally measured growth rate, substrate uptake rate, and metabolic waste secretion rate across different conditions are used to constrain the model and generate growth-associated (GAM) and non-growth associated (NGAM) ATP maintenance costs. The GAM is represented by the slope (m) of the linear regression over the conditions, while the NGAM is the Y-intercept (b) of that slope.

The coefficients of the monomers that compose RNA and proteins are obtained by weighting each metabolite frequency within a sequence (RNA transcript or translated protein) by the relative abundance of this transcript/protein in the corresponding omic dataset. This allows the user to calculate the relative abundance of each ribonucleotide and amino acid (AA) metabolites in the entire cell. To calculate stoichiometric coefficients, BOFdat thus uses the sequence of each RNA and protein along with the relative abundance of each transcript or translated protein. To ensure proper calculation of the stoichiometric coefficients, modelers must provide true relative abundances (i.e. not logarithmically transformed). The whole-cell relative abundances calculated using these input files are then converted into stoichiometric coefficients using the same calculation method as for the DNA metabolites (Fig 2C).

The lipidic constituents of the cell can also be incorporated into the BOF with BOFdat. To do so, modelers require access to a lipidomic dataset. The first input to provide is a conversion between BiGG identifier and the lipid generic name. This conversion step is intended to encourage a lipid network definition that matches experimental data. The second input file contains the relative abundances of each lipid species identified. Once provided with these two elements along with the lipidic MWF, BOFdat applies the conversion to BiGG and finds the molar weight of each lipid to compute their stoichiometric coefficients (Fig 2C). Lipid species may have varying tail length, which may impact their molar weight. In the case where the length of the chain is unknown, BOFdat will use a default R-chain weight that can be changed by the user.

The maintenance costs calculation is also included in the first step of BOFdat. Growth and non-growth associated maintenance (GAM and NGAM) are obtained from growth rate, substrate uptake rate and secretion rate on different media conditions as performed by Monk *et al.* [23] (Fig 2.2D). While this calculation is included in Step 1, we strongly advise to use a completed (or close to completion) model before generating the maintenance costs. The input file format used to perform the calculation is described in the BOFdat documentation

(http://bofdat.readthedocs.io). Briefly, growth-associated maintenance is represented as the slope of the curve obtained by linear regression, and the non-growth associated maintenance is the y-intercept of that curve (S1 Text).

2.7.3 Step 2: Identifying coenzymes and inorganic ions

While their mass fraction may not be as significant as the major macromolecules (Fig 2.3A), coenzymes and inorganic ions composing the soluble metabolite pool are key for cell growth, allowing many enzymatic reactions to take place. Since they are renewed after utilization, coenzymes may participate in many reactions and are often described as 'currency metabolites' [24]. The second step of the BOFdat workflow utilizes this property to identify high degree metabolites from the network (Fig 2.3B), which are considered as *bona fide* coenzymes (Fig S2.1).

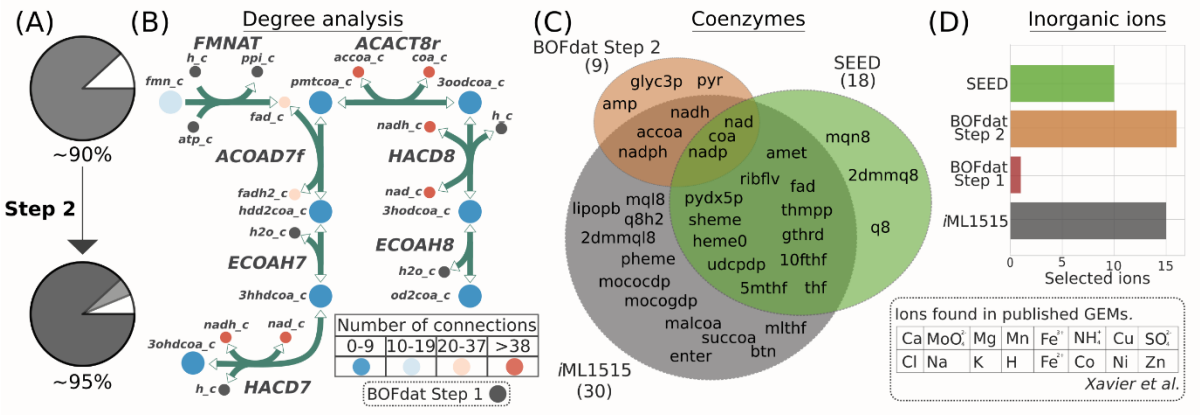

Figure 2.3 BOFdat Step 2: Identifying and calculating the stoichiometric coefficients of coenzymes and inorganic ions. (A) Pie graphs of the percent dry weight accounted for before and after BOFdat Step 2. (B) The coenzymes found by BOFdat Step 2 are metabolites with a higher degree than the established threshold (S1 Text). Shown is the degree analysis performed on a subset of 7 reactions in *i*ML1515. The metabolites are colored according to the number of reactions to which they participate in the model. Metabolites included in BOFdat Step 1 are removed from the Degree analysis (grey). (C) Venn diagram of the coenzymes found in Step 2 (orange) compared to SEED (green) and the original *i*ML1515 wild-type biomass (grey).

Manual curation was used to identify metabolites that qualify as coenzymes in both *i*ML1515 and SEED. (D) Bar chart showing the list of universal ions found by Rocha and colleagues [5] identified in each method. BOFdat Step2 finds the inorganic ions in the model by comparing the model metabolites against this list.

Xavier *et al.* [5] provided a table of biomass components used in 72 published metabolic models available on the BiGG Models database [25]. This valuable resource was integrated with the BOFdat workflow to add inorganic ions to the BOF since whole-cell experimental identification of inorganic ions is tedious. The 16 identified ions are mapped to the model (Fig 3D). Finding the ions in the metabolic reconstruction signifies their usage by the cell and is hence sufficient to add them to the BOF. To calculate the stoichiometric coefficients for these metabolites, BOFdat uses the MWF of the entire soluble pool. This value may vary from an organism to another, we therefore suggest obtaining the MWF experimentally or extracting the information from organism-specific literature. If this is not possible, BOFdat uses a default weight fraction of 0.05 (5%) of the cell for the entire soluble pool, a value that can be changed by users. The molecules that compose this category are assumed to be represented evenly, and the weight of each molecule is used to obtain the final stoichiometric coefficients, similarly to the method described above (S1 Text).

2.7.4 Step 3: Identifying organism-specific biomass precursors

The addition of the major macromolecules combined with the coenzymes and inorganic ions allow the first two steps of BOFdat to represent a high fraction of the cell weight (~95% for *E. coli*, Fig 2.4A). Using experimental gene essentiality data, BOFdat Step 3 aims at identifying condition and species-specific biomass precursors. These remaining metabolites are likely to vary from one species experimental condition to another, hence their addition via an unbiased approach ensures a context-specific composition of the BOF. BOFdat identifies these biomass precursors through multiple iterations of a genetic algorithm (GA). The implementation of the GA is based on the DEAP toolbox version 1.2 [26] as described in S1 Text.

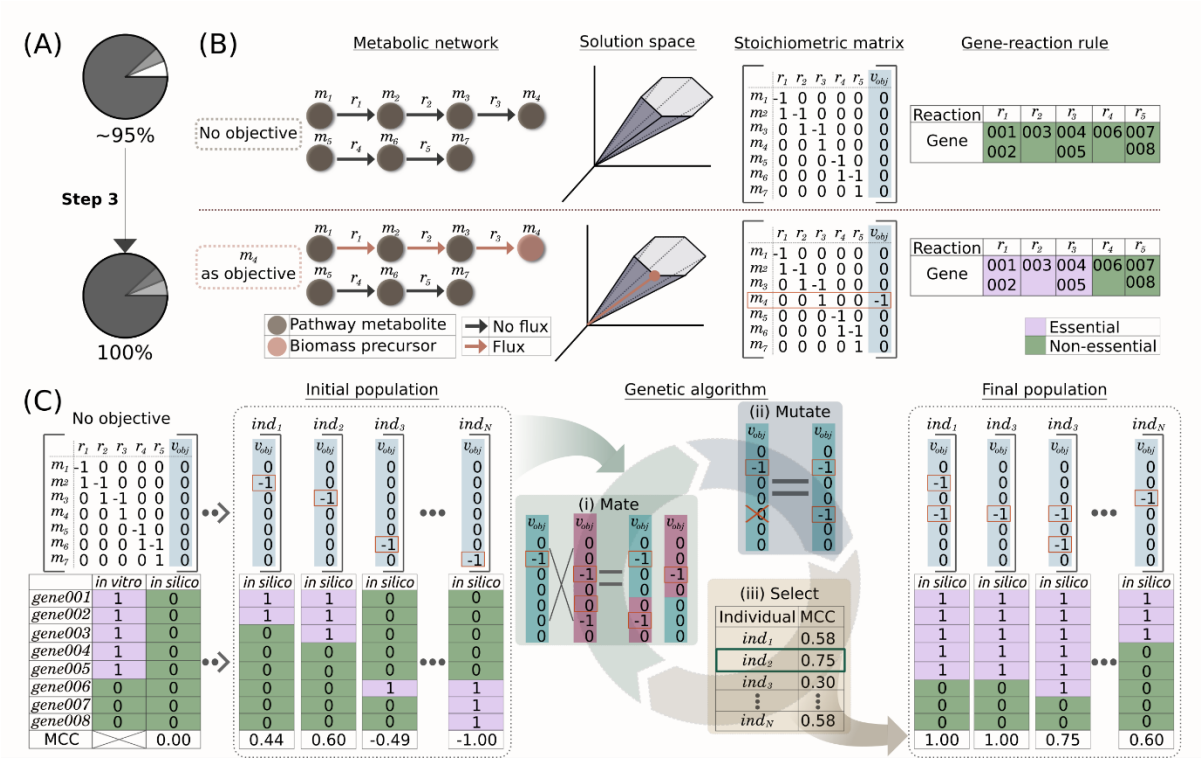

Figure 2.4 BOFdat Step 3: Identifying species-specific metabolic end goals. (A) After Step 3, the entire weight of the cell is accounted for by BOFdat. (B) Schematic description of the effect of adding a biomass precursor on the prediction of gene essentiality in the model. A simplified metabolic network composed of two linear pathways is depicted with its corresponding stoichiometric matrix S, in which the objective function is presented in the blue column (v_{obj}). The addition of the metabolite *m4* to the objective vector (orange dots and rectangle), forces the flux through reactions *r1*, *r2* and *r3* (orange arrows) and makes genes 001 to 005 computationally essential (purple boxes), defining a new line of optimality in the solution space. (C) Schematic representation of the implementation of the genetic algorithm (GA) using the metabolic network presented in B. The Matthews Correlation Coefficient (MCC) is used to compare *in vitro* (observed) and *in silico* (predicted) gene essentiality data. The MCC is calculated for each individual in the initial population. For simplicity, we represent each individual with a single biomass component. The genetic operators (mate, mutate and select) are then applied on a population to generate new individuals with higher MCC values (used here as a measure of fitness). At the end of the evolution, the final population is composed of different individuals with mainly high MCC values.

2.7.5 Concepts underlying the implementation of the GA.

The GA supposes that the addition of a metabolite to the objective function may change and improve the gene essentiality prediction of a model. As illustrated in a simplified network composed of two linear pathways (Fig 2.4B), the addition of a metabolite to the objective function sets a line of optimality on the solution space. To satisfy the new objective set by the addition of metabolite *m4* to the biomass, flux must go through reactions *r1*, *r2* and *r3*. Since *m4* can only be produced through these reactions, the model predicts them as essential along with the genes to which they are associated by the gene-protein-reaction rule (GPR). The COBRApy toolbox [27] allows users to generate model-wide single gene deletion predictions where each gene in the model is removed individually and the resulting growth is assessed by attempting to solve the model. A growth/no-growth phenotypic prediction can then be generated for every gene in the model similar to high-density transposon mutagenesis experiments [28] and other high-throughput approaches to assess gene essentiality *in vivo* [29]. For comparison purposes, the gene essentiality observations and predictions can be converted into binary vectors, enabling the use of common distance metrics for their comparison. The Matthews Correlation Coefficient (MCC) is a metric frequently used to evaluate GEMs' gene essentiality prediction, as it takes account of false and true positive and negative observations in a balanced way, and works with binary classifications [30]. Using this metric, the gene essentiality prediction resulting from a newly formulated BOF can be evaluated against experimental data, where an MCC equal to 0 would be equivalent to random whereas a MCC of 1 is an exact match between predictions and observations (Fig 2.4C). This concept allows users to define the main elements of a genetic algorithm where each newly generated BOF is defined as an individual and the MCC score can be used as a fitness metric. To ease the usability, BOFdat divides Step 3 into three different operations: 1) a group of individuals called an "initial population" is generated; 2) the GA is applied to the initial population by iteratively applying genetic operators to its individuals in a process termed *in silico* evolution, referred to simply as evolution throughout the text; and 3) the results are interpreted through spatial clustering to form metabolic end goals.

2.7.6 Definition of the initial population.

An initial population is generated, on which the evolution will be performed. Conceptually, each individual in the population may contain any combination of all metabolites in the model. In order to reduce the search space of the algorithm, BOFdat utilizes a series of feature selection operations. The metabolites from the output of BOFdat Step 2 are removed from the complete set of metabolites. Metabolites that cannot be produced individually by the model are also removed. Lastly, the impact on gene essentiality prediction for remaining metabolites is assessed by optimizing for the production of each individual metabolite and calculating the MCC score as described above (Fig S2.2). Metabolites with resulting MCC scores above a defined threshold are selected as input to Step 3 and referred to as the metabolites subset (S1 Text).

An individual is represented by a binary list referring to the presence (1) or absence (0) of the metabolites subset (Fig 2.4C). Individuals in the initial population are randomly generated to maximize diversity and coverage of the selected metabolites. It has been previously shown that the outcome of an evolution may be affected by the initial population [31]. Therefore, it is recommended to perform multiple evolutions. BOFdat hence requires the user to input the desired number of initial populations so that one evolution per generated initial population can be performed (Fig S2.3). The suggested number of evolutions to perform is discussed in the results section.

2.7.7 Application of the GA.

An evolution is performed by iteratively applying genetic operators to the initial population. The three genetic operators used in the GA are mutation, mating, and selection. Mating and mutation operators generate diversity within the population, allowing the GA to screen more combinations of BOFs. During an *in silico* mating event, crossovers can arise between two

individuals. In such case, a position in the list of two individuals is chosen and their elements are exchanged. An individual is an index of metabolites associated with an indicator of presence (1) or absence (0). This format allows to apply mutations which revert a 0 into a 1 and vice-versa. Selection is key to the GA since it ensures that a trait, referred to as fitness, is optimized throughout the evolution. In BOFdat Step 3, the fitness of each individual is measured by the MCC score, where a higher MCC score increases the chances of an individual being a member of the next generation. The process of applying the genetic operators on a population of individuals is called a generation.

The GA implementation also ensures that a limited number of metabolites is contained within individuals. To do so, the number of metabolites that a single individual can contain is restrained by applying a second objective to the fitness function used to select individuals. This maximizes the MCC score while minimizing the number of selected metabolites (Fig S2.4 and S1 Text). As mentioned in the DEAP documentation (https://deap.readthedocs.io), the output of the GA can be presented in the form of a logbook or hall of fame (HOF). The logbook records the statistics of the population fitness at each generation (mean, maximum and minimum on the individuals' fitness), allowing the user to follow the increase of fitness over generations. The HOF records the best individuals generated throughout an entire evolution and can be used to assess the metabolite content of optimal solutions.

2.7.8 Interpreting the result of multiple evolutions.

The results of multiple evolutions are interpreted through the clustering of significant metabolites identified by the GA based on their network distance (S1 Text). The stochasticity involved in the application of the genetic operators combined with the generation of multiple evolutions provides many optimal solutions with similar or equal fitness that nevertheless differ in metabolite content (Fig S2.5). Obtaining the same MCC score for different individuals is conceptually possible since the addition of different metabolites to the BOF may trigger the

same gene essentiality prediction. Conversion of optimal results into biological knowledge is therefore critical for modeling work since modelers ultimately need to decide which metabolites to add to their BOF. BOFdat provides a way to interpret the GA results by applying spatial clustering to the HOF of combined evolutions. The distance matrix is obtained by calculating the shortest path between metabolites in the network (S1 Text). Clustering of the distance matrix is performed with the DBSCAN algorithm for a list of significant metabolites selected based on their frequency of appearance in the HOF of the combined evolutions (Fig 2.5A, Fig S2.6). The term "species-specific metabolic end goal" is applied to describe these clusters as they link biological knowledge provided by the network reconstruction to relevant end goals found with the GA (Fig 2.5B).

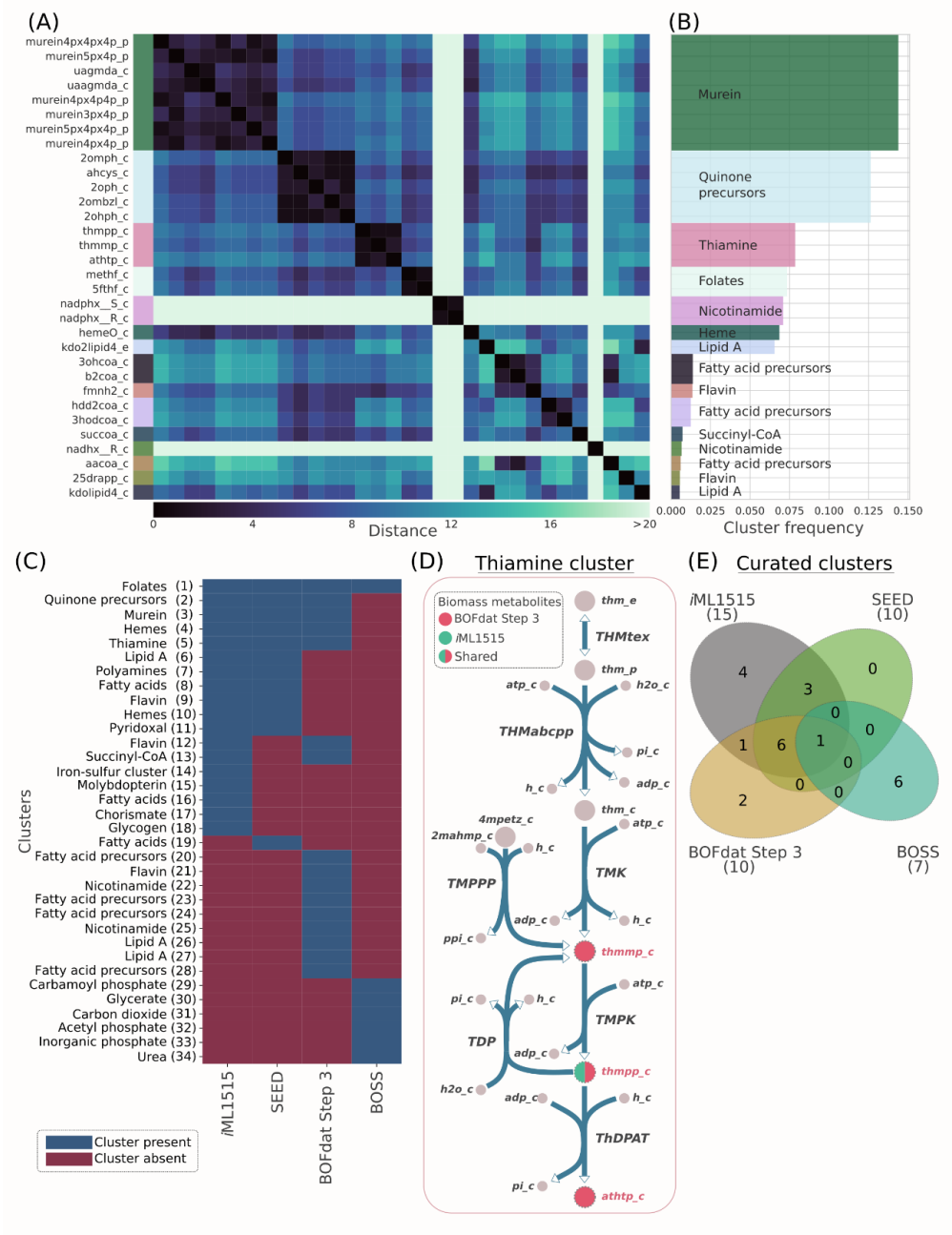

Figure 2.5 Identification of metabolic end goals by BOFdat Step 3. (A) The distance matrix for the metabolites selected based on their individual occurrence in HOFs for 150 different evolutions were clustered using DBSCAN (eps=8, S1 Text) (Fig S2.6). (B) Each of the 15 clusters identified in A were named by manually identifying the most frequent metabolite ontology in EcoCyc. The cluster frequency is the sum of each individual metabolite frequency within the cluster. (C) The metabolites from the original *i*ML1515, SEED, BOFdat and BOSS are pooled together, and clustered based on network distance using DBSCAN (eps=10, S1

Text). The clusters that are present (blue) and absent (red) for a given method are identified. The naming of the clusters was performed manually and the metabolite composition of each of them is available in S2 File. (D) Metabolic map of the thiamine cluster identified in B. The metabolites selected by BOFdat (red) are within 2 reactions of each other, and the biomass component from *i*ML1515 (green) lies in the middle. (E) Clusters shown in C were curated to group together those representing the same end goals and presented as a Venn diagram.

2.8 RESULTS AND DISCUSSION

To validate BOFdat, we reconstructed the BOF of *Escherichia coli* K-12 MG1655. We compared our obtained biomass compositions and phenotypic predictions with the recently updated *E.coli* model *i*ML1515 [23] and with two available methods to generate BOF: SEED and BOSS. The phenotypic predictions were formulated by setting the method's biomass composition as the new objective and FBA was performed. SEED is an automated platform for genome-scale metabolic reconstruction from genome information [9]. When generating a model with SEED, the user decides the biomass composition of the organism of interest from a set of predefined BOFs, based on phylogenetic relationship. For our purpose, we chose the gram-negative bacterial BOF and converted the obtained SEED metabolite identifiers to BiGG identifiers with MetaNetX [32] (S1 File). The metabolites for which no correspondence was obtained were manually converted. The Biological Objective Solution Search (BOSS) [11] algorithm uses 13C-based flux analysis data (MFA) to infer the stoichiometric coefficients for a new column of the stoichiometric matrix, which defines the BOF. In contrast to the SEED that uses *a priori* knowledge to determine the BOF composition and thus can be categorized as a biased method, BOSS uses a mixed-integer linear programming (MILP) optimization to extract biomass precursors and components from experimental data and is considered unbiased. In order to provide a valid point of comparison with the unbiased third step of BOFdat, we used an expanded version of the original BOSS that performs the optimization at the genome-scale (S1 Text) [33][33]. GEMs have been shown to efficiently predict cellular phenotypes such as growth rate and gene essentiality [34]. As illustrated in Fig 2.6A–B,

BOFdat provides a method to ensure that the BOF resulting from its use provides predictions as close as possible to experimental measurements.

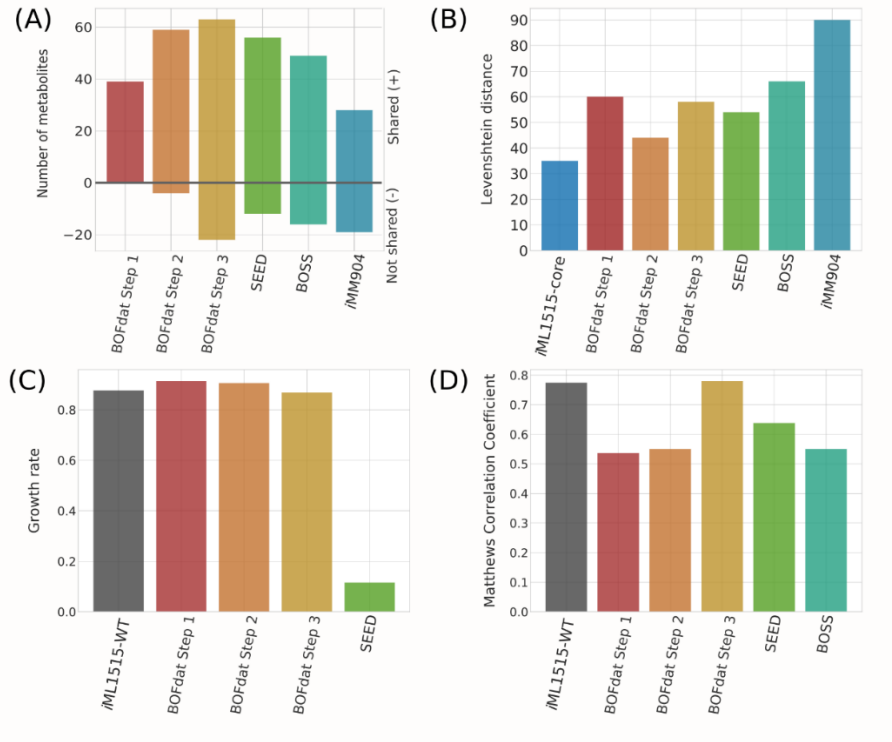

Figure 2.6 Comparison of phenotypic predictions and metabolite composition between the three steps of BOFdat, the original *i*ML1515 BOF, SEED, and BOSS. (A) Number of metabolites shared with the *i*ML1515 wild-type (WT) biomass (positive values), and specific to each of the method (negative values). (B) Levenshtein distance calculated between the original *i*ML1515-WT biomass and the BOF generated by each of the other methods, as well as with the *i*ML1515-core and the yeast model *i*MM904 BOF. The Levenshtein distance represents the number of additions, subtractions or substitutions that need to be applied on the list of metabolites from the compared BOF to retrace the reference BOF. (C) The predicted growth rates for all Steps of BOFdat are compared to the *i*ML1515-WT and SEED. BOSS imposes a fixed growth rate as part of the optimization problem and was hence not compared since it does not formulate a prediction. (D) Gene essentiality prediction as evaluated with a Matthews correlation coefficient when compared with experimental data generated on glucose minimal media [2].

2.8.1 Using omic datasets with macromolecular weight fractions allows to accurately calculate stoichiometric coefficients

The stoichiometric coefficients for four categories of major macromolecules were calculated using BOFdat (Fig 2.2C). We validated the use of omic datasets to calculate stoichiometric coefficients by comparing BOFdat generated stoichiometric coefficients against the original *i*ML1515 biomass and SEED. Fig 2.2C shows the stoichiometric coefficients for the four categories of macromolecules obtained from *i*ML1515, SEED or calculated with BOFdat Step 1. For all macromolecular categories, the median of the stoichiometric coefficients calculated with BOFdat is closer to the original *i*ML1515 median than the one obtained using SEED.

Given the availability of experimental data, BOFdat can be used to calculate stoichiometric coefficients for varying conditions. Here we used a high-quality quantitative proteomic dataset [35] to calculate the stoichiometric coefficients of each amino acid under 18 different conditions (Fig S2.7). For each condition, this dataset includes the MWF of proteins and quantitative proteomic measures, allowing to compare the impact of both parameters on the determination of the stoichiometric coefficients. We found that the MWF is significantly more important than the omic dataset itself (Pearson r = 0.984, *p-value* = 3.610e-14) (S1 Text). This result should incite modelers to query precise determination of the cell's composition under the studied condition [8]. This result also implies that, in a case where expression data is unavailable, BOFdat can still be used to calculate stoichiometric coefficients by replacing the relative abundance of each transcript/expressed protein by equal values (e.g. 1).

2.8.2 BOFdat identifies biomass precursors as clusters of metabolites

BOFdat Step 3 generates BOFs that allows the model gene essentiality prediction to match experimental data optimally. In our case study, the MCC of the generated BOF equaled or

exceed that of the original *i*ML1515 wild-type BOF (Fig 2.6B, Fig S2.3). However, the metabolite content varies from an optimal solution to another, a consequence of the stochasticity involved in the GA [31]. In order to generate a comprehensive and robust solution usable by modelers, BOFdat performs a final step of interpretation that involves clustering relevant metabolites, i.e. those that were frequently found in individuals with higher fitness, based on their relative distance in the metabolic network. The rationale supporting the clustering approach is that neighboring metabolites in the network may represent a single metabolic objective.

To evaluate the approach, we generated 150 different initial populations and evolved each of them for 500 generations. The optimal solutions from those 150 evolutions were filtered and clustered as described above (S1 Text and Fig S2.6) and 15 clusters were identified. The distance matrix (Fig 2.5A) allows discerning distinct clusters supporting the rationale that neighbor metabolites in the network represent metabolic objectives. The metabolite frequency is the number of times a metabolite appeared in an individual over the number of individuals with higher fitness. The cluster frequency is defined as the sum of all metabolite frequencies within this cluster. In Fig 2.5B, the clusters were ranked by order of cluster frequency and manually named according to the most representative ontology found on EcoCyc [36] (https://ecocyc.org/). Naming the clusters creates a link from identified metabolites to knowledge, where the name of a cluster represents a metabolic objective.

We describe those objectives and their importance for the first 3 clusters identified by order of cluster frequency. The first cluster represents the murein content of the cell, a key cell wall component of *E. coli* that maintains bacterial shape and allows resistance to intracellular osmotic pressure [37]. The second cluster is composed of metabolites involved in ubiquinone biosynthesis, also known as coenzyme Q10 (1,4-benzoquinone), critical for cellular respiration [38]. While both SEED and *i*ML1515 identified the production of quinone as biomass objective (Fig 5C), the metabolites chosen to represent it are different (see metabolites specific

to SEED, Fig 2.3C). The fact that different metabolites may represent the same biomass objective and create pseudo-diversity amongst BOFs is a concern raised by Rocha and colleagues [5]. We suggest that clustering metabolites based on network distance provides a solution to this recurrent issue. The third cluster is composed of thiamine (vitamin B1) related compounds, an essential coenzyme in *E. coli* [39] (Fig 2.5D). This cluster demonstrates how, in practice, BOFdat identifies metabolites in a linear pathway as demonstrated in Fig 2.4B. BOFdat identified 3 different metabolites in this pathway and one of them is present in *i*ML1515 BOF. Together, these observations support the fact that the identified clusters are consistent with current biological knowledge, with the more frequent clusters being species-specific essential components. Moreover, BOFdat provides an approach for the selection of biomass precursors and components that is condition-dependent and actively represents cellular objectives. Perhaps this type of information can be useful to modelers because it yields data-driven information on the cell's composition and its metabolic objectives, a challenge that was formulated before [12,13].

2.8.3 Benchmarking the clustering approach

We compared the metabolites found with BOFdat by applying spatial clustering to the complete list of metabolites found in each method (Fig 2.5C). In total 34 clusters were identified of which 18 associated with *i*ML1515, 12 with SEED, 16 with BOFdat and 7 with BOSS. These raw results revealed that SEED is the closest method to *i*ML1515 with 11 shared clusters, while BOFdat shares 7 and BOSS shares 1. The naming of the clusters found for each method was executed in the same way as for BOFdat alone (Fig 2.5B) and clusters sharing the same name were grouped together since they represent the same metabolic objective (Fig 2.5E). While SEED and *i*ML1515 share more biomass objectives together than the unbiased computational methods, BOFdat shares more curated clusters with *i*ML1515 (8) than BOSS (1). The number of non-shared clusters is also lower for BOFdat (2) than for BOSS (6). The underperformance of BOSS to identify biomass precursors is likely to be attributed to the coverage of the data used to generate the predictions. While state-of-the-art MFA data was

used for the prediction [40] (S1 Text), the number of covered reactions only reached ~60. As noted previously [11], MFA typically uses simplified models of central metabolism with lumped reactions, as is the case with the MFA data used here. Genome-wide gene essentiality data used by BOFdat hence circumvents these known issues of inferring metabolic objectives from MFA [11].

The two clusters specific to BOFdat are nicotinamide and fatty acid precursors. Nicotinamide was found in Step 2 (Fig 2.3C), but the damaged hydrated forms (nadhx and nadphx) were identified by BOFdat Step 3. Fatty acid precursors were identified despite the fact that lipids were identified in Step 1. These specific metabolic objectives are consistent with other methods (Fig. 2.5) and previous Steps of BOFdat.

BOFdat Step 3 aims to facilitate the identification of metabolic objectives by modelers by providing a data-driven way of selecting metabolites. An automated final solution is suggested by BOFdat following the clustering into biomass objectives (S1 Text). To provide users with recommendations on usage, we analyzed the impact of the number of evolutions on 1) the resulting MCC scores of these automated final solutions and 2) the number of clusters that were shared with the original *i*ML1515 biomass (Fig S2.8). The impact of hyper-parameters on the clustering algorithm were also studied (Fig S2.9, S1 Text). In brief, while the number of shared clusters steadily increased from 10 to 150 evolutions pooled together, the optimal MCC score could be reached after 20 evolutions (Fig S2.8A). A single evolution of 500 generations required ~480 hours (~21 hours on a 24 cores compute node) with a memory peak of ~10GB of RAM. To fulfill the requirements mentioned above (20 evolutions for 166 generations) would require ~3360 hours.

2.8.4 BOFdat generates biomass objective functions recapitulating key model predictions

We characterized the impact of each step of the procedure of BOFdat by evaluating the phenotypic prediction after each step and comparing with SEED and BOSS (Fig 2.6). We compared the metabolite content of all methods to the *i*ML1515-WT BOF (Fig 2.6A). As a negative control, we used the BOF from a distant organism, using the yeast model *i*MM904 [41]. The number of metabolites that were shared with the *i*ML1515 BOF increased at each step. While the number of shared metabolites is higher for BOFdat Step 3 than any other step or method, the metabolites that were not shared also increased, as expected, since BOFdat finds biomass objectives in clusters of metabolites (Fig 2.5). To account for correct additions, missed metabolites or wrongly added ones, we calculated the Levenshtein distance (S1 Text) of each biomass composition to the *i*ML1515-WT BOF (Fig 2.6B). This distance allows one to compare lists of different sizes which is the case for all BOF presented here. A smaller Levenshtein distance means less modifications need to be applied to trace back to the reference. The Levenshtein distance is minimized in Step 2 (from 60 to 44), by the addition of inorganic ions and coenzymes. The distance increases when adding the metabolites from Step 3 chosen after spatial clustering and the automated selection (58), slightly higher than SEED (54) but lower than the other unbiased approach BOSS (66). The predicted growth rate (Fig 2.6C) generated by BOFdat (0.868 h-1) is closer to the original prediction (0.877 h-1) than the one from SEED (0.116 h-1). This prediction is likely to be attributed to a correct prediction of GAM and NGAM costs. The BOSS approach was not compared for growth rate as the coefficients obtained with it requires fixing the growth rate to a certain score. The gene essentiality prediction was assessed using Matthews Correlation Coefficient (MCC, Fig 2.6D) and step three of BOFdat (MCC = 0.779) was found to be higher than the original prediction (MCC = 0.775) as well as both SEED (MCC = 0.638) and BOSS (MCC = 0.550). The performance advantage of BOFdat on that aspect is explained by the fact that the genetic algorithm (Step 3) specifically minimizes the distance from experimental data (Fig 2.4). Indeed, Steps 1 and 2 obtained MCC scores of 0.536 and 0.550, respectively. Together, these

elements suggest that the definition of a BOF from experimental data with BOFdat provides acceptable phenotypic predictions that are in par or exceed existing methods.

2.9 AVAILABILITY AND FUTURE DIRECTIONS

BOFdat is a systematic computational framework for the definition of biomass objective functions for genome-scale models from experimental data. BOFdat is implemented in Python and is conceived to be a part of the COBRApy ecosystem [27], an environment commonly used to build and analyze GEMs. Our package is accessible for installation through the Python package index (PyPI). The open-source code as well as an example usage and required input files are available on GitHub (https://github.com/jclachance/BOFdat) where a link to the full documentation and API is available. We are confident that its use will contribute to more reliable BOF and GEMs. By providing an unbiased, data-driven approach to defining biomass objective functions, BOFdat has the potential to improve the quality of new genome-scale models and also greatly decrease the time required to generate a new reconstruction. With the increasing number of omic datasets being generated and the realization of community models, we expect BOFdat to be leveraged for the generation of condition- and strain-specific BOF definitions, hereby increasing the quality of GEMs phenotypic predictions. While using experimental essentiality data provides metabolic end goals, other unbiased methods using different data types could be developed and added to the package, hence revealing more precisely the metabolic end goals of the cells.

2.10 ACKNOWLEDGMENTS

The authors would like to thank members of the Systems Biology Research Group that contributed to this work through software testing and discussions: Jared Broddrick, Erol Kavvas, Charles J. Norsigian, Saugat Poudel and Edward Catoiu.

2.11 REFERENCES

1. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. Nat Biotechnol. 2014;32: 447–452. doi:10.1038/nbt.2870
2. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010;13: 344–349. doi:10.1016/j.mib.2010.03.003
3. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5: 93–121. doi:10.1038/nprot.2009.203
4. Dikicioglu D, Kırdar B, Oliver SG. Biomass composition: the "elephant in the room" of metabolic modelling. Metabolomics. 2015;11: 1690–1701. doi:10.1007/s11306-015-0819-2
5. Xavier JC, Patil KR, Rocha I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. Metab Eng. 2017;39: 200–208. doi:10.1016/j.ymben.2016.12.002
6. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;330: 1099–1102. doi:10.1126/science.1192588
7. Volkmer B, Heinemann M. Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. PLoS One. 2011;6: e23126. doi:10.1371/journal.pone.0023126
8. Beck AE, Hunt KA, Carlson RP. Measuring Cellular Biomass Composition for Computational Biology Applications. Processes. Multidisciplinary Digital Publishing Institute; 2018;6: 38. doi:10.3390/pr6050038
9. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol Biol. 2013;985: 17–45. doi:10.1007/978-1-62703-299-5_2
10. Burgard AP, Maranas CD. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol Bioeng. Wiley Online Library; 2003;82: 670–677. Available: http://onlinelibrary.wiley.com/doi/10.1002/bit.10617/full
11. Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA. Predicting biological system objectives de novo from internal state measurements. BMC Bioinformatics. 2008;9: 43. doi:10.1186/1471-2105-9-43

12. Zhao Q, Stettner AI, Reznik E, Paschalidis IC, Segrè D. Mapping the landscape of metabolic goals of a cell. Genome Biol. 2016;17: 109. doi:10.1186/s13059-016-0968-2

13. Feist AM, Palsson BO. What do cells actually want? Genome Biol. 2016;17: 110. doi:10.1186/s13059-016-0983-3

14. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 2018;557: 503–509. doi:10.1038/s41586-018-0124-0

15. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248. doi:10.1038/nbt.1614

16. Varma A, Palsson BO. Metabolic capabilities of Escherichia coli: I. synthesis of biosynthetic precursors and cofactors. J Theor Biol. 1993;165: 477–502. Available: https://www.ncbi.nlm.nih.gov/pubmed/21322280

17. Chan SHJ, Cai J, Wang L, Simons-Senftle MN, Maranas CD. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. Bioinformatics. 2017;33: 3603–3609. doi:10.1093/bioinformatics/btx453

18. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics. 2007;8: 212. doi:10.1186/1471-2105-8-212

19. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. Exploring the diversity of complex metabolic networks. Bioinformatics. 2005;21: 1603–1609. doi:10.1093/bioinformatics/bti213

20. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, et al. Systems approach to refining genome annotation. Proc Natl Acad Sci U S A. 2006;103: 17480–17484. doi:10.1073/pnas.0603364103

21. Kumar VS, Maranas CD. GrowMatch: an automated method for reconciling *in silico*/*in vivo* growth predictions. PLoS Comput Biol. 2009;5: e1000308. doi:10.1371/journal.pcbi.1000308

22. Orth JD, Palsson BØ. Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng. 2010;107: 403–412. doi:10.1002/bit.22844

23. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes Escherichia coli traits. Nat Biotechnol. 2017;35: 904–908. doi:10.1038/nbt.3956

24. Gerlee P, Lizana L, Sneppen K. Pathway identification by network pruning in the metabolic network of Escherichia coli. Bioinformatics. 2009;25: 3282–3288. doi:10.1093/bioinformatics/btp575

25. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–22. doi:10.1093/nar/gkv1049

26. Fortin F-A, Rainville F-MD, Gardner M-A, Parizeau M, Gagné C. DEAP: Evolutionary Algorithms Made Easy. J Mach Learn Res. 2012;13: 2171–2175. Available: http://www.jmlr.org/papers/volume13/fortin12a/fortin12a.pdf

27. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013;7: 74. doi:10.1186/1752-0509-7-74

28. Kleckner N, Roth J, Botstein D. Genetic engineering in vivo using translocatable drug-resistance elements. New methods in bacterial genetics. J Mol Biol. 1977;116: 125–159. Available: https://www.ncbi.nlm.nih.gov/pubmed/338917

29. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, et al. A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. Cell. 2016;165: 1493–1506. doi:10.1016/j.cell.2016.05.003

30. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One. 2017;12: e0177678. doi:10.1371/journal.pone.0177678

31. Diaz-Gomez PA, Hougen DF. Initial Population for Genetic Algorithms: A Metric Approach. GEM. 2007. pp. 43–49. Available: http://www.cameron.edu/~pdiaz-go/GAsPopMetric.pdf

32. Moretti S, Martin O, Van Du Tran T, Bridge A, Morgat A, Pagni M. MetaNetX/MNXref--reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 2016;44: D523–6. doi:10.1093/nar/gkv1117

33. Yang L, Bento J, Lachance J-C, Palsson BO. Genome-scale estimation of cellular objectives [Internet]. arXiv [q-bio.QM]. 2018. Available: http://arxiv.org/abs/1807.04245

34. O'Brien EJ, Monk JM, Palsson BO. Using Genome-scale Models to Predict Biological Capabilities. Cell. 2015;161: 971–987. doi:10.1016/j.cell.2015.05.019

35. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent *Escherichia coli* proteome. Nat Biotechnol. 2016;34: 104–110. doi:10.1038/nbt.3418

36. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, et al. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. Nucleic Acids Res. 2017;45: D543–D550. doi:10.1093/nar/gkw1003

37. Höltje JV. Growth of the stress-bearing and shape-maintaining murein sacculus of Escherichia coli. Microbiol Mol Biol Rev. 1998;62: 181–203. Available: https://www.ncbi.nlm.nih.gov/pubmed/9529891

38. Cox GB, Newton NA, Gibson F, Snoswell AM, Hamilton JA. The function of ubiquinone in *Escherichia coli*. Biochem J. 1970;117: 551–562. Available: https://www.ncbi.nlm.nih.gov/pubmed/4192611

39. Bazurto JV, Farley KR, Downs DM. An Unexpected Route to an Essential Cofactor: Escherichia coli Relies on Threonine for Thiamine Biosynthesis. MBio. 2016;7: e01840–15. doi:10.1128/mBio.01840-15

40. Haverkorn van Rijsewijk BRB, Nanchen A, Nallet S, Kleijn RJ, 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

41. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol. 2008;26: 1155–1160. doi:10.1038/nbt1492

2.12 SUPPLEMENTARY TEXT

2.12.1 BOFdat Step 1

2.12.1.1 Generating stoichiometric coefficients from omic datasets and macromolecular weight fractions

The implementation of BOFdat is modular, meaning that each function can be operated individually depending on the availability of data. Each module of BOFdat Step 1 calculates

the stoichiometric coefficients for a different category of macromolecule. DNA, RNA, proteins and lipids stoichiometric coefficients can be calculated. To obtain the coefficients BOFdat requires the input of the macromolecular weight fraction (MWF) of the intended category. The MWF represents the total weight of a category of macromolecule within the cell dry weight. For example if 5% of the cell is composed of DNA, the DNA_{MWF} is 0.05 and stipulates that, in a steady-state growth scenario, for 1g of cell produced 0.05g of DNA is produced. To obtain a growth rate prediction, flux-balance analysis (FBA) relies on the definition of a basis where the product of cell mass by time is equal to 1g of dry weight per hour [1]. Using BOFdat and ensuring that all MWF used through the process add up to one ensures that this basis is respected and that the formulated growth rate prediction is valid.

2.12.1.2 DNA

The calculation of stoichiometric coefficients is made by counting the number of each bases in the provided fasta genome file such that:

$$R_{AT} = (A_f + T_f) / (L * 2) \quad \text{(équation 2.1)}$$

$$R_{CG} = (C_f + G_f) / (L * 2) \quad \text{(équation 2.2)}$$

where R_{AT} and R_{CG} are the ratios of the bases A, T and C, G, L is the length of the genome, A_f, T_f, C_f, G_f are the number of A, T, C and G bases counted in the single-strand genome sequence. Equation 2 is used to calculate the stoichiometric coefficient of each metabolite:

$$m_{sc} = R_m * DNA_{MWF} / (m_{MW} - Pi_{MW}) * 1000 \quad \text{(équation 2.3)}$$

where m_{sc} is the stoichiometric coefficient of a given base ($mmol * gDW$), R_m is the ratio of the metabolite for which the coefficient is determined, let it be R_{AT} or R_{CG}, within its category,

$DNA_{MWF}$is the MWF of DNA, $m_{MW}$is the molar weight of the metabolite, $Pi_{MW}$the molar weight inorganic di-phosphate (removed in the polymerization reaction).

2.12.1.3 RNA

The calculation of stoichiometric coefficients is made by counting the number of each bases in every gene in the provided GenBank annotation file such that:

$$r_b = \left(\sum_{s_i}^{s_L} [s = b] \right) / l \qquad \text{(équation 2.4)}$$

where r_b is the ratio of base b in a gene of sequence length l. The ratio for each gene is then normalized by the abundance of each gene provided in the transcriptomic file (2 column CSV file where the first column represents the gene identifier and the second is the relative abundance of the transcript). The RNA content of the cell is divided in the three main categories present in the cell: rRNA, tRNA and mRNA. Since transcriptomic experiments usually deplete the rRNA and tRNA contents, the abundance of each rRNA or tRNA coding genes is considered equal to 1. Equation 4.1 allows to calculate the ratio of a given base in mRNA. Equation 4.2 is the simplification of equation 4.1 for the ratio of a given base in rRNA and tRNA, assuming each rRNA and tRNA have an abundance of 1.

$$R_m = \sum_g (r_g * a_g) \; / \sum_g a_g \qquad \text{(équation 2.5)}$$

$$R_r = R_t = \sum_g r_g \; /N \qquad \text{(équation 2.6)}$$

where R_m, R_r, R_t are the total fraction of the base b in mRNA, rRNA and tRNA respectively, $r_g$is the ratio of base b in gene g obtained in (3), $a_g$is the abundance of gene g . The total fraction of each base is then calculated by adding the content in each RNA category:

$$R_b = (mRNA_f * R_m) + (tRNA_f * R_t) + (rRNA_f * R_r) \quad \text{(équation 2.7)}$$

where $mRNA_f, tRNA_f, rRNA_f$are the fraction of mRNA, tRNA and rRNA in the total RNA content of the cell. These values are set by default to $rRNA_f$= 0.9,$tRNA_f$= 0.05 and $mRNA_f$ = 0.05. The stoichiometric coefficient of a given base can then be determined:

$$m_{sc} = R_b * RNA_{MWF} / (m_{MW} - Pi_{MW}) * 1000 \quad \text{(équation 2.8)}$$

where $m_{sc}$is the stoichiometric coefficient of a given base ($mmol * gDW$), $R_b$is the total ratio a given base in the entire RNA content of the cell obtained in (5), $RNA_{MWF}$is the total MWF of RNA, $m_{MW}$is the molar weight of base b and $Pi_{MW}$is the molar weight of inorganic diphosphate released as a product of the polymerization reaction.

2.12.1.4 Protein

The calculation of stoichiometric coefficients for amino acids is based on the proteomic dataset and the MWF of proteins:

$$r_a = \left(\sum_{s_i}^{s_L} [s = a] \right) / l \quad \text{(équation 2.9)}$$

where r_a is the ratio of amino acid a in a protein of sequence length l. The ratio for each gene is then normalized by the abundance of each gene obtained by the proteomic dataset such that:

$$R_a = \sum_p (r_a * a_p) \; / \sum_p a_p \qquad \text{(équation 2.10)}$$

where R_a is the total fraction of the amino acid a, $r_a$is the ratio of amino acid a for protein p obtained in (6), $a_g$is the abundance of gene g obtained via the proteomic data file (2 column CSV file where the first column represents the gene identifier and the second is the relative or absolute abundance of the transcribed protein). The total fraction of each amino acid can then be used to calculate their respective stoichiometric coefficient:

$$m_{sc} = R_m * \; PROT_{MWF} \; / \; (m_{MW} - \; H_2O_{MW}) * \; 1000 \qquad \text{(équation 2.11)}$$

where $m_{sc}$is the stoichiometric coefficient of a given amino acid ($mmol \; * \; gDW$), $R_b$is the total ratio a given amino acid in the entire protein content of the cell obtained in (8), $PROT_{MWF}$is the total MWF of proteins, $m_{MW}$is the molar weight of amino acid a and $H_2O_{MW}$is the molar weight of water released as a product of the polymerization reaction.

2.12.1.5 Lipids

The organism's lipid composition can be extracted through lipidomics experiments. Lipidomics allow to obtain the identity of the lipids that compose the cell membrane as well as their relative abundances. Unlike DNA, RNA and proteins that are composed of standard metabolites, the lipids may vary from a species to another and the mapping of the identifiers provided in the experimental data to model identifiers should be provided as a conversion file (2 column CSV file). The calculation of the lipid stoichiometric coefficients is then directly calculated from the relative abundances and the lipid MWF:

$$m_{sc} = R_m * \; LIP_{MWF} \; / \; m_{MW} \; *1000 \qquad \text{(équation 2.12)}$$

where m_{sc} is the stoichiometric coefficient of a given lipid ($mmol * gDW$), R_m the relative abundance of a given lipid obtained from the lipidomic dataset, LIP_{MWF} the lipid MWF and m_{MW} the molar weight of lipid m.

2.12.1.6 Growth and non-growth associated maintenance

Maintenance costs represent the unspecified energy expenditure that are not directly part of the biomass precursor synthesis reactions. These costs lie outside of the scope of the metabolic network but are necessary to incorporate growth rate prediction in the model. The growth-associated maintenance (GAM) represents the ATP cost of generating another cell while the non-growth associated maintenance (NGAM) is associated with basal cellular processes. A standardized format for the organization of phenotypic data allows BOFdat to calculate GAM and NGAM (https://bofdat.readthedocs.io/).

Phenotypic data is parsed by BOFdat to constrain the provided model with the experimental growth rate, substrate uptake rate and the secretion rate of metabolic wastes. For a given experiment, a bound is set to the biomass objective function to set the growth rate to the experimental value. Similarly a flux is forced through the exchange reactions associated with either the uptake rate and the secretion rate(s). The model is then optimized for ATP production and the recorded flux through the ATP consumption is recorded. Going through all conditions allows to obtain a scatter plot where each growth rate is associated with an ATP cost. Linear regression is then applied and the slope of the curve is the GAM while the y-intercept is the NGAM.

2.12.2 BOFdat Step 2

2.12.2.1 Finding coenzymes

The degree of each metabolite is assessed by obtaining the number of reactions to which each metabolite takes part. The degree distribution for all metabolites in the *E. coli* metabolic network is shown in Fig S2.1. The threshold was set as the mean + one standard deviation for this distribution. Metabolites that were a result of BOFdat Step 1 and potentially included in the maintenance costs such as ATP, H+, ADP and PPi are removed from the list of potential metabolites to be added but are still included in the calculation of statistic parameters of the distribution. One of the main reason to add this step is that metabolites with a high degree are likely to be missed by the genetic algorithm as many paths may lead to their production. Network topology nevertheless clearly informs on the importance of those metabolites.

2.12.2.2 Determining stoichiometric coefficients

The calculation of the stoichiometric coefficients is simplified for BOFdat Step 2 and Step 3 where the main focus is set on the qualitative choice of metabolites to add to the BOF. Hence, the assumption is that all metabolites within this category (i.e.: coenzymes and inorganic ions) are in equal abundance within the cell. The ratio of a given metabolite R_m is therefore:

$$R_m = 1 \,/\, N \qquad \text{(équation 2.13)}$$

where, $N$is the number of metabolites within the category. The stoichiometric coefficient m_{sc} of a given metabolite is then:

$$m_{sc} = R_m * \frac{CI_{MWF}}{m_{MW}} * 1000 \qquad \text{(équation 2.14)}$$

where CI_{MWF} is the molecular weight fraction of coenzymes and inorganic ions category and m_{MW} is molar weight of the metabolite.

2.12.3 BOFdat Step 3

2.12.3.1 Generation of initial populations

The initial population is generated by extracting the list of metabolites from the provided model in order to generate an index of metabolites for every individual in the population. If provided, the metabolites that belonged to the previous steps of BOFdat are removed from the list of all metabolites contained in the model. Then, each metabolite is tested individually for solvability by removing the existing objective function in the model and applying a single objective on the production of that given metabolite. A stoichiometric coefficient of 0.1 is attributed to the metabolite, making it a reactant. The reaction is a sink and has no product. The model is then optimized. A threshold for *in silico* growth rate value is set above the numerical error (1e-9) to determine whether or not the metabolite can be produced by the model. BOFdat does not change the media conditions of the model. Hence, modellers should provide the model with the exchange reactions set as the media used for the generation of the experimental data. Once a list of solvable metabolites is generated, a pre-feature selection step is executed that selects the metabolites that are most likely to affect the measure of fitness in the machine learning approach. Much like the selection of solvable metabolites this step iteratively sets each metabolite as an objective function. The difference here is that the gene essentiality is tested for each individual metabolite, allowing to attribute a Matthews Correlation Coefficient (MCC) value to each metabolite. The MCC is obtained using the SciKit Learn method (http://scikit-learn.org/). The metabolites are thus ranked in order of MCC and a threshold of mean + one standard deviation is set to determine those that should be used in the populations and the upcoming evolution (if a metabolite is absent from the initial population it will not be present in the evolution). Each individual is then generated randomly. The individual size is

set to 20 meaning that 20 metabolites are identified as present in the population. The size of the population is adapted to get a sufficient coverage:

$$P = C * M/I \qquad \text{(équation 2.15)}$$

where P is the population size, C is the coverage (set to 10 by default), M is the length of the index and I is the individual size. This adaptation of the population size to the number of features in the metabolite index is intended to reduce the initial population bias. For each individual the list of 0 and 1 defining the presence or absence of a metabolite is converted into an objective vector by applying a stoichiometric coefficient to the row of the stoichiometric matrix corresponding to a metabolite identified as present.

2.12.3.2 Implementation of the genetic algorithm

The DEAP toolbox is used to implement a genetic algorithm that maximizes the MCC between *in vitro* and *in silico* gene essentiality data. The initial population generated by the previous function is imported. Each column of the initial population is a different individual. Before the first generation each individual is evaluated and ranked by fitness value. The fitness is parametrized by both MCC value (weight of 1.0) and by the size of the individual (weight of -0.25), which is the number of metabolites contained in the individual (Fig S2.4). This multi-parameter fitness function allows to select individuals with a minimal number of metabolites for a higher MCC value ensuring that selected metabolites are significant. The weights attributed to each parameter of the fitness function is one of the hyper-parameters that may be changed by the user. The individuals are selected with the tournament method (http://deap.readthedocs.io/). The mutation method applied is set to FlipBit (http://deap.readthedocs.io/) which is designed for binary individuals. The probability of selecting an individual to be mutated is set to 0.1 while the probability of flipping a feature is set to 0.005. The cross-over method applied is one point (http://deap.readthedocs.io/). This method randomly selects a location on the chromosome and exchanges both branches of the 2

individuals selected to breed together. The probability of crossover is set to 0.5. A common problem with genetic algorithms is the premature convergence to a local optimum. The Random Offspring Generation (ROG) method is applied to avoid it [2]. Like other methods that avoid premature convergence, the ROG favors the generation of genetic diversity. To assess the loss of diversity, the offsprings produced by crossover are compared to their parents. If no difference is recorded between children and parents, identical twins were bred and the population did not gain diversity. One parent is then replaced by a completely new individual, generated in a random fashion as done for the initial population individuals and the crossover is re-applied. The algorithm selection, mutation and crossover process will run for a given number of generations as parametrized by the users.

2.12.3.3 Clustering into metabolic end goals

Each evolution generates a hall of fame (HOF) that contains the best individuals that were produced within it (http://deap.readthedocs.io). The default size of the HOF is 1000, meaning that the 1000 best individuals generated in an evolution are kept. To interpret the data generated by BOFdat over multiple evolutions, a spatial clustering method is implemented (Fig S2.6). The HOF of each evolutions are pooled together and an arbitrary threshold is applied to select the best individuals. By default, the 20% best are chosen. These selected individuals are then analyzed for their metabolite content. The frequency of apparition of each metabolite across all individuals is calculated and a threshold is set to keep only the most frequent, hereby reducing noise in data. The threshold is set to the mean of the frequency distribution, in the reconstruction of the *E.coli* biomass presented here (150 evolutions over 500 generations) the mean frequency was 0.00538. A distance matrix is generated using the Dijkstra algorithm finding the shortest path between nodes. Hubs are known to introduce bias in paths, greatly shortening paths between nodes that are not closely related. Therefore, nodes with more than 15 connections are removed from network distance analysis. This number may vary from a metabolic network to another depending on the number and degree of each metabolite, hence the related threshold may be set by users as described in the documentation

(https://bofdat.readthedocs.io/). The adjacency matrix obtained for the selected metabolites is clustered using the DBSCAN [3] algorithm from the SciKit Learn library (http://scikit-learn.org/). Unlike hierarchical clustering methods, DBSCAN does not require *a priori* definition of the number of clusters but requires the minimum number of elements to include in a cluster, (suggested and the default value of 0.5), and the maximum distance between elements to form a cluster, termed eps in the DBSCAN documentation. This second value greatly impacts the cluster formation as a greater eps would group all metabolites into a single cluster (i.e.: 30) whereas a shorter eps (i.e.: 1) separates every metabolite into a unique cluster. The impact of varying the eps is studied in Fig S2.9. As mentioned, when a low eps is given, the number of clusters is maximized to a point where the number of clusters is equivalent to the number of metabolites, providing no mechanistic explanation whereas a greater eps yields very few clusters, also making the interpretation difficult. For this study we have used eps values that yield a number of clusters at half the number of provided metabolites and have come with results that could be interpreted by a modeller with knowledge of the organism.

BOFdat provides a way to automatically select metabolites from the formed clusters. The cluster frequency is calculated by adding metabolite frequency:

$$C_F = \sum_{m_i}^{i} m_F \qquad \text{(équation 2.16)}$$

The z-score of each cluster is than calculated:

$$C_z = \frac{C_F - \overline{x_c}}{S_C} \qquad \text{(équation 2.17)}$$

where $C_z$is the z-score of a cluster, $\underline{x_C}$the average of all cluster frequencies, $S_C$the standard deviation of all cluster frequencies and$C_F$the individual cluster frequency obtained from (1).

Each cluster is then weighted according to its z-score:

$$\begin{aligned} &C_z \geq 1, w_C = 3, \\ &1 > C_z \geq 0, w_C = 2, \\ &C_z < 0, w_C = 1. \end{aligned} \qquad \text{(équation 2.18)}$$

The weight of each cluster determines the number of metabolites that should be added to the final BOF suggested by BOFdat. If the weight is superior to the number of metabolites in the cluster, the number of metabolites added is equal to the size of the cluster. The stoichiometric coefficients of each metabolite added in BOFdat Step 3 are determined in a similar way as for BOFdat Step 2 where the ratio of each metabolite is by default distributed equally across all metabolites present in that step (Eq. 11 and 12).

2.12.4 Methods

2.12.4.1 Using omic datasets to calculate stoichiometric coefficients

BOFdat uses both omic datasets and macromolecular weight fractions (MWF) to calculate stoichiometric coefficients. The omic datasets allow to get the relative abundance of transcribed proteins or RNA transcripts. To evaluate the importance of these parameters, we calculated stoichiometric coefficients for amino acids under 18 different experimental conditions. A high-quality proteomic dataset for *E.coli* was generated by Heinemann and colleagues [4] under multiple experimental conditions. Along with the precise quantity of each protein present in the cell, the protein weight fractions were also measured for every growth condition making the data perfectly suitable for our study. Using equation 17 stoichiometric coefficients for all growth conditions were compared to the data obtained on glucose minimal media (Fig S2.7):

$$e_a = \frac{|n_a - g_a|}{g_a} * 100 \qquad \text{(équation 2.19)}$$

where $e_a$is the percentage difference between the reference coefficient and the new coefficient for a given amino acid, n_a is the new coefficient in the given experimental condition, g_a is the original coefficient on glucose minimal media. When using condition-specific protein weight fractions, the mean difference from glucose was between 4.1% (Acetate) and 24.2% (Stationary phase 1 day). The weight fraction differences correlated with the average percent difference for each condition (Pearson r = 0.984, *p-value* = 3.610e-14) indicating that the weight fraction is the most prevailing parameter to determine stoichiometric coefficients. The use of appropriate weight fractions has a greater impact on the stoichiometric coefficients generated and should be considered as the most significant experimental measurement in the biomass composition of GEMs.

2.12.4.2 Required number of evolutions

To provide a recommendation for the number of evolutions that should be performed by a user of BOFdat we generated the MCC values and the number of shared clusters with *i*ML1515 wild-type BOF for a number of evolutions pooled together ranging from 10 to 150 pooled evolutions (Fig S2.8). 20 random samples were formed for each number of pooled evolutions. For each sample, the evolutions were randomly selected from 187 individual evolutions. A selection of metabolites was operated as described above for each sample generated. The metabolite selection was then used to evaluate the MCC value of the given solution (Fig S2.8A). Similarly, for each sample, the selected metabolites were pooled with the metabolites from the original *i*ML1515 BOF and clustered using the method described above (eps = 8, min_sample = 0.5) (Fig S2.8B). The maximum MCC value obtained was 0.779 and was observed as the median for more than 40 evolutions pooled together. For users with fewer computational resources, we note that pooling 20 evolutions had a median MCC higher than the baseline established by *i*ML1515 at 0.775. The percent shared clusters with *i*ML1515

steadily increased between 10 evolutions (26.61%) and 150 evolutions (31.19%), meaning that the accuracy of the metabolites found by BOFdat is dependent to the number of evolutions generated.

2.12.4.3 Multiple correspondence analysis

Multiple correspondence analysis (MCA) was used to compare feature selected by different evolutions [5]. The presence or absence of a metabolite in an individual is defined as a nominal variable with two levels (present = 1, absent = 0) and the total number of nominal variables K is the number of unique metabolites in all selected individuals (119 metabolites). The number of observations I is defined as the total number of individuals compared to each other. We selected 35 individuals with fitness values ranging from the lowest to the highest recorded in the 150 evolutions used in this study. The 35 individuals were extracted from 10 different HOFs. The X matrix of size K feature columns by I observations rows was processed following the usage guide from the Python MCA package (https://pypi.org/project/mca/).

2.12.4.4 Biomass objective function from SEED

The SEED platform for automated reconstruction of genome-scale models of metabolism allows for the selection of pre-defined biomass objective function that is based on knowledge (File S1). The default gram-negative bacteria BOF was chosen when reconstructing the model for *Escherichia coli* MG1655 on the SEED platform. The objective function was extracted from the model and the SEED metabolite identifiers were converted to BiGG identifiers using MetaNetX [6]. The remaining unconverted metabolites were converted manually to the best matching BiGG identifier.

2.12.4.5 Using BOSS on a genome scale

We used the BOSS (Biological Objective Solution Search) method [7] to generate a biomass objective reaction from fluxomics data. We used ^{13}C metabolic flux analysis data for *E. coli* grown on glucose minimal medium [8]. To solve BOSS for the genome-scale model [9], we used a recent algorithmic extension of BOSS, which solves the nonconvex optimization problem of BOSS using a distributed optimization method called ADMM (alternating direction method of multipliers) [10]. We made three additional modifications to the extended BOSS for this study. First, we constrained BOSS to keep the biomass coefficients from BOFdat Step 2, to facilitate the comparison between BOFdat Step 3 vs. BOSS. Second, because ADMM can require many iterations to reach a high-precision solution, we implemented a solution "polishing" step. The polishing step enables taking a medium-precision solution from ADMM (e.g., feasibility tolerance ~ 10^{-6}) and subsequently solving one optimization problem to obtain a high-precision solution (e.g., feasibility tolerance ~ 10^{-9}). The polishing step solves the following quadratic program (QP):

$$\begin{aligned} &\min_{v,y} ||y - y^B||^2{}_2 \\ &subject\ to\ Sv + yz^b = 0, \\ &l \le v \le u \end{aligned} \qquad \text{(équation 2.20)}$$

where v is the vector of fluxes, y is the vector of stoichiometric coefficients for the generated objective, y^B is the stoichiometric vector from the BOSS solution, z^B is the flux through the BOSS objective reaction, S is the stoichiometric matrix for the metabolic reconstruction, and l and u are lower and upper flux bounds, respectively. The solution to this problem provides the objective reaction that most closely matches the BOSS solution that still satisfies the FBA constraints to a higher precision. Third, we generated a sparse objective using both an L1-norm regularization during the BOSS computation, and also a post-processing step. For post-processing, we sparsified the vector, y^B, in the polishing step (QP) above by only keeping the largest 200 coefficients ordered by magnitude and zeroing out the rest of the coefficients.

Therefore, the solution to the polishing step generated a sparsified objective that still satisfies FBA constraints.

2.12.4.6 Levenshtein distance calculation

The Levenshtein distance allows to compute distance between strings of different lengths. The BOF generated by all three steps of BOFdat, SEED, BOSS and the biomass of a completely different organism (yeast, *i*MM904) were compared to the original *i*ML1515 wild-type biomass using this distance metric. The list of metabolites from the reference (original *i*ML1515) and the compared biomass are converted to strings where each letter corresponds to the presence or absence of a metabolite in the reference. The reference as a presence indicator for each metabolite whereas the compared biomass as both presence or absence. In that case, the Levenshtein distance is equivalent to the number of absence indicators.

2.12.5 Supplementary references

1. Varma A, Boesch BW, Palsson BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. Appl Environ Microbiol. 1993;59: 2465–2473. Available: https://www.ncbi.nlm.nih.gov/pubmed/8368835

2. Rocha M, Neves J. Preventing Premature Convergence to Local Optima in Genetic Algorithms via Random Offspring Generation. Multiple Approaches to Intelligent Systems. Springer Berlin Heidelberg; 1999. pp. 127–136. doi:10.1007/978-3-540-48765-4_16

3. Ester M, Kriegel H-P, Sander J, Xu X, Others. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996. pp. 226–231. Available: http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

4. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent *Escherichia coli* proteome. Nat Biotechnol. 2016;34: 104–110. doi:10.1038/nbt.3418

5. Abdi H, Valentin D. Multiple correspondence analysis. Encyclopedia of measurement and statistics. 2007; 651–657. Available: https://www.researchgate.net/profile/Dominique_Valentin/publication/239542271_Multi

ple_Correspondence_Analysis/links/54a979900cf256bf8bb95c95.pdf

6. Moretti S, Martin O, Van Du Tran T, Bridge A, Morgat A, Pagni M. MetaNetX/MNXref--reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 2016;44: D523–6. doi:10.1093/nar/gkv1117

7. Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA. Predicting biological system objectives de novo from internal state measurements. BMC Bioinformatics. 2008;9: 43. doi:10.1186/1471-2105-9-43

8. Van Rijsewijk BRBH, Nanchen A, Nallet S, Kleijn RJ, Sauer U. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. Mol Syst Biol. EMBO Press; 2011;7: 477. Available: http://msb.embopress.org/content/7/1/477.short

9. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. Nat Biotechnol. 2017;35: 904–908. doi:10.1038/nbt.3956

10. Yang L, Bento J, Lachance J-C, Palsson BO. Genome-scale estimation of cellular objectives [Internet]. arXiv [q-bio.QM]. 2018. Available: http://arxiv.org/abs/1807.04245

2.13 SUPPLEMENTARY FIGURES

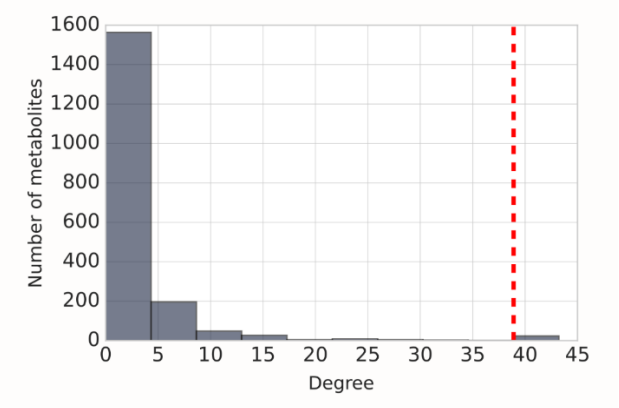

Figure S2.1. Distribution of the metabolites degree in the *E. coli* metabolic network used to identify coenzymes (BOFdat Step 2). As established in the literature, the degree is the number of metabolites to which a given metabolite is linked in the network. The red dashed line represents the threshold to identify a coenzyme, set as the mean (5.63) + one standard deviation (33.26) of 1877 metabolites of the network.

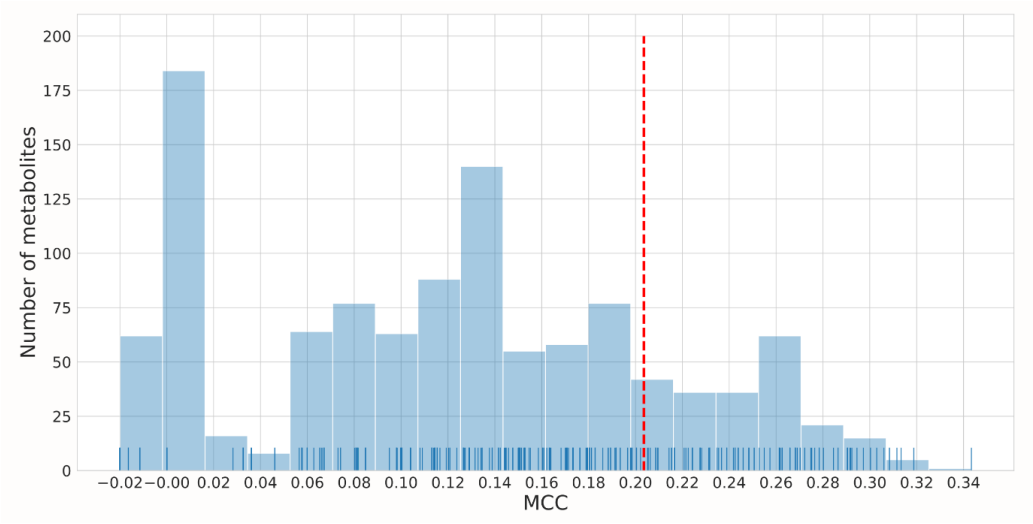

Figure S2.2 Distribution of individual MCC values for metabolites. The pre-feature selection for the generation of initial populations of BOFdat Step 3 includes the evaluation of the MCC value of each remaining metabolites (dark blue bars). The *i*ML1515 model contains 1877 metabolites, 63 metabolites were removed because they were present in BOFdat Step 2 and 766 were removed because the model could not solve when they were set as an objective. The MCC was determined for the 1048 remaining metabolites. The threshold is set at the mean of the distribution + one standard deviation (MCC = 0.20) allowing to select 186 metabolites with the highest potential and referred to as the metabolites subset.

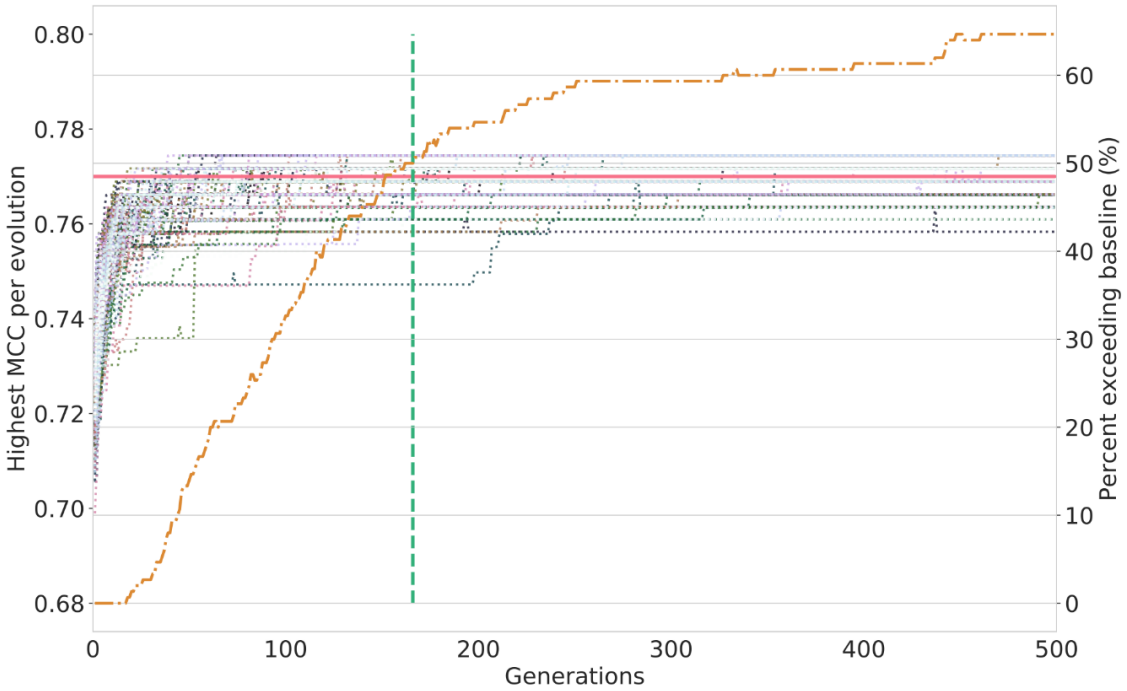

Figure S2.3 Distribution of the genetic algorithm output for 150 evolutions over 500 generations. At each generation, the highest MCC value of the population was plotted for each evolution (left Y-axis). The solid horizontal red line represents the baseline established by the MCC value of the original *i*ML1515 BOF (0.775). The orange dashed line represents the percentage of evolutions in which the best individual exceeds the baseline (right Y-axis). The vertical green line shows the number of generations necessary to obtain 50% of evolutions exceeding the baseline, in this case 166.

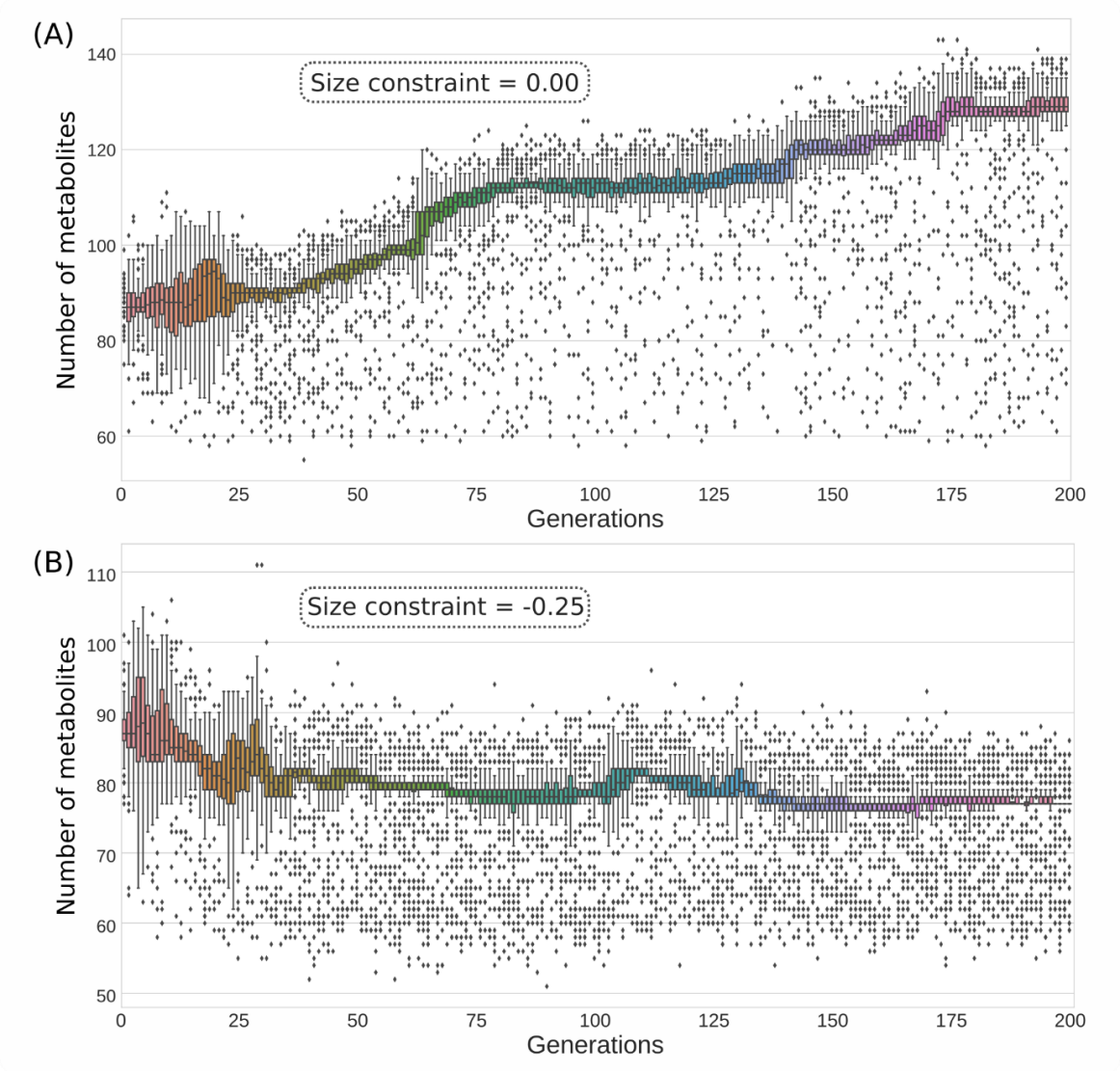

Figure S2.4 Impact of the size constraint in the genetic algorithm. Number of metabolites contained in each individual (Y-axis) present in every generation over a typical evolution. The size constraint is a second weight applied on the fitness function. Here a weight of 1.0 is applied to maximize the MCC while a weight of 0.0 (A) and -0.25 (B) is applied to evolutions performed for 200 generations. In both cases, the individuals are initialized with a number of metabolites randomly varying from 60 to 100. Without applying a size constraint (A) the size of the individuals is ever increasing through the evolution. The application of a weight of -0.25 penalizes the individuals for the addition of metabolites and the individual size is stabilized (B).

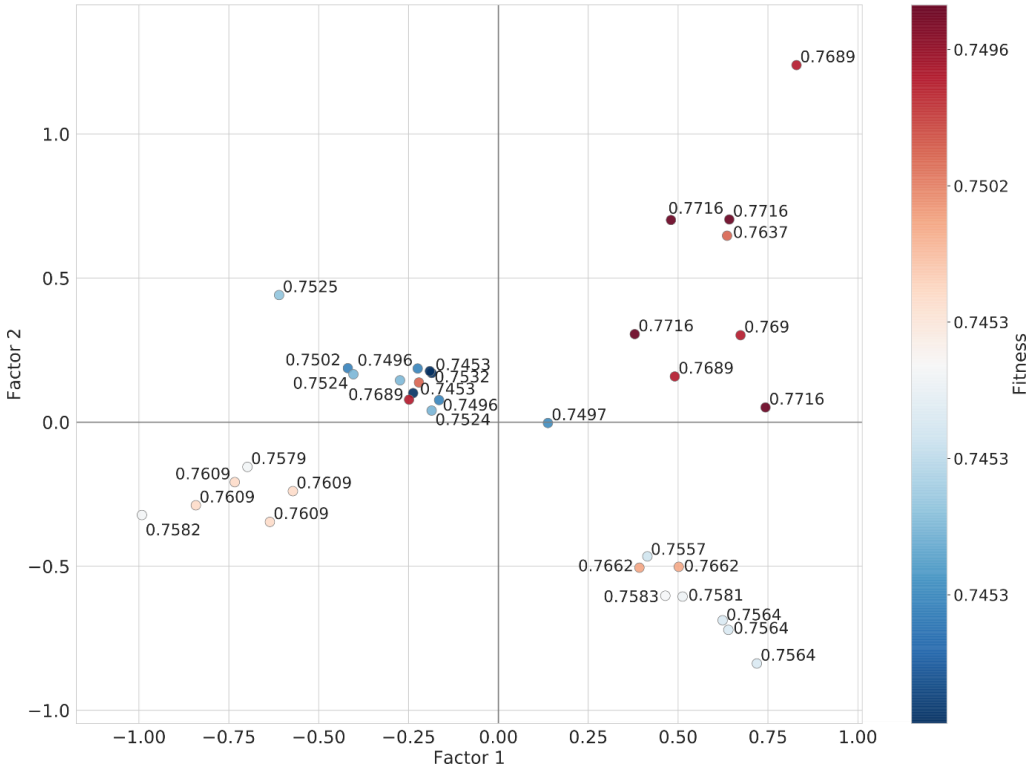

Figure S2.5 Multiple correspondence analysis (MCA) of individuals generated with BOFdat Step 3. MCA is used to compare feature similarity between 35 individuals selected from HOFs from different evolutions. The fitness value is represented from the minimum value recorded (0.7453, blue) to the maximum (0.7716, red). The percent variability explained by the first two factors is 19.54% for factor 1 and 13.92% for factor 2. This analysis shows that similar fitness values can be obtained from diverse sets of metabolites.

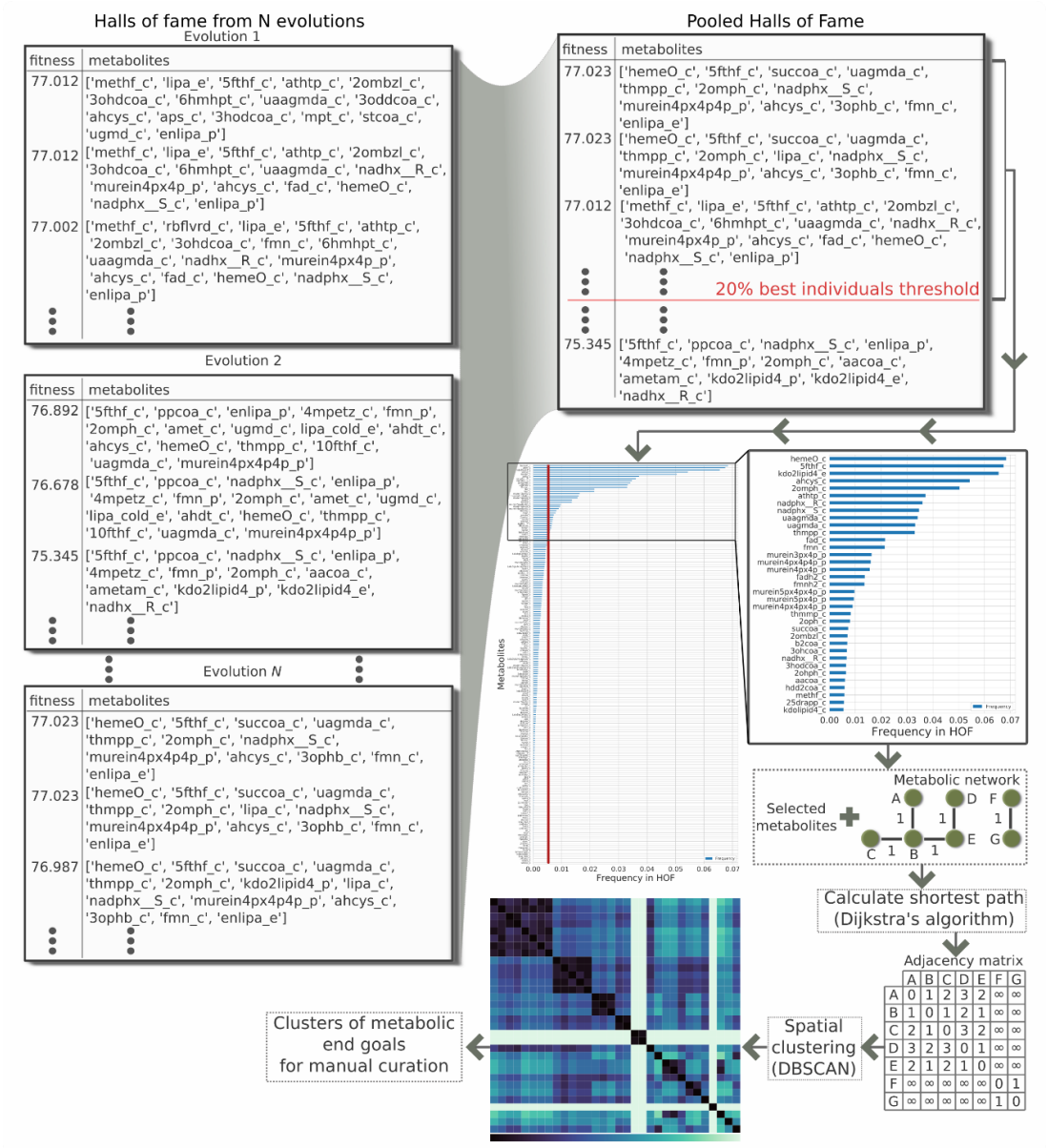

Figure S2.6 Schematic description of spatial clustering in BOFdat Step 3. Each evolution generates a Hall of Fame (HOF) that, by default, stores the 1000 best individuals generated in the course of that evolution. The individuals produced by all the evolutions are pooled together and, arbitrarily, the 20% best individuals are selected based on their fitness value. The metabolites that compose these individuals are then ranked based on their frequency of apparition in those individuals. The metabolites with a frequency of apparition above average are selected. The Dijkstra algorithm is then used to calculate the distance in the metabolic network between the selected metabolites, assuming that a single reaction is one unit of distance. The distance matrix is then clustered using DBSCAN to identify clusters of metabolic

end goals. The removal of highly connected metabolites may break the connection of certain metabolites to the rest of the network. For visualization purposes, the maximum distance is attributed to such metabolites.

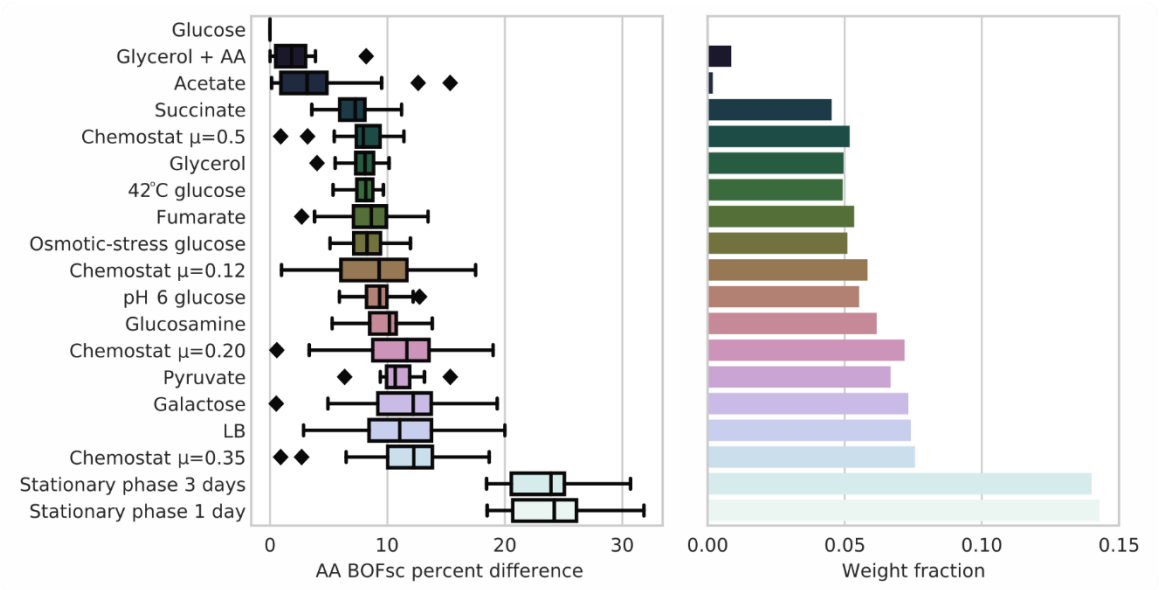

Figure S2.7 BOFdat Step 1 allows calculating stoichiometric coefficients under different experimental conditions. A high-quality quantitative proteomic dataset was used to generate stoichiometric coefficients for 18 different growth conditions [35]. Each boxplot represents the BOFsc percent difference of each amino acid (AA) in a given condition compare to glucose (left). The dataset also includes the macromolecular weight fraction (MWF) of protein in the cell for each condition. The histograms represent the difference between the protein weight fraction of each condition from the same dataset, against the fraction determined in the glucose growth condition (right). This shows that the stoichiometric coefficients are mainly affected by the MWF.

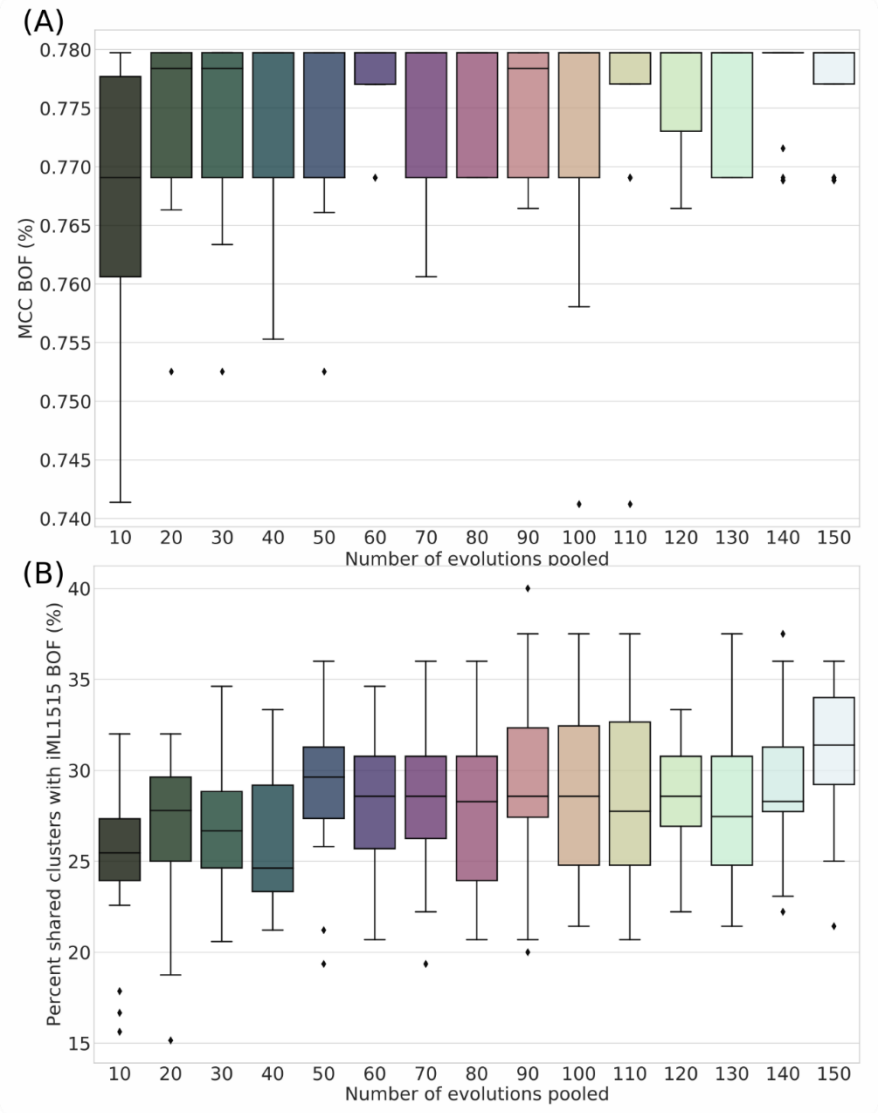

Figure S2.8 Impact of the number of evolutions on the clustering results. HOFs from the 150 evolutions over 500 generations are sampled, such that 20 samples are formed per number of evolutions pooled together (x-axis). The number of evolutions is the number of HOF pooled together to form clusters and select metabolites. (A) The MCC value of the automatic selection of metabolites following the clustering in BOFdat Step 3 for an increasing number of HOF from different evolutions. The median MCC value exceeds the baseline set by *i*ML1515 for 20 evolutions pooled together and reaches a maximum of 0.779 after 40 evolutions. (B) The selected metabolites are compared against the original biomass of *i*ML1515. New clusters are formed and the distribution of percentage of shared clusters is shown (y-axis). The use of more evolution yields an increase in the percentage of shared clusters with the original BOF from 26.61% for 10 evolutions to 31.19% at 150 evolutions.

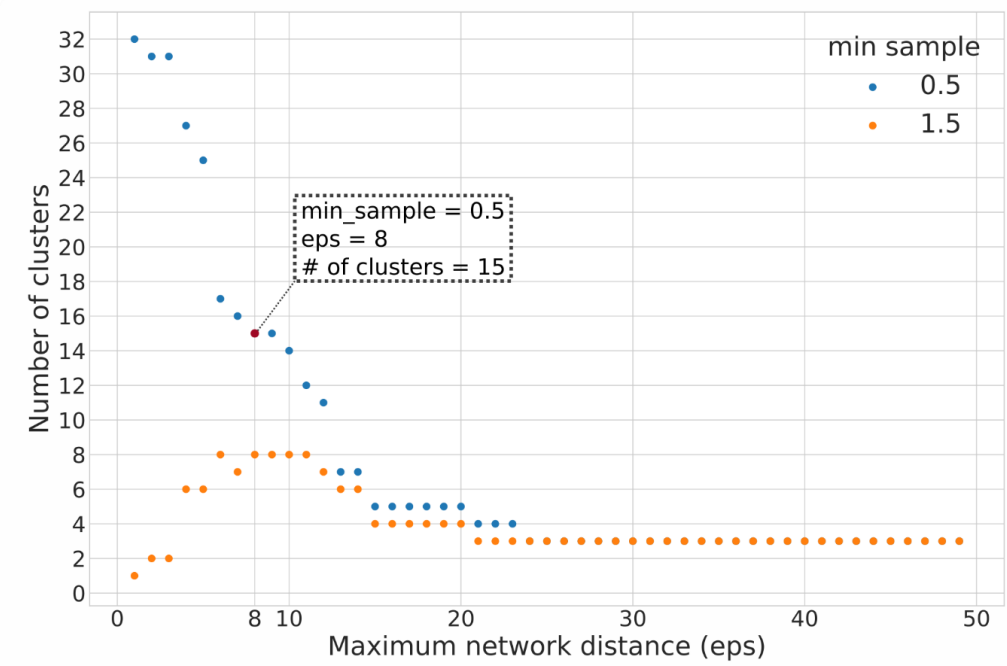

Figure S2.9 Impact of hyperparameters on the clustering results. Selected metabolites from the 20% best individuals generated over 150 generations were clustered using the DBSCAN algorithm while varying the hyperparameters (eps and min_sample). As shown in the legend, two different values of min_sample were used (0.5 in blue, and 1.5 in orange). For these two min_sample values, the eps was varied from 1 to 50 and the number of clusters generated for these hyperparameters recorded (y-axis). Varying the "min_sample" for values 0 to 1 generated a single curve, while using a min_sample above 1 generated another. Hence only two values were represented. In this study, we systematically used a min_sample value inferior or equal to 1 (blue curve) and eps values that yielded a number of clusters around half of the number of metabolites provided.

2.14 SUPPLEMENTARY FILES

Supplementary files are available online:

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006971

S1 File. SEED metabolite identifiers of the gram-negative bacterial BOF converted to BiGG identifiers.

S2 File. Description of clusters for Fig 2.5C.

S1 Text. Supplementary methods and implementation details.

CHAPITRE 3
GENOME-SCALE METABOLIC MODELING REVEALS KEY FEATURES OF A MINIMAL GENE SET

3.1 CONTEXTE

L'une des promesses de la biologie synthétique est de faire de la biologie moléculaire une discipline d'ingénierie prédictive (Lachance et al., 2019a), un exploit qui requiert des connaissances biologiques larges et détaillées. Atteindre ce degré de précision sera probablement plus accessible dans des cellules minimales qui contiennent un faible nombre de gènes de fonction inconnue. Même dans ces organismes très simples, la consolidation de toutes les connaissances dans un format informatique unique capable de faire des prédictions phénotypiques sera essentielle pour concevoir des génomes viables en pratique. Largement utilisés pour générer des prédictions phénotypiques à partir d'informations génotypiques, les modèles métaboliques à l'échelle du génome (GEM) ont des applications allant de la découverte au génie métabolique et peuvent servir les objectifs de la biologie synthétique (Monk et al., 2014).

Mesoplasma florum est un organisme non pathogène à croissance rapide et quasi minimal, pour lequel des techniques de génie génétiques telles que le clonage et la transplantation de génome entier ont été développées (Baby et al., 2018b; Matteau et al., 2017). Dans ce manuscrit, nous décrivons *i*JL208, le premier GEM pour *M. florum*. Pour formuler la BOF et valider les prédictions phénotypiques du modèle, nous nous sommes appuyés sur une étude de caractérisation intégrative détaillée des paramètres physico-chimiques de la cellule (Matteau et al., 2020). Nous avons également développé spécifiquement un milieu semi-défini qui a permis la mesure des taux d'absorption et de sécrétion de métabolites à partir du milieu ainsi que la validation de l'utilisation des différents sucres par *M. florum*. La forte dépendance aux

données expérimentales est une caractéristique de cette étude qui distingue *i*JL208 des modèles métaboliques des mollicutes publiés précédemment (Bautista et al., 2013; Breuer et al., 2019; Suthers et al., 2009b; Wodke et al., 2013).

*i*JL208 a également été utilisé pour prédire un génome minimal pour *M. florum*. Une étude similaire a prédit l'ensemble de gènes minimal de *Mycoplasma genitalium* (Rees-Garbutt et al., 2020). Cependant, le nombre de gènes dans ces prédictions (360 et 380, respectivement) était fortement inférieur au nombre contenu dans JCVI-syn3.0 (473) (Hutchison et al., 2016), ce qui suggère que de telles conceptions de génome entraîneraient probablement des cellules non viables. Heureusement, le nombre de protéines de *M. florum* ayant des homologues dans JCVI-syn3.0 est assez élevé (409/680) et a permis de comparer notre prévision au seul ensemble minimal de gènes validé expérimentalement disponible à ce jour. Notre génome prédit de 552 kbp contenait 535 gènes codant pour des protéines et en partageait 343 avec JCVI-syn3.0. Alors que la majorité des protéines supprimées (76/145) étaient spécifiques à *M. florum*, 37 d'entre elles avaient un homologue dans JCVI-syn3.0, ce qui suggère que d'autres ensembles de gènes minimaux existent probablement mais conservent de nombreuses caractéristiques clés.

3.2 CONTRIBUTION DES AUTEURS

Ce projet est le fruit d'une collaboration entre les laboratoires des Pr Sébastien Rodrigue et Pierre-Étienne Jacques de l'Université de Sherbrooke et des Pr Adam M. Feist et Bernhard O. Palsson de l'Université de Californie à San Diego. Dans ce projet, j'ai utilisé des approches bio-informatiques afin d'évaluer l'annotation du génome de *M. florum* avant de reconstruire son réseau métabolique complet. À cette étape, l'expertise du Dr Nathan Mih en analyse de structure tridimensionnelle de protéines a été mis à contribution avec l'utilisation des logiciels ssbio et I-TASSER. J'ai ensuite reconstruit l'ensemble du réseau métabolique de *M. florum* en révisant minutieusement les données obtenues, permettant de générer un modèle fonctionnel. Les connaissances des Dr Jonathan M. Monk et Colton J. Lloyd ont été particulièrement utile

à cette étape afin de vérifier que les décisions relatives au choix de réactions soient cohérentes avec l'approche de modélisation par analyse de flux à l'équilibre. La contribution de Joëlle Brodeur a permis de générer un milieu expérimental semi-défini. La définition de ce milieu a été complétée par le Dr Dominick Matteau qui a aussi généré l'ensemble des tests de croissance sur différents sucres et les données de chromatographie liquide à haute-performance (HPLC). Ces données m'ont permis de déterminer les taux de consommation de sucrose et de production de lactate et acétate, que j'ai par la suite utilisé comme contraintes sur le modèle. La conformité du modèle avec la base de données BiGG a été confirmée par le Dr Zachary King. Au moment de l'écriture et des corrections du manuscrit, la contribution conjointe du Dr Dominick Matteau et des Pr Pierre-Étienne Jacques et Sébastien Rodrigue a été monumentale.

Référence bibliographique: Lachance, J.-C., Matteau, D., Brodeur, J., Lloyd, C.J., Mih, N., King, Z.A., Knight, T.F., Feist, A.M., Monk, J.M., Palsson, B.O., Jacques, P.-E, Rodrigue, S. Genome-scale metabolic modeling reveals key features of a minimal gene set. Soumis au journal Molecular Systems Biology [MSB-2020-10099].

3.3 TITLE PAGE

Genome-scale metabolic modeling reveals key features of a minimal gene set

Jean-Christophe Lachance[1], Dominick Matteau[1], Joëlle Brodeur[1], Colton J. Lloyd[2], Nathan Mih[2], Zachary A. King[2], Tom F. Knight[3], Adam M. Feist[2,5], Jonathan M. Monk[2], Bernhard O. Palsson[2,4,5,6], Pierre-Étienne Jacques[1] and Sébastien Rodrigue[1*]

1. Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada
2. Department of Bioengineering, University of California, San Diego, La Jolla, USA
3. Ginkgo Bioworks, Boston, Massachusetts, USA
4. Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, USA
5. Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA
6. Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Building 220, 2800 Kongens, Lyngby, Denmark

*Corresponding author

Sébastien Rodrigue: sebastien.rodrigue@usherbrooke.ca

3.4 ABSTRACT

Mesoplasma florum, a fast-growing near-minimal organism, is a compelling model to explore rational genome designs. Using sequence and structural homology, the set of metabolic functions its genome encodes was identified, allowing the reconstruction of a metabolic network representing ~30% of its protein-coding genes. Growth medium simplification enabled substrate uptake and product secretion rates quantification which, along with experimental biomass composition, were integrated as species-specific constraints to produce the functional *i*JL208 genome-scale model (GEM) of metabolism. Genome-wide expression and essentiality datasets as well as growth data on various carbohydrates were used to validate and refine *i*JL208. Discrepancies between model predictions and observations were mechanistically explained using protein structures and network analysis. *i*JL208 was also used to propose an *in silico* reduced genome. Comparing this prediction to the minimal cell JCVI-syn3.0 and its parent JCVI-syn1.0 revealed key features of a minimal gene set. *i*JL208 is a stepping-stone towards model-driven whole-genome engineering.

3.5 INTRODUCTION

The increased efficiency of *in vitro* DNA synthesis and assembly methods [1] have enabled the development of organisms living with either large fractions of or completely synthetic genomes [2–4]. The capability to physically write entire chromosomes from synthetic DNA is an outset for synthetic genomics, but the ability to predict whether or not this assembly will produce a viable cell remains a substantial challenge. This difficulty is linked to the inherent complexity of living organisms and the incomplete knowledge of the molecular functions they entail [5].

Minimal cells are simple organisms containing the fewest number of genes necessary to support self-replicating life [6]. The number of unknown molecular functions within these small genomes is proportional to their size [7], which makes them especially amenable to the

exhaustive characterization of their content. JCVI-syn3.0A, a working approximation of a minimal cell, was recently reported to contain only 91 proteins of unknown function [8]. This number was considerably higher for the phylogenetically related *Mycoplasma pneumoniae* (311 unknowns) and even higher in the model organism *Escherichia coli* (1,780 unknowns) [8].

Addressing the lack of knowledge in an organism can be aided by a computational framework [9] and could lead to a complete understanding of its molecular functions, an important milestone for reliable biological engineering [5,10]. Furthermore, such computational frameworks can be used directly for the prediction and design of minimal gene sets [11,12]. The development of a genome-scale model (GEM) of metabolism, which details all known metabolic reactions catalyzed by an organism in a reaction matrix, represents a promising strategy to face this challenge [13,14]. GEMs have been previously produced for other naturally occurring and synthetic minimal cells from the mollicutes' phylogenetic group [8,15–17] but not for the fast-growing and non-pathogenic *Mesoplasma florum*. Using this mathematically structured knowledgebase, key phenotypic predictions (e.g. gene essentiality, metabolic flux states, growth medium requirements) can be obtained from the genotype without the need for precise enzyme kinetics data [13,14].

Here we present *i*JL208, the first GEM for *M. florum*. The 208 genes in the model account for ~30% of the total gene count in the genome. We thoroughly investigated and reviewed the genome annotation using a combination of computational approaches, resulting in a metabolic network composed of 372 reactions. A recent deep characterization study of *M. florum* [18] was leveraged to define a species-specific biomass composition. A novel semi-defined growth medium was developed, enabling the identification of the main energy sources that can be metabolized by *M. florum*. Both substrate uptake and product secretion rates were determined in this medium, allowing the definition of constraints on the model. Flux-state and gene essentiality predictions were validated against genome-wide expression and essentiality datasets, reaching an accuracy of ~78% and ~77%, respectively.

Finally, we took advantage of the phylogenetic proximity of *M. florum* to the minimal cell *Mycoplasma mycoides* JCVI-syn3.0 [3] to assess the predictive power of GEMs for the design of minimal genomes. We previously reported that an alternate minimal gene set was likely for *M. florum* [19], which motivated our model-driven search. Given that whole-genome cloning and transplantation techniques were developed for *M. florum* [20,21], minimal genome designs could be put to the test imminently. This contrasts with other mycoplasma for which predictions were made [12] but genetic engineering techniques remain unavailable. The experimentally validated *i*JL208 model was therefore used to formulate a minimal genome prediction that also accounts for both transcription unit architecture and genome-wide essentiality.

3.6 RESULTS

3.6.1 Identification of protein molecular functions in *M. florum*

The reconstruction of a high-quality GEM for *M. florum* requires a comprehensive identification of the molecular functions encoded in its genome. We used a combination of three different computational approaches relying on both sequence and structural homology to review the annotation of all open reading frames (Supplementary text). Proteome comparison (Figure 3.1A and B), structural homology (Figure 3.1C), and the probabilistic identification of enzyme commission (EC) numbers (Figure 3.1D) were combined to define a final annotation score for each of the 676 *M. florum* predicted proteins (Figure 3.1E, Material and methods). Basic, medium, and high confidence levels could be attributed to 258, 292, and 126 proteins, respectively (Figure 3.1F).

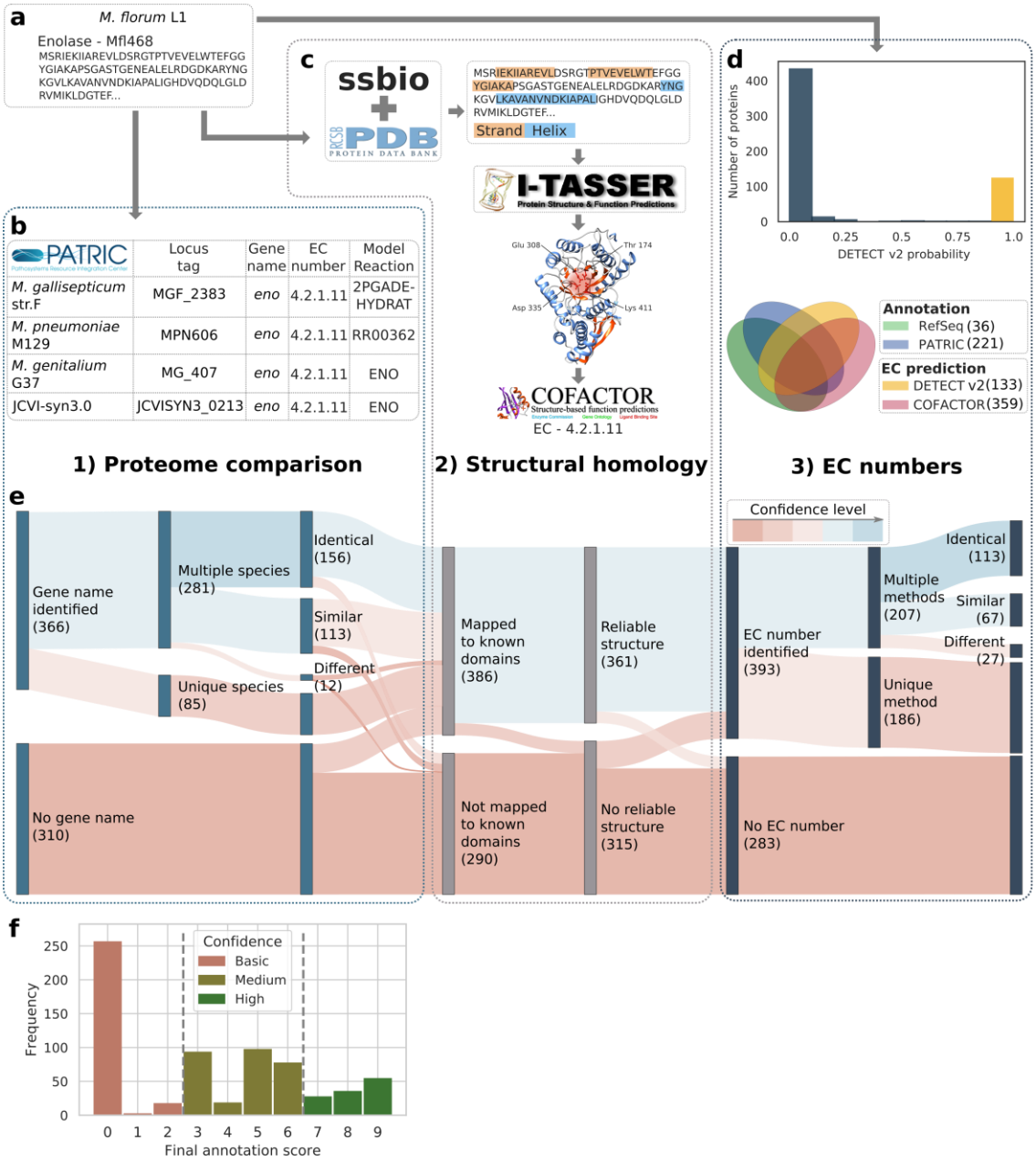

Figure 3.1 Computational identification of molecular functions in *M. florum*. An example of the characterization process is provided with the enolase gene. **a** Predicted amino acid sequence of the *M. florum* enolase (Mfl468). **b** The PATRIC proteome comparison tool allowed the identification of orthologs in the four mollicutes species for which a metabolic model was available, including the reactions to which they were associated in other models as well as their gene names and/or enzyme commission (EC) numbers. **c** The ssbio software was used to search for known protein domains in the amino acid sequence of proteins using the Protein DataBank (PDB) as a reference. If any domain were detected, the I-TASSER suite was

used to generate a tridimensional model of the protein, which was in turn used to obtain an EC number prediction with COFACTOR. **d** Gold standard EC number identifications (yellow bar) were found using DETECT v2 which provides the likelihood of correct annotation (top). These predictions were compared to those from RefSeq, PATRIC, and COFACTOR (bottom). **e** Sankey diagram presenting the sequential steps used to interrogate the 676 predicted coding sequences. The level of confidence specific to each approach was determined and is represented by the red (low) to blue (high) color gradient. **f** Distribution of the final annotation score for the 676 predicted *M. florum* protein-coding genes. Based on this score, a basic (< 3; red), medium (>= 3 and < 7; kaki) or high (>= 7; green) confidence level could be attributed to each predicted protein.

3.6.2 Genome-scale metabolic network reconstruction

Public databases were queried to identify the reactions associated with gene annotations included in the reconstruction [22–26]. To ensure consistency between the identified reactions and available knowledge on mollicutes metabolism, an extensive literature search was also conducted (Supplementary text). The small size of the *M. florum* genome allowed for a manual curation of the putative function for every gene. The resulting metabolic reconstruction, *i*JL208, contains 208 protein-coding genes, 370 reactions, and 351 metabolites, a count similar to other mollicutes models [8,15–17] (Figure 3.2, Tableau 3.1, Supplementary files 4 and 6).

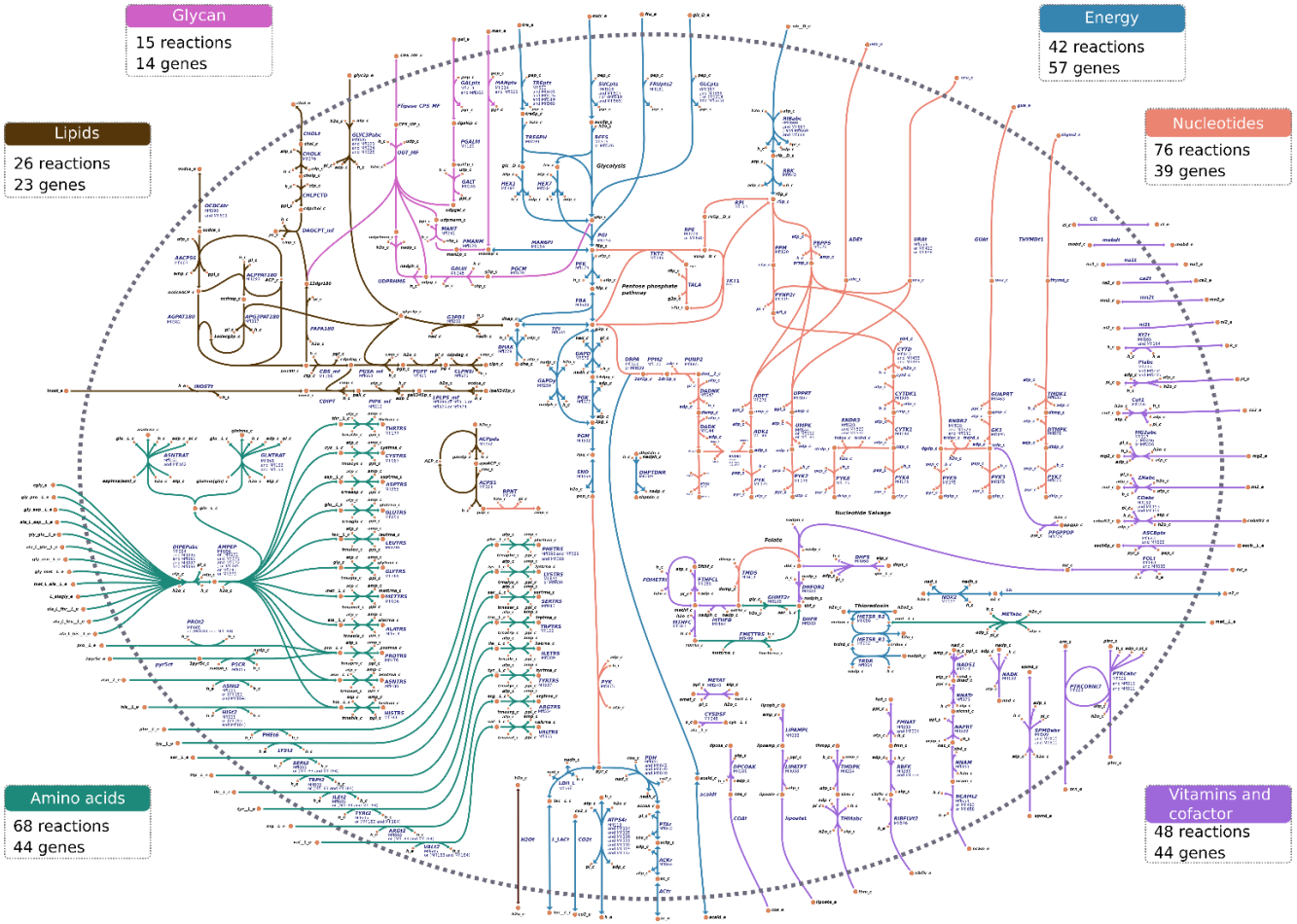

Figure 3.2: Map of the genome-scale metabolic network of *M. florum*. Circles represent metabolites and connecting lines indicate metabolic reactions. Metabolite names are indicated in black, while reaction names and associated gene names are in dark blue. Reaction directionality is represented by arrows. The dotted line shows the cell membrane, with transport reactions linking the intracellular milieu with the extracellular environment. The reactions are color-coded according to the 6 main modules. The outer colored boxes describe the number of genes and reactions. An interactive Escher [67] version of this map is available (Supplementary file 5).

Tableau 3.1 Number of protein-coding genes, reactions, and metabolites in mollicutes metabolic models.

Species	Model	Genes: Model/Total (%)	Total reactions in model	Reactions: shared with *M. florum* (%)*	Total metabolites in model
M. genitalium	*i*PS189	126/507 (24.9%)	351	79/174 (45.4%)	324
M. pneumoniae	*i*JW145	145/691 (20.1%)	306	74/156 (47.4%)	346
M. gallisepticum	N/A**	198/747 (26.5%)	322	83/260 (31.9%)	444
JCVI-syn3.0A	N/A**	155/473 (32.8%)	338	87/338 (25.7%)	304
M. florum	*i*JL208	208/680 (30.6%)	370	---	351

The number of genes and reactions shared with *M. florum* is represented for each model.
*The percentage of shared reactions applies only to the gene-associated reactions.
**No model name was provided by the authors.

Overall, 236 of the 370 reactions are gene-associated in *i*JL208 (Figure 3.2). Of those, 156 reactions are linked to a single gene while 80 are linked to more than one (enzyme complex or isozymes). Of the 134 orphan reactions, 95 are pseudo-reactions (exchange, demand, ATP maintenance, and biomass reactions) while 39 are necessary orphans (i.e. CO_2 transport). Notably, about a third (84/278) of the total number of reactions (excluding pseudo-reactions) in the model are transport reactions.

*i*JL208 reactions were grouped into six different modules: Energy, Amino acids, Lipids, Glycan, Nucleotides, and Vitamins & Cofactors (Figure 3.2). An extensive description of the composition of each module as well as the mechanistic description of all reactions included in the model is provided in the Supplementary text. Each module describes a general metabolic objective and contains between 15 (Glycan) and 76 (Nucleotides) reactions, and 14 (Glycan) to 57 (Energy) genes.

The extent of missing knowledge in each module was estimated using the number of orphan reactions required to generate a functional model. Energy, Nucleotides, and Amino acids modules had the lowest proportion of orphan reactions (<14%, Figure 3.3A) and described well-known aspects of the metabolism of mollicutes. Conversely, the Glycan, the Lipids, and the Vitamins & Cofactors modules had the highest percentage of orphans relative to the total number of reactions in the module (between 25 and 33%). In proportion to their total number of genes, these modules also displayed a lower gene annotation confidence level than the other three (Figure 3.3B).

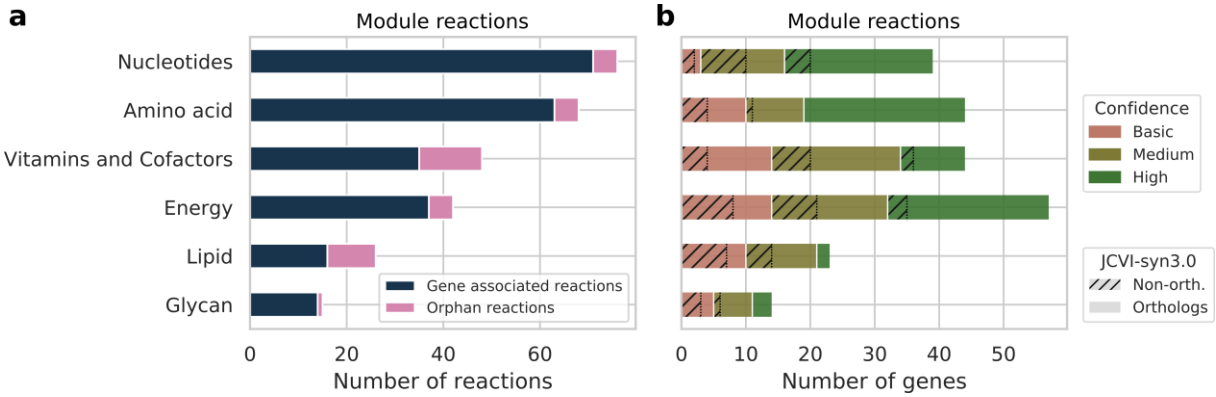

Figure 3.3 Characteristics of the *M. florum* metabolism as revealed by the genome-scale network reconstruction. a Distribution of gene-associated and orphan reactions in the 6 modules defined in *i*JL208. **b** Distribution of genes and their associated final annotation score (see figure 3.1F) in the 6 metabolic modules. JCVI-syn3.0 orthologs are plain while the non-orthologs are hatched.

Interestingly, about 70% (146/208) of the protein-coding genes included in the metabolic reconstruction were orthologous to JCVI-syn3.0 (Figure 3.3B), a slightly higher fraction than the number of orthologs present in the entire genome (~60%, 411/685) (Figure S3.1). The Glycan and Amino acids modules were more conserved than the average for the entire model, while Energy, Vitamins & Cofactors, and Nucleotides had a distribution similar to the model. The Lipids module was the least conserved with 52% of orthologs.

3.6.3 Medium simplification and growth kinetics

While the genomic complexity of mollicutes is remarkably low [27], the number of necessary medium components to sustain their growth is rather high [28]. The metabolic reconstruction reflects this reality with 84 extracellular metabolites associated with transport reactions defining the complete *in silico* medium (Figure 3.2, Supplementary text). We sought to elaborate a simplified growth medium for *M. florum* which, before this study, was commonly grown in the complex and undefined ATCC 1161 medium containing horse serum (HS), yeast extract (YE), and heart infusion broth (Materials and methods).

A particular problem faced when using the ATCC 1161 medium was apparent growth when no sucrose was added (Figure 3.4A, top), which prevented the assessment of *M. florum*'s metabolic capabilities when supplemented with different carbohydrates. To circumvent this issue, the concentrations of HS and YE were lowered by adding a completely defined rich medium base to the mixture (CMRL 1066). This allowed a 64-fold reduction in the concentrations of HS (to 0.313%) and YE (to 0.02%) required for significant growth (Figure S3.3), as well as the complete removal of heart infusion broth. This CMRL 1066-based semi-defined medium, referred to as CSY, allows visible growth only when sucrose is added (Figure 3.4A, bottom).

We observed that reducing the concentration of HS and YE impacted the doubling time of *M. florum* (Figure 3.4B), suggesting that nutrients contained in these undefined components are rate-limiting in *M. florum*. When varying the initial sucrose concentrations in CSY medium, the total biomass produced followed an asymptotic behavior (Figure 3.4C), with a predicted maximum concentration of 5.95 x 1e8 colony forming unit per ml (CFU/ml), corresponding to 0.013 grams of dry weight per liter (gDW/L). Given that *M. florum* cell densities typically reach ~1e10 in ATCC 1161 medium[18] , these observations confirmed that nutrients other than sucrose are rate-limiting in CSY.

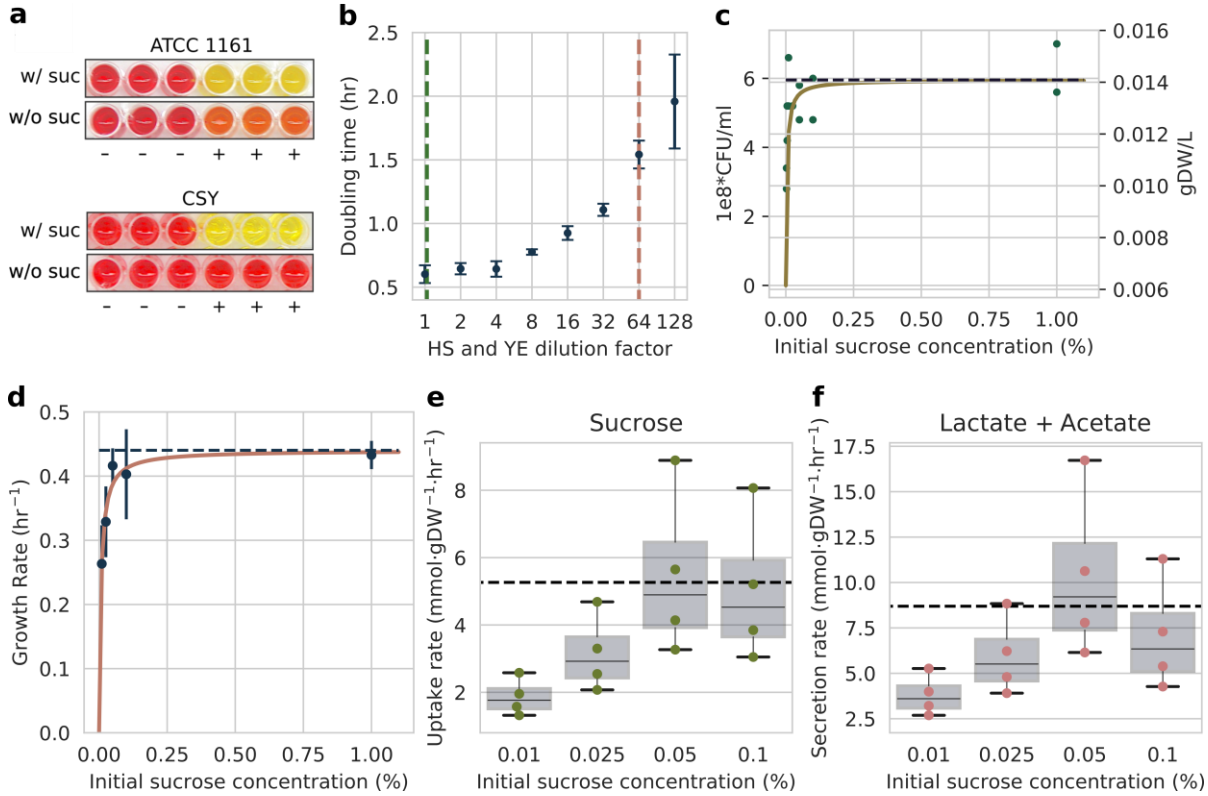

Figure 3.4 Impact of medium composition on *M. florum* growth kinetics. **a** Bacterial growth assessed by color change due to medium acidification for both the undefined ATCC 1161 and semi-defined CSY media with or without sucrose, 4% and 1% respectively. (+), Inoculated wells; (-), non-inoculated control. The picture was taken after a 24 hour incubation period. **b** Impact of horse serum (HS) and yeast extract (YE) dilution on *M. florum* doubling time. The green and red dotted lines indicate HS and YE concentrations found in ATCC 1161 medium (20% HS, 1.35% YE) and CSY medium (0.313% HS, 0.02% YE), respectively. Doubling times were measured in a CMRL 1066 base medium using colorimetric assays. Dots and error bars indicate the mean and standard deviation calculated from three biological replicates. **c** The maximal biomass concentration observed for *M. florum* cultures growing in CSY medium with varying initial concentrations of sucrose. Biomass was measured using colony forming units (CFU/ml; left axis), and converted to grams of dry weight (gDW/L; right axis). A rectangular hyperbola fit is shown (yellow), and the dotted line represents the maximal biomass value predicted by the fit (1e8 x 5.95 CFU/ml; 0.013 gDW/L). **d** Relationship between varying initial sucrose concentration and growth rate in CSY medium. Growth rates were determined by fitting a simple exponential growth model to CFU/ml data from time-course experiments (see Figure S4 and 5) and error bars indicate the standard deviation associated with each value. A rectangular hyperbola fit is shown, and the predicted maximal growth rate (0.44 hr^{-1}) is indicated by the dotted line. **e** Sucrose specific uptake rate and **f** combined lactate/acetate specific secretion rates at varying initial sucrose concentrations in CSY medium. Boxplots represent the median and interquartile range of uptake or secretion rate values calculated at different time intervals during the exponential growth phase of *M. florum*

cultures. Dotted lines indicate the selected uptake (-5.26 mmol gDW^{-1} h^{-1}) and secretion rate (8.6 mmol gDW^{-1} h^{-1}) values used for modeling.

Nevertheless, at low sucrose concentrations, initial sucrose and growth rate displayed a Monod-like relation [29], with a maximal growth rate in CSY found at 0.44 hr^{-1} (Figure 3.4D, Figure S3.4 and 3.5A). We used this range of dependency between the growth rate and the initial sucrose concentrations to define substrate uptake and by-product secretion rates. The sucrose and combined lactate/acetate variation of concentration over time were measured by high-performance liquid chromatography (HPLC). Substrate-specific uptake rate (sucrose) and metabolic by-product secretion rates (lactate/acetate) were calculated using linear regression in the exponential growth phase (Material and methods, Figure S3.5 and 3.6). The range of possible rates within the exponential phase was calculated for each initial sucrose concentration (Figure 3.4E and 3.4F). Interestingly, a tendency towards both a maximum uptake and secretion rate (qS_{max}) could be observed, which is desired for modeling purposes where the optimal conditions are assumed. Accounting for experimental variability and the number of data points available, the maximum sucrose uptake rate was set at -5.26 mmol gDW^{-1} h^{-1} and the combined lactate/acetate secretion rate at 8.69 mmol gDW^{-1} h^{-1} (see Material and methods).

3.6.4 Conversion into a mathematical format and sensitivity analysis

The biomass objective function (BOF) is a reaction of the stoichiometric matrix used to simulate an organism's growth [30]. Previously reported experimental macromolecular composition characterizing 98.8% of *M. florum*'s dry mass, as well as multiple omics datasets, [18] were used as input into the BOFdat software [31] to define the *M. florum* specific BOF (Material and methods, Supplementary text). DNA, RNA, and protein stoichiometric coefficients were determined by the first step of BOFdat and accounted for 76.4% of the total cellular dry weight [18] (Figure 3.5A, left). Coenzymes and inorganic ions were next identified, finding 12 metabolites previously defined as universally essential cofactors in prokaryotes [32],

as well as seven other metabolites with high connectivities (Figure 3.5A, middle). These 19 metabolites were considered as the soluble pool and their stoichiometric coefficients were determined using the remaining 1.2% *M. florum* biomass.

Using the Step3 of BOFdat, the correspondence between single-gene essentiality prediction and genome-wide transposon mutagenesis data [19] was improved by the addition of nine metabolites, two of which were also identified in previously published lipidomic data [18] (Figure 3.5A, right; Figure S3.7, Material and methods). A metabolite corresponding to the *M. florum* capsular polysaccharide (CPS) was also added during Step3. In the pathways leading to the production of the four most frequently identified metabolites in Step3, 10 out of 15 genes had their essentiality prediction modified compared to Step2 (Figure 3.5B). Among these 10, a single gene was wrongfully identified as essential compared to transposon mutagenesis data.

The model was then constrained using the experimental rates defined above (Figure 3.4D, E, and F). Given the complexity of the growth medium, growth and non-growth associated maintenance costs [33] (GAM and NGAM, respectively) could not be obtained directly and had to be inferred from known parameters. To simulate growth on CSY, the *in silico* minimal medium was defined using the COBRApy toolbox [34] and a set of key initial parameters were selected (Material and methods). A phenotypic phase plane (PPP) analysis [35] was performed to identify the ATP maintenance value that allowed the model to reproduce the experimentally determined growth rate and define both GAM (Figure 3.5C) and NGAM (Figure 3.5D) values. With these constraints, the model sensitivity to the sucrose uptake rate was assessed, revealing three different growth phases ending with a plateau (Figure 5E). This is similar to the experimental observations that, in CSY, *M. florum*'s growth is not restricted by sucrose availability alone (Figure 3.4).

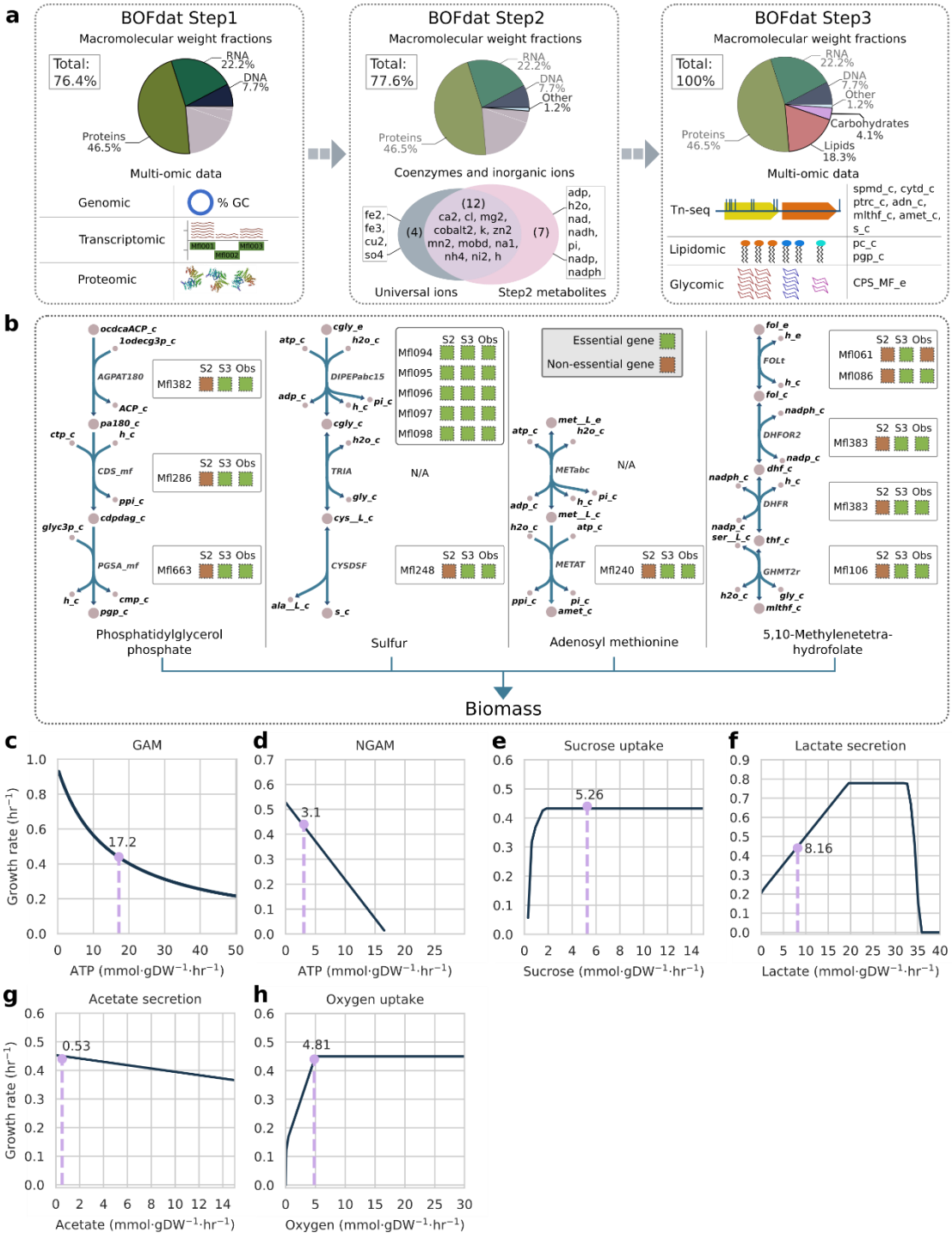

Figure 3.5 Conversion into a mathematical format. a The biomass objective function (BOF) was defined using the BOFdat software [31] and experimental data from Matteau *et al.* [18]. The stoichiometric coefficients for the major cellular macromolecules were generated with BOFdat

Step1 (left), the inorganic ions and coenzymes with BOFdat Step2 (middle) and the remaining components with BOFdat Step3 (right), which identifies the metabolites with the greatest impact on the BOF gene essentiality prediction accuracy. **b** Depiction of the metabolic context of four complete pathways supporting the output of BOFdat Step3. Metabolite names are shown in black and reaction names in grey. Boxes contain the predicted essentiality of the protein-coding genes associated with a given reaction following BOFdat Step2 (S2) and Step3 (S3), along with the observed essentiality (Obs). N/A corresponds to orphan reactions. **c-h** Sensitivity analysis of the *i*JL208 model to growth-associated maintenance (GAM), non-growth associated maintenance (NGAM), sucrose uptake rate, lactate and acetate secretion rates, and oxygen uptake rate. The light purple dotted line represents the corresponding rate at the predicted maximal growth rate in CSY (0.44 hr^{-1}).

The metabolic reconstruction revealed the capability of *M. florum* to produce both lactate and acetate as fermentation products (Figure 2). Since only their combined secretion could be measured by HPLC (Figure S5C, Material and methods), individual production rates had to be inferred. Previously reported differential expression of key enzymes[18], such as the lactate dehydrogenase (LDH, Mfl596), showed an approximately 4 to 8 fold increase in the protein expression levels compared to genes of the pyruvate dehydrogenase complex (PDH, Mfl039, Mfl040, Mfl041, and Mfl042). We thus used an 8:1 initial ratio (lactate:acetate) for further PPP analysis (Material and methods). For lactate production, this analysis revealed a positive linear relationship between the predicted growth rate and lactate secretion rate (Figure 5F). Conversely, the production of acetate was detrimental to the predicted growth rate (Figure 5G), and its secretion rate had to be lowered to match the experimental growth rate in CSY. While both secretion routes generate ATP, acetate production requires oxygen to regenerate the NAD+ pool through the NADH oxidase (NOX2, Mfl037) (Figure 3.2). To simultaneously ensure that *M. florum* was not able to produce oxygen and that the growth rate was not linearly dependent on oxygen uptake, the NOX2 reaction was bounded between 0 and 5 mmol gDW^{-1} h^{-1}, resulting in an optimal oxygen uptake rate intersecting a plateau (Figure 3.5H).

3.6.5 Validation of model phenotypic predictions

The development of the CSY medium enabled an experimental validation of the capability of *M. florum* to grow on 14 different carbohydrates and to compare these observations with the constrained model's phenotypic predictions (Figure 3.6A and Figure S3.8). The growth/no-growth phenotype was correctly predicted by *i*JL208 for 12 out of the 14 sugars tested. The two remaining sugars, maltose and mannose, were used by *M. florum* while the model predicted no growth. To address these discrepancies, the alternate carbon metabolism of *M. florum* was studied, seeking enzymes that would likely carry a promiscuous activity. Particularly, the specificity of three enzymes was challenged using the FATCAT 2.0 server [36] to compare the tridimensional structures reconstructed with I-TASSER in this study to crystallographic structures from the PDB (Figure S3.9, Tableau S3.12).

The similarity between maltose and trehalose suggested that the trehalose hydrolase (Mfl499) could also hydrolyze maltose. This hypothesis was supported by the very high structural similarity between the reconstructed Mfl499 and a *Bacillus sp*. α-glucosidase shown to have a high-specificity for α-(1-4)-glucosidic linkage [37] (Figure 3.6B (i) and Figure S3.9A). Similarly, the capacity of *M. florum* to metabolize mannose could be explained by the capability of the glucose-6-phosphate isomerase (Mfl254) to convert mannose-6-phosphate into fructose-6-phosphate, hereby entering glycolysis (Figure 3.6B (ii) and Figure S3.9B). While the promiscuity of the transporter complexes could not be tested *in silico* (Material and methods), the addition of both the promiscuous transport and digestion reactions were sufficient to provide a growth prediction on maltose and mannose. As reported previously, both glucose and mannose were detected in the *M. florum* polysaccharide layer [18]. The presence of a phosphomannomutase (Mfl120) in the genome annotation suggested the conversion of mannose-6-phosphate to mannose-1-phosphate, a necessary precursor for glycan synthesis [38]. The very high structural similarity between the reconstructed Mfl120 and an enzyme necessary for the production of exopolysaccharides [39] in *Pseudomonas aeruginosa* (Figure 3.6B, (iii); Figure S3.9C) supported this hypothesis.

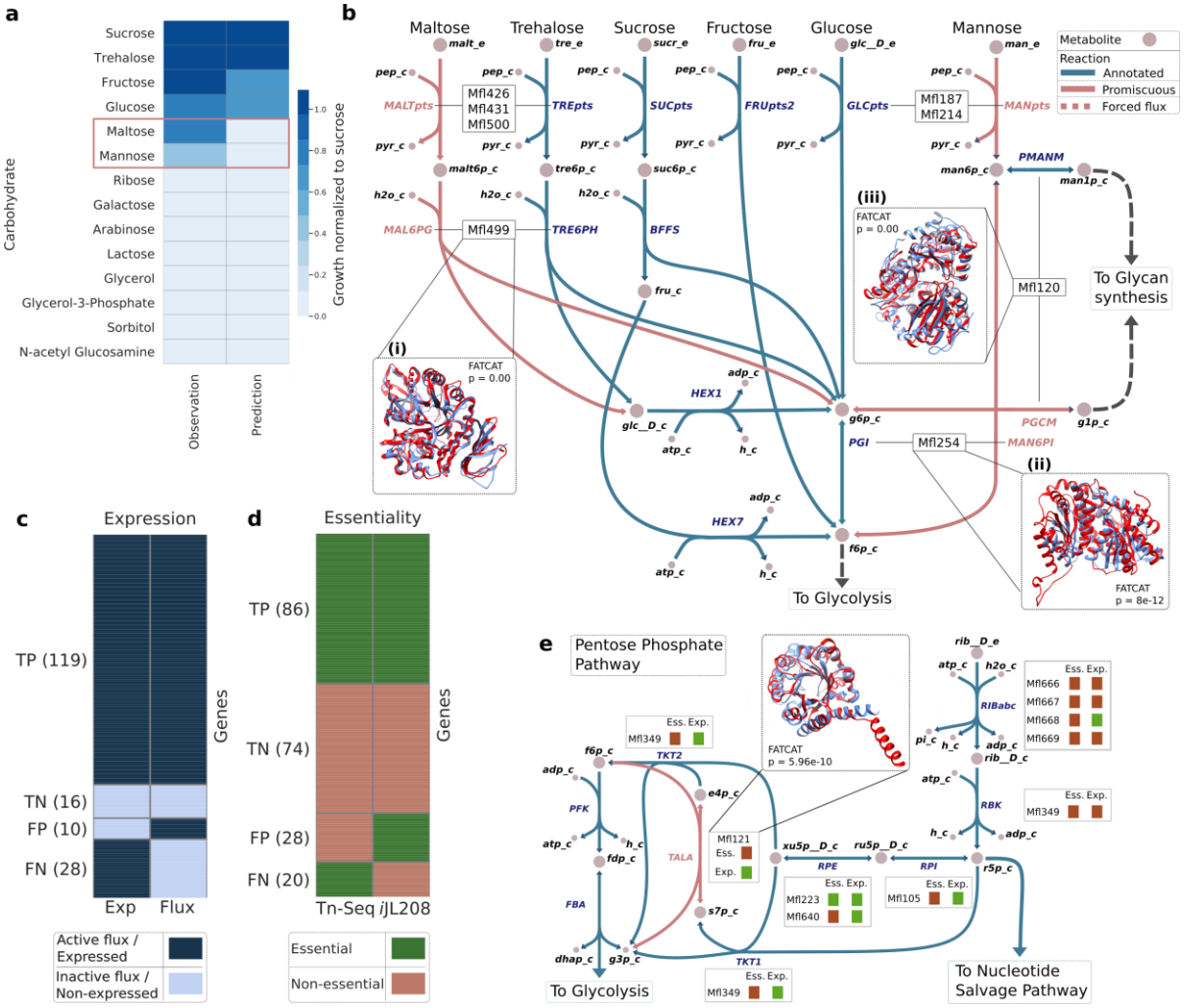

Figure 3.6 Validation of the model's phenotypic predictions. a Growth phenotype observed on CSY medium supplemented with different carbohydrates (1% final concentration) compared to *i*JL208 predictions. Growth observations and predictions were normalized by growth on sucrose. Discrepancies between experimental data and predictions are highlighted (red rectangle). **b** The metabolic network reconstruction allowed the identification of potential candidates carrying promiscuous reactions (pink) responsible for maltose and mannose catabolism. 3D structures of candidates (Mfl499, Mfl120, and Mfl254; red) are superimposed to available structures in the PDB (light blue) for which the suspected enzymatic activity is annotated. FATCAT p-values are shown. **c** The 173 genes where the transcriptomic and proteomic expression data (Exp) are consistent (Figure S3.10) are compared with parsimonious FBA (pFBA) flux states predictions (Flux). True Positives (TP) and True Negatives (TN) correspond to genes where both predictions and observations are consistent, while False Positives (FP) and False Negatives (TN) where they are inconsistent. Genes considered

expressed or with an active flux are represented in dark blue, and silent in light blue. **d** Revised gene essentiality data (Figure S3.11) compared to essentiality predictions from *i*JL208. Essential and non-essential protein-coding genes are represented in green and red, respectively. TP, TN, FP, and TN are as in C. **e** Mfl121 (red) shows structural similarity with a transaldolase (blue) and justified the addition of a TALA reaction in the model.

Genome-wide expression [18] and transposon mutagenesis [19] datasets available for *M. florum* were used as a reference for the validation of model flux states and gene essentiality predictions [40]. From the 208 protein-coding genes present in *i*JL208, 173 showed a consistent expression between transcriptomic and proteomic datasets (Figure S3.10; Material and methods). Of these genes, 135 occurrences had an expression profile in agreement with the metabolic fluxes predicted by the model, which corresponds to an accuracy of 78.03% (Figure 3.6C). In parallel, the original transposon mutagenesis data were reanalyzed, resulting in the re-assignment of 79 coding genes previously considered as non-essential, for a total of 332 *M. florum* genes determined as essential (Figure S3.11, Supplementary file 9, and Material and methods). 160 single-gene essentiality predictions from *i*JL208 were consistent with that revised experimental data, for an overall accuracy of 76.92% (Figure 3.6D).

Essentiality and expression comparison with model predictions provided a context for the refinement of the model. Targeted false negatives, genes simultaneously expressed and essential while no flux or essentiality was predicted in *i*JL208 (Figure 3.6C and D), together with a single false positive were manually curated by the addition of specific constraint(s) (Supplementary text, Figure S3.12). Among those, genes of the pentose phosphate pathway were specifically investigated since they all showed expression but carried no metabolic flux (Figure 3.6E). In *M. florum* and other mollicutes, this pathway is incomplete because no gene is typically attributed to the transaldolase (TALA) reaction [8,15,16,41]. Here, the I-TASSER reconstructed structure of 2-deoxyribose-5-phosphate aldolase (Mfl121) was queried against the PDB to find potential matches with transaldolase structures (Figure S9D). Of the 12 transaldolase identified, the structure from *Thermotoga maritima* had the most significant match and highest similarity (Figure 3.6E and Tableau S3.12). Based on these results, the

TALA reaction was assigned to Mfl121 in *i*JL208, thereby allowing flux through the pentose phosphate pathway, which is consistent with expression data.

3.6.6 Model-driven prediction of a minimal genome

The validated *i*JL208 GEM was used together with experimental gene essentiality and transcription unit (TU) architecture [18] to infer and characterize a minimal gene set for *M. florum* using the MinGenome algorithm [11]. To generate this prediction, MinGenome incorporates both experimental data and model constraints into an optimization problem that finds the largest possible deletion in the genome which, applied iteratively, defines the minimal gene set. Interestingly, a minimum growth rate can be imposed as a constraint on the MinGenome optimization problem. We investigated the impact of a range of imposed growth rates on predicted genome reduction scenarios, and three possible genome reduction scenarios were obtained (Supplementary text). The smallest genome size was identified at the lowest imposed growth rate, corresponding to 562 kbp and 152 deleted coding and non-coding genes, respectively (Figure 3.7A). The resulting genome designs were compared to the gene content of JCVI-syn3.0 [3], which has a high proportion of orthologs in *M. florum* (Figure S3.1) and provides a compelling validation framework for a minimal genome. Interestingly, lowering the growth rate constraint increased the similarity of the predicted minimal gene set to JCVI-syn3.0 (Figure S3.13).

The smallest predicted genome was further analyzed by comparing its protein-coding genes to both JCVI-syn3.0 and its parent strain JCVI-syn1.0 [2]. Of the 145 proteins deleted in this scenario, 37 are present in JCVI-syn3.0, while 32 are JCVI-syn1.0 proteins that were also deleted in the process of generating JCVI-syn3.0. In total, 108 (74%) of these proteins were not in JCVI-syn3.0 (Figure 3.7B). Also, the number of JCVI-syn3.0 proteins with no ortholog in *M. florum* (58) was similar to the number of JCVI-syn1.0 proteins deleted to generate JCVI-syn3.0 and the number of proteins shared with *M. florum* (63).

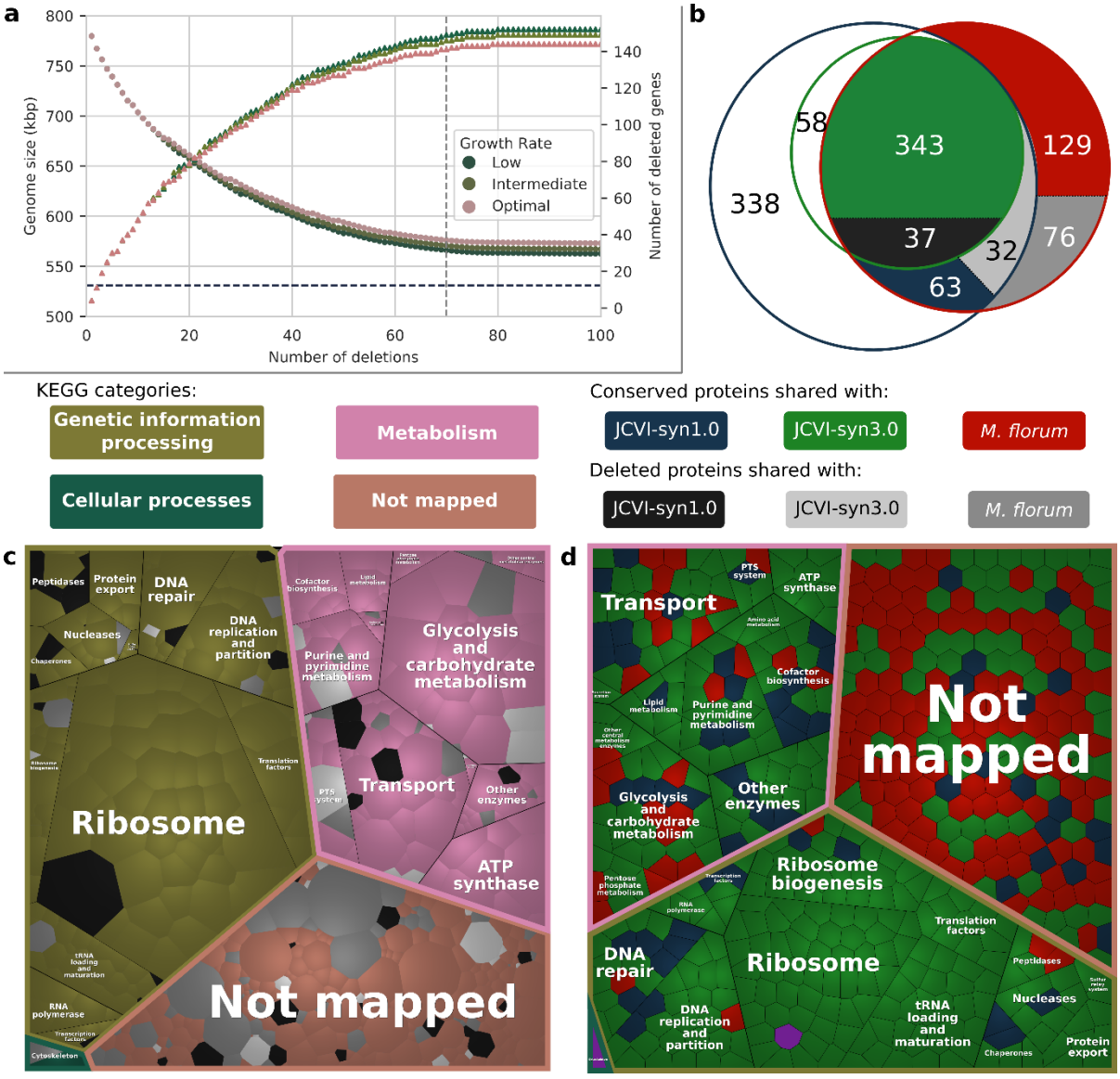

Figure 3.7 Model-driven prediction of a minimal genome for *M. florum*. a Iterative deletions using the MinGenome algorithm. Faded circles represent genome size and bright triangles the number of deleted genes. Low, intermediate, and optimal growth rate constraints correspond to 60, 90, and 100% of the *M. florum* growth rate measured in CSY (0.44 hr^{-1}). **b** Venn diagram showing the proteins shared between *M. florum* (red circle), JCVI-syn3.0 (green circle), and its parent JCVI-syn1.0 (blue circle). The different shades of grey represent the proteins targeted for deletions in the Low growth rate constraint presented in A. **c** Voronoi diagram showing the functional distribution of 145 proteins targeted for deletions present in JCVI-syn3.0 (black), *M. florum* only (dark grey), and JCVI-syn1.0 but not JCVI-syn3.0 (light grey). The shapes are sized according to transcriptomic data and the KEGG categories are

represented by bright colors (NM: not mapped, GIP: genetic information processing, M: metabolism, EIP: environment information processing, CP: cellular processing). **d** Voronoi diagram showing the functional distribution of the conserved proteins in the predicted reduced genome. Colors are as in panel b, purple represent proteins shared with JCVI-syn3.0A.

We further investigated the functional categories of the deleted proteins and combined this information with the reported protein expression level (Figure 3.7C, Figure S3.14A), revealing that eight key cellular functions were simply not hit by any deletions: ATP synthase, Translation factors, RNA polymerase, Protein export, Cofactor biosynthesis, Sulfur relay system, Lipid metabolism, and the Pentose phosphate metabolism. The Ribosome category was nearly untouched except for the highly expressed ribosomal protein L31 (*rpmE*, Mfl648). While functionally important, this protein was categorized as non-essential in our dataset (Supplementary file 9). In *E. coli*, it was also shown that the deletion of both this protein and its paralog would yield a viable cell, albeit one with significant growth defects [42]. Like in *M. florum*, this protein was hit by transposons in JCVI-syn1.0 and caused minimal growth disadvantage [3] but was likely conserved to increase the robustness of JCVI-syn3.0.

The key features retained in the proposed minimal gene set were similarly detailed by mapping their functions to KEGG categories and their homology to JCVI-syn3.0 or JCVI-syn1.0 (Figure 3.7D, Supplementary text). The 535 proteins retained in the minimal gene set were similarly distributed between the three functional categories. Of the 191 proteins that did not map to a KEGG category, the majority (106) were specific to *M. florum* (Figure S3.14B). Three of the 12 sub-categories contained in the Metabolism category (ATP synthase, Amino acid metabolism, Secretion system) were exclusively composed of proteins having orthologs in JCVI-syn3.0 (Supplementary file 10). While all glycolysis enzymes were retained and common with JCVI-syn3.0, enzymes responsible for the assimilation of sucrose in the three sub-categories where they were found (Transport, PTS system, and Glycolysis and carbohydrate metabolism) had no orthologs in JCVI-syn3.0, meaning that energy sources are interchangeable in mollicutes' minimal genomes.

We also observed the absence of the E1 component of the pyruvate dehydrogenase complex in JCVI-syn3.0 [3]. The conservation of these proteins in our minimal genome prediction relies on the forced secretion of acetate in *i*JL208 (Figure 3.5), which best represented the experimental setting. Consistent with our observation that varying the imposed growth rate generated three different genome reduction scenarios, this reveals that alternate minimal genome designs may be obtained by varying the constraints imposed on the input GEM.

Finally, the majority of the retained proteins in the genetic information processing category is composed of 13 sub-categories and have orthologs in JCVI-syn3.0 (174/195). The fewest number of JCVI-syn3.0 orthologs were found in the DNA repair sub-category. The conserved proteins in our prediction entailed recombination proteins, glycosylases, and the DNA polymerase IV. In all cases, a single occurrence was still available in JCVI-syn3.0, meaning that these genome integrity functions should remain in a minimal gene set. Transcription factors involved in the assimilation of fructose were also conserved in our prediction while being absent from JCVI-syn3.0, which is consistent with the observations made for the metabolism that revealed energy sources could be swapped in minimal genomes.

3.7 DISCUSSION

Computer-aided design is crucial for the development of synthetic biology on a large-scale. In genome-writing projects [43], the predictive power of GEMs could be leveraged to reduce the overall design and engineering efforts required to produce a viable strain. Still, the applicability of computational models for genome design is tightly linked to the level of knowledge available for the organism of interest.

In *M. florum*, the level of knowledge was examined by cross-validating the identification of molecular functions from different computational methods, establishing a confidence

hierarchy for protein annotation (Figure 3.1). Using predicted protein functions, we reconstructed the metabolic network of *M. florum*. Overall, *i*JL208 shares many similarities to previously reconstructed mollicutes models (Tableau 3.1), including the requirement of a rich medium as a key feature to support the cell's growth, typically associated with a scavenger lifestyle [44,45]. This metabolic regime is mainly characterized by the abundance of transport reactions (84), the absence of a respiratory system, and the fact that biosynthesis occurs mostly through salvage pathways (Figure 3.2). These elements are closely linked, with glycolysis providing a low ATP yield that is nonetheless sufficient to fuel the import of nutrients from the medium via various PTS systems. *M. florum* then assembles premade molecular building blocks, which considerably lowers the energetic cost of building cellular biomass.

Combining this information with the six modules proposed in the metabolic reconstruction (Figure 3.2) revealed that the Lipids, Glycan, and the Vitamins & cofactors modules had fewer genes identified with scarcer reliable information (Figure 3.3). It is possible that enzymes evolved by mollicutes to integrate lipids in their membranes and assemble glycans into capsular polysaccharides are not very similar to more thoroughly studied proteins in model organisms. Corollary to this hypothesis is the lost ability to synthesize a cell wall, a landmark of mollicute evolution [27]. Therefore, lipid and glycan synthesis is probably performed by currently un-annotated but likely essential enzymes. Additional work will be needed to describe the exact lipid and glycan composition and the genes involved in their processing.

The need for additional knowledge was also highlighted in our minimal genome prediction (Figure 3.7). While the majority of the genes targeted for deletion were not mapped to KEGG functional categories, a significant proportion (~36%) of the retained proteins were of unknown function (Figure 3.7D, Figure S3.14B). This fraction is very similar to its counterpart in JCVI-syn3.0 where it was initially reported that 149 genes out of the 473 were uncharacterized (~32%) [3].

Amongst the key features entirely retained in the minimal genome scenario were the cofactor biosynthesis-related proteins (Figure 3.7C), which is consistent with the fact that BOFdat Step2

enforced the addition of all vitamins and cofactors essential to sustain prokaryotic life [31,32]. Whether or not all these cofactors are effectively necessary to support *M. florum*'s growth could be addressed upon the definition of a completely defined medium. Given that the Vitamin & Cofactors module contained a higher fraction of proteins with a lower confidence level (Figure 3.3), such a study could provide highly valuable information for the complete understanding of *M. florum* molecular functions. BOFdat Step3 was used to find metabolites that contribute to increasing *i*JL208 gene essentiality prediction when added to the BOF (Figure 3.5). While this approach was not applied to other mollicutes models, all but one metabolite identified in this step were found in these models' BOF (Tableau 3.2). Taken together, these comparisons support the proposed BOF composition for *M. florum*.

Tableau 3.2 Comparison of the metabolites identified in BOFdat Step3 to those included in other mollicutes' model biomass compositions.

Metabolite	*M. genitalium*	*M. pneumoniae*	*M. gallisepticum*	JCVI-syn3.0A
Sulfur	**Present***	Absent	**Present***	Absent
Spermidine	**Present**	Absent	**Present**	**Present**
Cytidine	Absent	**Present**	Absent	Absent
Putrescine	**Present**	Absent	**Present**	Absent
Phosphatidylglycerol	Absent	Absent	**Present**	**Present**
Adenosine	Absent	**Present**	Absent	Absent
Methyltetrahydrofolate	**Present**	Absent	**Present**	**Present**
S-adenosyl-methionine	**Present**	**Present**	**Present**	Absent
Phosphatidyl-glycerophosphate	Absent	Absent	Absent	Absent

*Sulfate was identified instead of Sulfur in these models.

The CSY semi-defined medium allowed the assessment of *M. florum* growth on various energy sources and identified discrepancies between observations and model predictions (Figure 3.6). Comparing selected 3D protein structures reconstructed with I-TASSER to the PDB using FATCAT 2.0 revealed similarities with proteins of known promiscuity (Figure S3.9). While this approach does not constitute a direct validation of enzyme promiscuity, it does provide contextual hypotheses for reducing the search space for eventual biochemical characterizations. Yet our study was not the first to use such *ad hoc* reconstruction of 3D structures for genome-wide identification of protein functions [46–48]. We foresee that faced with the great challenge of identifying numerous molecular functions required for synthetic biology, combining the increasing reliability of structure prediction algorithms [49–51] to the predictive power of GEMs is likely to play an important role in organism design in the coming years.

M. florum-specific uptake and secretion rates were defined in CSY medium, a measurement performed for only two of the four modelled mollicutes acknowledged in our study [16,17]. While the calculated substrate uptake rate was slightly lower than both values recorded in other mollicutes, the combined lactate and acetate secretion rate was within the previously measured values (Tableau 3.3). In our datasets, expression of both lactate dehydrogenase (LDH) and pyruvate dehydrogenase (PDH) complex-forming genes was observed, which led to the hypothesis that both products would be secreted in *M. florum*. The initial lactate/acetate secretion rate ratio (8:1) chosen based on the expression data was exacerbated following sensitivity analysis (~15:1). This shift in ratios can be explained by the upper flux limit applied to the NADH oxidase reaction, which utilizes oxygen to recycle NADH cofactor produced when generating acetate and was necessary to ensure a non-linear relationship between oxygen uptake and growth rate. This shift also reflected the difference in LDH and PDH expression levels as well as the stoichiometric disparity of the two active protein complexes [52,53]. Our genome reduction scenario conserved the E1 component of the PDH whereas it was absent in JCVI-syn3.0. Hence, the individual and unequivocal quantification of *M. florum* lactate and

acetate secretion rates could reveal the conditions under which this pathway is essential, hereby shedding light on alternate genome reduction paths.

Tableau 3.3 Comparison of the main model constraints with those of other mollicutes' models.

Constraint	*M. florum*	*M. genitalium*	*M. pneumoniae*	*M. gallisepticum*	JCVI-syn3.0A
GAM	17.2	9.7*	25	9.7**	46.54**
NGAM	3.1	8.4*	3.3	8.4**	3.3**
Substrate uptake rate	5.26	5*	7.37	16.53	7.4**
Acetate secretion rate	0.53	Unconstrained	6.93	N/A	6.9**
Lactate secretion rate	8.16	Unconstrained	N/A	10.29	Unconstrained

All units in mmol gDW^{-1} h^{-1}.

*Extrapolated from other species.

**Extrapolated from other mollicutes models.

In *M. florum*, the high proportion of JCVI-syn3.0 orthologs (Figure S3.1) provided an interesting validation for the GEM-driven prediction of a minimal gene set featured in our study (Figure 3.7). The rational design of minimal genomes using the *M. genitalium* whole-cell model reported minimal genes sets considerably lower than the 473 genes contained in JCVI-syn3.0 (360 and 380) [12]. Given that JCVI-syn3.0 was pleomorphic, had a doubling time three times greater than its parent JCVI-syn1.0, and that 19 genes had to be re-inserted to produce the more robust JCVI-syn3.0A [8], predictions containing fewer genes than this organism are not likely to be viable. Our reduction scenario contained 90 more genes than JCVI-syn3.0, which we attribute to the constraint-based framework that makes deleting essential genes impossible [11].

Comparing our genome reduction scenario to JCVI-syn3.0 revealed the possibility that minimal genomes could use alternate carbohydrates to fuel their cellular needs. We also found that varying the growth rate constraint resulted in a reduced genome more similar to JCVI-syn3.0. A growth rate set to 60% of optimal was also notably similar to the growth rate ratio between JCVI-syn3.0 and the more robust JCVI-syn3.0A (50%) [8]. The absence of the E1 complex from JCVI-syn3.0 but its presence in our minimal gene set suggests that varying the constraints imposed on the input GEM could result in different genome reduction scenarios. Some proteins from the genetic information processing category differed from JCVI-syn3.0. The impact of these different chaperones, peptidases, ribosome methylases, and ribosome composition (i.e. : *rpmE*) could be assessed by generating a model that includes the expression machinery (ME-model) [54,55].

In conclusion, the quality of a GEM heavily depends on the level of biochemical knowledge available for the organism. While *i*JL208 was built on a revised annotation obtained from several computational approaches, we also noted that missing or incomplete knowledge could lead to false or inaccurate predictions. *i*JL208 will provide a framework to generate hypotheses, guide future experiments to address this challenge, and reach an exquisite understanding of cellular mechanisms in *M. florum*. With recent advances enabling complex genome manipulation in *M. florum* [20,21], *i*JL208 will also contribute to whole-genome engineering studies in this emerging model organism.

3.8 MATERIAL AND METHODS

3.8.1 Bacterial strains, data, and Memote report availability

All experiments described in this study were performed using *M. florum* strain L1 (ATCC 33453). The complete genome sequence of this strain is available in GenBank under RefSeq

accession number NC_006055.1. Genome annotations were either based on RefSeq (NC_006055.1), PATRIC (Genome ID: 265311.5), or both depending on specific analysis context and needs. The transposon mutagenesis dataset was taken from Baby and colleagues [19]. *M. florum* biomass composition, gene expression datasets, and lipidomic profile were taken from Matteau and colleagues [18]. Original transcriptomic and proteomic data are accessible through the Gene Expression Omnibus (GEO) under Series accession number GSE152985 and via the PRIDE partner repository with the dataset identifier PXD019922 and 10.6019/PXD019922, respectively. The final version of *i*JL208 was processed through the Memote software [56] to ensure compliance with the current standards for metabolic modeling. This report, the final *i*JL208 along with all code necessary to generate the results presented in this study are available on GitHub (https://github.com/jclachance/iJL208).

3.8.2 Proteome comparison

The proteome comparison tool from PATRIC [24] (https://www.patricbrc.org) was used to identify orthologous proteins between *Mesoplasma florum* L1 (Genome ID: 265311.5) and the following strains: *Mycoplasma gallisepticum* str. F; Genome ID: 708616.3, *Mycoplasma pneumoniae* M129; Genome ID: 272634.6, *Mycoplasma genitalium* G37; Genome ID: 243273.27, and *Mycoplasma mycoides* JCVI-Syn3.0; Genome ID: 2102.8. The parameters to identify orthologous proteins were the following: minimum positives of 0.2, minimum sequence coverage of 0.3, a minimum identity of 0.1, and a maximum E-value of 1e-5. In the case of pairwise proteome comparisons, both unidirectional and bidirectional best hits are used to define orthologous genes. Gene names were considered similar if they shared the same initial 3 characters in at least two species.

3.8.3 Homology modeling

3D protein structures were reconstructed for *M. florum* L1 coding sequences from RefSeq using the I-TASSER Suite 5.1 [57,58]. To provide relevant homology for functional predictions,

a pre-screening step was applied. This step used the Structural Systems Biology software (ssbio) [59] to compare the sequence of each *M. florum* protein to the Protein Data Bank (PDB) of crystallized structures [60] (www.rcsb.org). HMMER was then used to determine structural domain coverage in the PDB and identified an initial set of 459 proteins with a match to known domains. The following parameters were applied to filter this initial set to determine proteins that were likely to provide reliable structures from homology modeling: E-value < 1e-4, domain sequence identity > 10%, and domain sequence similarity > 30%. The transmembrane proteins (20) were also discarded given the limited capability of I-TASSER to produce a relevant model for such proteins [61]. All in all, a 3D structure was reconstructed for a total of 386 *M. florum* proteins. Structures with a "C-score" higher than -1.5 and a "Tm-score" higher than 0.5 were defined reliable (361).

The FATCAT 2.0 [36] software was used to find similar enzymes to selected structures reconstructed by I-TASSER. Structures were compared against the Protein DataBank (90% non-redundant set) using the Database search tool with the flexible FATCAT alignment parameter. Alignments with a P-value < 0.05 were considered as significant hits.

3.8.4 Identification of enzyme commission (EC) numbers

EC numbers were retrieved from the *M. florum* L1 RefSeq genome annotation (NC_006055.1) using the DETECT v2 [62] software and from the reliable protein structures reconstructed by I-TASSER using COFACTOR [63]. Identifications above the default probability threshold from DETECT v2 (90%) were considered as the gold standard. For cross-validation, EC numbers were considered similar if the first three digits were identical in at least two identification methods, i.e. using COFACTOR, DETECT v2, or from the RefSeq and PATRIC annotations.

3.8.5 Confidence level and final annotation score

A scoring system was established to integrate all information gathered through the computational identification of molecular functions in M. florum. For each method, the following score was attributed based on the level of precision that could be attributed to each gene, as presented in Figure 3.1E. Proteome comparison (Identical = 3, Similar = 2, Different = 1, No gene name or unique = 0), Structural homology (Reliable structure = 3, No reliable structure = 0), EC numbers (Identical EC = 3, Similar EC = 2, Different EC = 1, One method or no EC = 0). The cumulative score attributed to each predicted protein was defined as the final annotation score. Based on this score, a basic (< 3), medium (>= 3 and < 7) or high (>= 7) confidence level was attributed.

3.8.6 Reconstruction of the metabolic network

The draft *M. florum* metabolic model was reconstructed using the SimPheny platform [64] to ensure reaction conformity with a standard database and quality control. The reactions issued from the scaffold generated through the comparative approach were added first. Both RefSeq and PATRIC annotations, along with the identified EC numbers found with DETECT v2 and COFACTOR, were screened manually to identify metabolic candidates. Metabolic reactions associated with these genes were determined based on the information available in public databases [22,25,26,65,66]. When multiple reactions were possible, preference was given to the terms matching the most detailed genome annotation. Reaction and metabolite names used in the model followed the modeling specific nomenclature of the BiGG database [23]. The initial SimPheny model was imported in COBRApy [34] for further manipulations (i.e.: addition of species-specific reactions and metabolites, definition of the BOF, flux balance analysis. etc.).

The subsystems from the *Escherichia coli i*ML1515 model were assigned to reactions in *i*JL208 when their identifiers had a perfect match with an *i*ML1515 reaction. The subsystems

were then grouped together to form the six modules. Reactions that did not match an identifier in the *i*ML1515 model were manually assigned.

An interactive map of the entire reconstructed *M. florum* metabolic network was built using Escher [67]. The central metabolism map of the *Escherichia coli i*JO1366 model was used as an initial template on which the *i*JL208 model was mapped. The map was manually expanded using the reactions available in the model. This interactive map is provided as Supplementary file 5.

3.8.7 Flux-balance analysis

Flux balance analysis (FBA) is a mathematical approach to simulate cellular phenotype [13]. The metabolic network is represented as a stoichiometric matrix (***S***) where every row or column represents a unique metabolite or reaction, respectively. The stoichiometry of each metabolite in a reaction is given as a coefficient in the matrix. If we assume a vector of metabolic fluxes ***v***, the variation of metabolite concentration over time becomes:

$$\frac{dX}{dt} = S \cdot v \qquad \text{(équation 3.1)}$$

Where ***X*** is the vector of metabolites in the network. FBA assumes that the metabolic network will reach a steady-state. In this case, the concentration of metabolites over-time should be in equilibrium where the inputs are equal to the outputs so that:

$$0 = S \cdot v \qquad \text{(équation 3.2)}$$

Defining a physiologically meaningful objective (***Z***) allows the formulate an optimization problem on which constraints apply:

$$maximize Z,$$
$$0 = S \cdot v \qquad \text{(équation 3.3)}$$
$$a_i < v_i < b_i$$

Where ***a*** and ***b*** are the flux bounds on every reaction. This mathematical formulation can be solved using linear programming and allows finding the optimal solution of a given metabolic network at steady-state.

3.8.8 Biomass objective function

The *M. florum* BOF was defined using the BOFdat software [31] and leveraging the previously reported biomass composition of the cell and available omics datasets (transcriptomic, proteomic and lipidomic) [18]. The first and second steps of BOFdat were used to determine the precise stoichiometric coefficients of the major cellular macromolecules as well as inorganic ions and coenzymes, respectively. The third step of BOFdat was used to identify metabolites most improving the essentiality prediction accuracy of the model. Revised gene essentiality data previously published for *M. florum* was used in that context (see identification of essential genes section of the Material and methods). 50 evolutions were performed for 200 generations each. The biomass compositions (individuals) with the highest MCC were saved into a “Hall of Fame" for each evolution. The frequency of apparition of metabolites within the individuals saved in the Hall of Fame was determined and the ones appearing most frequently were added as part of the BOFdat Step3. From the nine metabolites identified, seven were considered part of the metabolite pool category (MWF: 1.2%). The stoichiometric coefficients of every metabolite in this category were re-computed using the BOFdat with this MWF and assuming that each metabolite is represented equally. Two metabolites belonged to the lipids category so the same procedure was employed but using the lipids MWF (18.3%). Finally, the metabolite representing the capsular polysaccharide of *M. florum* was added and its stoichiometric coefficient determined using the carbohydrates MWF (4.1%).

3.8.9 Development of a semi-defined growth medium

An exponential growth phase *M. florum* preculture grown at 34°C in ATCC 1161 medium (1.75% (w/v) heart infusion broth, 4% (w/v) sucrose, 20% (v/v) HS, 1.35% (w/v) YE, 0.004% (w/v) phenol red, 200 U/ml penicillin G) [21] was centrifuged for 1 min at 21,000 x *g* and washed twice with PBS 1X. Washed cells were inoculated at an initial concentration of ~1e5 CFU/ml into three different medium bases containing decreasing concentrations of HS and YE (from 20% HS/1.35% YE to 0.01% HS/ 0.0006% YE) and either 1.75% (w/v) heart infusion broth (ATCC 1161 base), PBS 1X (PBS base) or CMRL 1066 chemically defined medium (C5900-02A, US Biological; CMRL 1066 base), all supplemented with 0.004% (w/v) phenol red and 200 U/ml penicillin G. Medium bases were adjusted to a pH of ~7.5 and transferred into a 96-well microplate for growth assays. Half of wells were also supplemented with sucrose at a final concentration of 4% (w/v) for the ATCC 1161 base and 1% (w/v) for PBS and CMRL 1066 medium bases. The inoculated microplate was incubated with shaking at 34°C in a Multiskan GO microplate reader (Thermo Scientific) and the OD_{560nm} was measured every 10 min for 24 hours. Changes in the absorbance of phenol red at 560 nm caused by the metabolic activity of *M. florum* (medium acidification) were previously shown to correlate with the number of CFU/ml. Color fold change was then evaluated by comparing the minimal OD_{560nm} value observed over the entire incubation period to a non-inoculated control of identical medium composition. The color fold change observed for sucrose containing wells was then compared to wells not supplemented with sucrose, resulting in a normalized growth index for each tested medium base. The medium composition showing the highest normalized growth index for the lowest concentrations of HS and YE (CMRL 1066 base supplemented with 0.313% HS and 0.02% YE, Figure S3.3), referred to as CSY, was selected for the evaluation of growth sustaining carbohydrates. All conditions were tested in technical triplicate.

3.8.10 Experimental evaluation of *M. florum* growth on different carbohydrates

A 96-well microplate was filled with CSY supplemented with either one of the following carbohydrates, at a final concentration of 1% (w/v): sucrose, trehalose, fructose, glucose, maltose, mannose, glycerol, sorbitol, lactose, galactose, ribose, arabinose, N-acetylglucosamine, and glycerol-3-phosphate. A no-sugar control was also performed. Half of wells were inoculated at an initial concentration of ~1e5 CFU/ml with an *M. florum* preculture prepared and washed as described in the previous section. The microplate was incubated at 34°C without shaking for 24 hours. The OD_{560nm} at 24 hours was measured using a Multiskan GO microplate reader (Thermo Scientific) and compared between inoculated and non-inoculated conditions, resulting in a growth index for each carbohydrate tested. All conditions were tested in technical triplicate. For comparison with the model's predictions, the growth index measured for each carbohydrate was finally normalized to growth on sucrose.

3.8.11 *In silico* prediction of carbohydrates utilization

The formulated model with defined biomass composition and constraints was used for the prediction of carbohydrates utilization. An exchange reaction, the model equivalent of the medium composition was added for each carbohydrate tested experimentally. When tested, a lower bound of -10 was applied to the exchange reaction and the model was optimized for biomass production. The predicted growth rates for each carbohydrate were saved and normalized over sucrose to facilitate comparison with experimental results.

3.8.12 Measurement of *M. florum* doubling time and growth rate calculation

The doubling time of isolated transposon insertion mutants as well as *M. florum* growing in CMRL 1066 base medium with variable HS and YE concentrations was measured using colorimetric assays as described in Matteau *et al* [18]. Cultures were performed in technical duplicate and incubated at 34°C with shaking. Growth data of insertion mutants is available in

Supplementary file 9. For *M. florum* growing in CSY medium (0.313% HS and 0.02% YE) with variable initial concentration of sucrose, the doubling time was measured according to CFU counts of time course experiments. Briefly, a simple exponential growth model was fit to the mean CFU counts measured over time for each initial sucrose concentration (Eq. 4):

$$A_t = A_0 e^{rt} \qquad \text{(équation 3.4)}$$

Where A_0 is the initial number of bacteria and k is the growth rate. In simple exponential growth, the relation between growth rate (r) and doubling time (d) is given by:

$$d = \frac{ln(2)}{r} \qquad \text{(équation 3.5)}$$

3.8.13 Quantification of sucrose uptake rate and fermentation products secretion rate

An exponential growth phase *M. florum* preculture grown at 34°C in ATCC 1161 medium was centrifuged for 1 min at 21,000 x g and washed twice with PBS 1X. Washed cells were inoculated at an initial concentration of ~1e5 CFU/ml into different CSY media containing variable concentrations of sucrose. Inoculated media were adjusted to a pH of ~7.5 and transferred into a 96-well microplate for growth experiments. Cultures were incubated with shaking at 34°C in a Multiskan GO microplate reader (Thermo Scientific) and growth was followed by measuring CFU counts every ~90-120 min until late exponential phase. CFU were evaluated by spotting serial dilutions of the cultures on ATCC 1161 solid medium and counting colonies after an incubation of 24-48 hours at 34°C. For modeling purposes, CFU/ml were converted to gDW/L (biomass) according to the previously determined *M. florum* dry weight (Matteau et al, 2020). In addition to CFU/ml measurements, sucrose and fermentation products (lactate and acetate) were also quantified throughout the entire experiments by HPLC. For this task, cultures were filtered through 0.2 uM PES filters and frozen at -80°C until quantification. HPLC analysis was performed by the Laboratoire des Technologies de la Biomasse at the

Université de Sherbrooke. Briefly, sucrose and lactate/acetate quantification were performed using a Dionex ICS-5000+ ion chromatography system equipped with a KOH eluent generator, an analytical gradient pump, a thermostated AS-AP autosampler, and an electrochemical detector. The stability of the signal was ensured by a 200 mM KOH post-injection using a Dionex GP 50 gradient pump. a Dionex CarboPac SA10–4 µM column was used for sucrose quantification, while a XXX column was used for lactate and acetate quantification. Columns were set at 45 °C and the electrochemical detector was operated at 30 °C. Mobile phase was composed of aqueous KOH solution and the elution gradient mode was set as follows: 1 mM for 12 min, 10 mM for 5 min, 1 mM for 10 min. The flow rate was maintained at 1.25 ml·min^{-1}, and the injection volume was set to 5 µL. Quantifications were performed by external calibration using ~99% pure compounds (Acros). Since lactate and acetate peaks were indiscernible, corresponding peak areas were combined, resulting in a combined lactate/acetate estimate. Following quantification, substrate (sucrose) specific uptake rate and fermentation product (lactate/acetate) specific secretion rates were calculated according to the following equation:

$$qS = \frac{\Delta S \cdot r}{X_{t2}} \qquad \text{(équation 3.6)}$$

Where qS is the substrate (or product) specific rate, ΔS is the variation of substrate concentration over time and X_{t2} is the biomass concentration at the end of the time interval [68]. To calculate qS, simple exponential fits were applied to the mean of biomass, sucrose, and lactate/acetate concentration data points (Figure S3.5). A linear regression in exponential phase (14hr to 16hr) was then applied to these fits (Figure S3.6), and ΔS and X_{t2} were calculated for each 1-hour time interval using the parameters of these regressions. Growth rates (r) were obtained from the exponential fits applied to biomass data of each condition (Eq. 4). The maximum substrate and product specific rates used for modeling purposes were determined by computing the average of the possible rates obtained for the two highest initial sucrose concentrations (0.05 and 0.1%).

3.8.14 Sensitivity analysis

The model sensitivity to maintenance costs as well as uptake and secretion rates was assessed by setting initial parameters and further varying each of them individually. The initial parameters used were as follow: Growth associated maintenance (GAM): -5 mmol gDW^{-1} h^{-1}, Non-growth associated maintenance (NGAM): 3 mmol gDW^{-1} h^{-1}, sucrose uptake rate: -5.26 mmol gDW^{-1} h^{-1}, lactate secretion rate: -7.65 mmol gDW^{-1} h^{-1}, acetate secretion rate: -0.96 mmol gDW^{-1} h^{-1}, oxygen uptake rate: -10 mmol gDW^{-1} h^{-1}. The initial sucrose uptake rate used is the rate defined experimentally in CSY medium. The sum of the initial lactate and acetate secretion rates is equal to the combined secretion rate measured experimentally at -8.61 mmol gDW^{-1} h^{-1}. The choice was made to favor lactate over acetate secretion since its path to secretion in the metabolic network is simpler. The initial oxygen uptake rate represents half the rate reported for *E. coli* growing in minimal medium batch cultures [69].

Maintenance costs represent the amount of energy necessary to support the cell aside from the production of biomass components from the metabolic network. While GAM is the ATP hydrolysis reaction within the biomass objective function, the NGAM is represented by the ATP maintenance reaction which also consumes ATP but is independent of biomass production. These rates were defined first as they are most impactful for growth rate prediction [31]. Using the initial parameters, the GAM was determined by varying the ATP maintenance in the BOF from 0 to 50 mmol gDW^{-1} h^{-1} and predicting the growth rate with standard FBA optimization for each GAM value. The theoretical GAM value was identified matching the experimental growth rate in CSY (0.44 hr^{-1}). Similarly, the NGAM was identified by varying its value from 0 to 30 mmol gDW^{-1} h^{-1}, fixing the theoretical GAM and keeping the initial uptake and secretion rates. This allowed the identification of the NGAM value at which the predicted growth rate fits its experimental counterpart.

The model sensitivity to uptake and secretion rates was then evaluated using the fixed theoretical maintenance costs. Using the initial oxygen uptake as well as lactate and acetate secretion rates, the sucrose uptake rate was varied between 3 and 15 mmol gDW^{-1} h^{-1}, which encompasses the experimentally measured uptake rate of 5.26 mmol gDW^{-1} h^{-1} (absolute value). Sensitivity to lactate and acetate production was evaluated on specific ranges (0 to 40 mmol gDW^{-1} h^{-1}, 0 to 15 mmol gDW^{-1} h^{-1}, respectively) with the determined sucrose uptake rate and initial oxygen uptake rate. The lactate secretion rate was determined as the minimum rate matching the experimental growth rate in CSY (0.44 hr^{-1}). This value was found at 8.16 mmol gDW^{-1} h^{-1}. The acetate secretion rate was defined as the difference between the average experimentally determined value of 8.61 mmol gDW^{-1} h^{-1} for both products and the theoretical lactate secretion rate, corresponding to 0.53 mmol gDW^{-1} h^{-1}. Finally, the impact of oxygen uptake rate was assessed using all constraints and rates previously determined. The oxygen uptake rate varied between 0 and 30 mmol gDW^{-1} h^{-1}. The final uptake rate (4.81 mmol gDW^{-1} h^{-1}) was chosen to match the experimental growth rate value in CSY (0.44 hr^{-1}).

3.8.15 Identification of expressed genes

Transcriptomic and proteomic expression datasets were available for *M. florum* [18]. To determine the set of expressed genes, the reported number of protein molecules per cell and fragments per kilobase per million of mapped reads (FPKM) associated to each gene were compared with each other. An expression threshold spanning the entire range of measured values was iteratively applied to each dataset resulting in a list of expressed and unexpressed genes. The resulting binary vectors were compared using the MCC by setting the transcriptomic data as a reference and generating a distinct MCC value per pair of thresholds. The correlation matrix containing all MCC values was multiplied by the number of genes expressed at the given thresholds to produce a correlation score:

$$Score = \prod_{MCC}^{j} \overline{x_{i,j}} \qquad \text{(équation 3.7)}$$

This score accounts for the correlation between each dataset while maximizing the number of expressed genes. The thresholds providing the optimal score were used for this study and were found at 23 proteins per cell (proteomics) and 168 FPKM (transcriptomics).

3.8.16 Identification of essential genes

Previously published experimental essentiality data [19] generated by transposon mutagenesis was re-analyzed in this study. The doubling time measured for individual mutants (Supplementary file 9) was used along with the relative position of the insertion site within the interrupted gene to re-evaluate gene essentiality (Figure S3.15). The growth data was filtered to include only mutants for which the standard deviation of doubling time between replicates was within 30% of the average doubling time measured. Mutants for which a reliable doubling time could be obtained were defined as non-viable if their measured doubling time exceeded the sum of the median and median absolute deviation. For insertions not impairing the growth of *M. florum*, interrupted genes were considered essential only if the transposons were strictly restricted to the terminal region of genes, defined as the last 20% of the gene length. Both Hutchison *et al.* [3] and Breuer *et al.* [8] used the location of transposon insertion to nuance their essentiality observations.

3.8.17 Prediction of metabolic flux state

The flux state through the metabolic network was obtained by optimizing the production of biomass using parsimonious flux balance analysis (pFBA), a version of FBA that allows the generation of a unique flux state prediction through minimization of enzyme usage [70]. This method is best suited for the comparison of predicted fluxes to gene expression [71]. A reaction flux was defined as active when the predicted value exceeded the numerical error (1e-8) and

the flux was attributed to every gene that could catalyze the reaction via the gene-reaction rule. The objective was set to the BOF from BOFdat Step3 and the *in silico* medium set with sucrose as the main energy source.

3.8.18 Model-driven prediction of a minimal gene set and identification of functional features

The MinGenome algorithm [11] was used to sequentially identify the longest possible deletions in the *M. florum* genome. The transcription units (relationship between gene and promoter locations) were obtained from the integrative characterization of *M. florum* [18]. The *i*JL208 model along with the experimentally determined essential genes revised in this study were also used as input. The algorithm extracts constraints from these inputs and writes a bi-level linear program where the lower level optimizes the production of biomass and the higher level maximizes the DNA length (bp) of the deletion to be performed. To identify the minimal gene set for *M. florum*, the optimization was performed iteratively 100 times. Deleted genes and promoters were encompassed between a deletion start and end site.

The *M. florum* proteome was compared to that of JCVI-syn1.0 (fasta file reconstructed from DataSetS1 from Hutchison *et al.* [3] and JCVI-syn3.0 (Genome ID: 2102.8) using the PATRIC proteome comparison (see above). Since multiple comparisons were executed, only bidirectional best hits were used to define homologous genes. Mapping of *M. florum* genes to KEGG functional categories [72] was performed as described previously [18], where the automated attributions were manually curated to fit the context of *M. florum*. The composition of the cell was depicted using the proteomap software [73] and the protein abundance was obtained from available transcriptomic data [18].

3.9 ACKNOWLEDGMENTS

Funding for this research was provided by the Natural Sciences and Engineering Research Council of Canada: 2020-06151, by the Fonds de recherche du Québec – Nature et technologies: 2018-PR-206064, and by the Novo Nordisk Foundation: NNF10CC1016517. The authors would like to acknowledge and thank for their indirect but substantial contribution to this work, members of the Systems Biology Research Group that were not listed as authors. Specifically, Dr. Jared Broddrick, Dr. Erol Kavvas, Dr. Yara Seif and Charles J. Norsigian, for the multiple computational modeling related discussions. Edward Catoiu for additional support with 3D modeling of protein structures. Marc Abrams for the revision of written english. A special thanks to Pr. Laurence Yang and Dr. Anand V. Sastry for conceptualization of the study.

3.10 REFERENCES

1. Hughes, R. A. & Ellington, A. D. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harb. Perspect. Biol.* **9**, (2017).
2. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
3. Hutchison, C. A., 3rd *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
4. Venetz, J. E. *et al.* Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8070–8079 (2019).
5. Danchin, A. & Fang, G. Unknown unknowns: essential genes in quest for function. *Microb. Biotechnol.* **9**, 530–540 (2016).
6. Glass, J. I., Merryman, C., Wise, K. S., Hutchison, C. A. & Smith, H. O. Minimal Cells—Real and Imagined. *Cold Spring Harb. Perspect. Biol.* (2017) doi:10.1101/cshperspect.a023861.
7. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
8. Breuer, M. *et al.* Essential metabolism for a minimal cell. *Elife* **8**, (2019).

9. Yurkovich, J. T. & Palsson, B. O. Solving Puzzles With Missing Pieces: The Power of Systems Biology. *Proc. IEEE* **104**, 2–7 (2016).
10. Lachance, J.-C., Rodrigue, S. & Palsson, B. O. Minimal cells, maximal knowledge. *Elife* **8**, (2019).
11. Wang, L. & Maranas, C. D. MinGenome: An *In Silico* Top-Down Approach for the Synthesis of Minimized Genomes. *ACS Synth. Biol.* **7**, 462–473 (2018).
12. Rees-Garbutt, J. *et al.* Designing minimal genomes using whole-cell models. *Nat. Commun.* **11**, 836 (2020).
13. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
14. O'brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693 (2013).
15. Suthers, P. F. *et al.* A Genome-Scale Metabolic Reconstruction of *Mycoplasma genitalium*, iPS189. (2009) doi:10.1371/journal.pcbi.1000285.
16. Wodke, J. A. H. *et al.* Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.* **9**, 653 (2013).
17. Bautista, E. J. *et al.* Semi-automated Curation of Metabolic Models via Flux Balance Analysis: A Case Study with Mycoplasma gallisepticum. (2013) doi:10.1371/journal.pcbi.1003208.
18. Dominick Matteau, Jean-Christophe Lachance, Frédéric Grenier, Samuel Gauthier, James M. Daubenspeck, Kevin Dybvig, Daniel Garneau, Thomas F. Knight, Pierre-Étienne Jacques, & Sébastien Rodriguea. Integrative characterization of the near-minimal bacterium Mesoplasma florum. (2020).
19. Baby, V. *et al.* Inferring the Minimal Genome of Mesoplasma florum by Comparative Genomics and Transposon Mutagenesis. *mSystems* **3**, (2018).
20. Baby, V. *et al.* Cloning and Transplantation of the Mesoplasma florum Genome. *ACS Synth. Biol.* (2017) doi:10.1021/acssynbio.7b00279.
21. Matteau, D. *et al.* Development of oriC-based plasmids for Mesoplasma florum. *Appl. Environ. Microbiol.* (2017) doi:10.1128/AEM.03374-16.
22. Artimo, P. *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **40**, W597–603 (2012).
23. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–22 (2016).
24. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* **45**, D535–D542 (2017).
25. Placzek, S. *et al.* BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **45**, D380–D388 (2017).

26. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–62 (2016).
27. Sirand-Pugnet, P., Citti, C., Barré, A. & Blanchard, A. Evolution of mollicutes: down a bumpy road with twists and turns. *Res. Microbiol.* **158**, 754–766 (2007).
28. Keçeli, S. A. & Miles, R. J. Differential inhibition of mollicute growth: an approach to development of selective media for specific mollicutes. *Appl. Environ. Microbiol.* **68**, 5012–5016 (2002).
29. Monod, J. The growth of bacterial cultures. *Annu. Rev. Microbiol.* **3**, 371–394 (1949).
30. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344–349 (2010).
31. Lachance, J.-C. *et al.* BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLOS Computational Biology* vol. 15 e1006971 (2019).
32. Xavier, J. C., Patil, K. R. & Rocha, I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. *Metab. Eng.* **39**, 200–208 (2017).
33. Varma, A. & Palsson, B. O. Metabolic capabilities of Escherichia coli: I. synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.* **165**, 477–502 (1993).
34. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7**, 74 (2013).
35. Edwards, J. S., Ramakrishna, R. & Palsson, B. O. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol. Bioeng.* **77**, 27–36 (2002).
36. Li, Z., Jaroszewski, L., Iyer, M., Sedova, M. & Godzik, A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Research* vol. 48 W60–W64 (2020).
37. Auiewiriyanukul, W., Saburi, W., Kato, K., Yao, M. & Mori, H. Function and structure of GH13_31 α-glucosidase with high α-(1→4)-glucosidic linkage specificity and transglucosylation activity. *FEBS Letters* vol. 592 2268–2281 (2018).
38. Bertin, C. *et al.* Highly dynamic genomic loci drive the synthesis of two types of capsular or secreted polysaccharides within the Mycoplasma mycoides cluster. *Appl. Environ. Microbiol.* **81**, 676–687 (2015).
39. Regni, C., Tipton, P. A. & Beamer, L. J. Crystal structure of PMM/PGM: an enzyme in the biosynthetic pathway of P. aeruginosa virulence factors. *Structure* **10**, 269–279 (2002).
40. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010).
41. Miles, R. J. Catabolism in mollicutes. *J. Gen. Microbiol.* **138**, 1773–1783 (1992).

42. Lilleorg, S., Reier, K., Remme, J. & Liiv, A. The Intersubunit Bridge B1b of the Bacterial Ribosome Facilitates Initiation of Protein Synthesis and Maintenance of Translational Fidelity. *J. Mol. Biol.* **429**, 1067–1080 (2017).
43. Ostrov, N. *et al.* Technological challenges and milestones for writing genomes. *Science* **366**, 310–312 (2019).
44. Arraes, F. B. M. *et al.* Differential metabolism of Mycoplasma species as revealed by their genomes. *Genet. Mol. Biol.* **30**, 182–189 (2007).
45. Fisunov, G. Y. *et al.* Reconstruction of Transcription Control Networks in Mollicutes by High-Throughput Identification of Promoters. *Front. Microbiol.* **7**, 1977 (2016).
46. Yang, Z. & Tsui, S. K.-W. Functional Annotation of Proteins Encoded by the Minimal Bacterial Genome Based on Secondary Structure Element Alignment. *J. Proteome Res.* **17**, 2511–2520 (2018).
47. Yang, Z., Zeng, X. & Tsui, S. K.-W. Investigating function roles of hypothetical proteins encoded by the Mycobacterium tuberculosis H37Rv genome. *BMC Genomics* **20**, 394 (2019).
48. Antczak, M., Michaelis, M. & Wass, M. N. Environmental conditions shape the nature of a minimal bacterial genome. *Nat. Commun.* **10**, 3100 (2019).
49. Billings, W. M., Hedelius, B., Millecam, T., Wingate, D. & Della Corte, D. ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. doi:10.1101/830273.
50. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
51. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in CASP13. *Proteins: Struct. Funct. Bioinf.* (2019).
52. Mattevi, A. *et al.* Atomic structure of the cubic core of the pyruvate dehydrogenase multienzyme complex. *Science* **255**, 1544–1550 (1992).
53. Wigley, D. B. *et al.* Structure of a ternary complex of an allosteric lactate dehydrogenase from Bacillus stearothermophilus at 2·5 Å resolution. *J. Mol. Biol.* **223**, 317–335 (1992).
54. Liu, J. K. *et al.* Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Syst. Biol.* **8**, 110 (2014).
55. Lloyd, C. J. *et al.* COBRAme: A computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput. Biol.* **14**, e1006302 (2018).
56. Lieven, C. *et al.* MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* **38**, 272–276 (2020).
57. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
58. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725 (2010).

59. Mih, N. *et al.* ssbio: a Python framework for structural systems biology. *Bioinformatics* **34**, 2155–2157 (2018).
60. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
61. Koehler Leman, J., Ulmschneider, M. B. & Gray, J. J. Computational modeling of membrane proteins. *Proteins* **83**, 1–24 (2015).
62. Nursimulu, N., Xu, L. L., Wasmuth, J. D., Krukov, I. & Parkinson, J. Improved enzyme annotation with EC-specific cutoffs using DETECT v2. *Bioinformatics* **34**, 3393–3395 (2018).
63. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* **45**, W291–W299 (2017).
64. Schilling, C. H., Thakar, R., Travnik, E., Van Dien, S. & Wiback, S. SimPheny^{TM}: A Computational Infrastructure for Systems Biology.
65. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
66. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
67. King, Z. A. *et al.* Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput. Biol.* **11**, e1004321 (2015).
68. Sauer, U. *et al.* Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *J. Bacteriol.* **181**, 6679–6688 (1999).
69. Andersen, K. B. & von Meyenburg, K. Are growth rates of Escherichia coli in batch cultures limited by respiration? *J. Bacteriol.* **144**, 114–123 (1980).
70. Lewis, N. E. *et al.* Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).
71. Machado, D. & Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* **10**, e1003580 (2014).
72. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–80 (2004).
73. Liebermeister, W. *et al.* Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8488–8493 (2014).

3.11 SUPPLEMENTARY TEXT

3.11.1 Identification of molecular functions in *M. florum* L1

Mesoplasma florum L1 is bacterium of the Mollicute class that was isolated from a lemon tree flower [1]. It is phylogenetically associated with the Mollicutes class, a group of wall-less bacteria that was identified as an ideal candidate for the study of the minimal combination of components capable of sustaining cellular life [2]. More specifically, the Mollicutes are autonomously replicating bacterias with genome sizes varying from ~500kb in the case of *Mycoplasma genitalium* [3] to more than a million (~1.5Mbp) for the larger *Acheloplasma laiidwali* [4]. Multiple studies have provided a general understanding of the metabolism of these near-minimal cells [5,6]. While this understanding is useful and has previously been put to profit for genome-scale metabolic reconstructions of different Mollicute species [7–9], studies specifically targeted at *M. florum* are rare. Hence, we set to extract a maximal amount of functional information from the *M. florum* genome by:

1. comparing the proteome of *M. florum* with that of previously published genome-scale models,
2. extracting enzyme commission numbers,
3. and generating three-dimensional structures through homology modelling.

Tableau S3.1 Comparison of common Mollicute species' characteristics.

	Genome size*	**Number of coding genes***	**Host***	**Pathogenicity***	**Doubling time in rich medium**
***Mesoplasma florum* L1 (AE017263)**	793 224	685	Insects or flower	No	0.6 hr**
JCVI Syn3.0 (CP014940.1)	531 490	452	N/A	N/A	3 hr[10]

Tableau S3.1 Comparison of common Mollicute species' characteristics. (suite)

***Mycoplasma mycoides* JCVI-syn1.0 (CP002027.1)**	1 203 804	1138	Cattle	Yes	1 hr[11]
***Mycoplasma genitalium* G37 (AAGX00000000)**	559 388	653	Human	Yes	12 hr[12]
***Mycoplasma pneumoniae* M129 (U00089)**	816 394	755	Human	Yes	6 hr[12] to 20 hr[8]
***Mycoplasma gallisepticum* str. F (CP001873)**	977 612	796	Chicken	Yes	2 hr[13]
***Acholesplasma laidlawii* NCTC10116 (LS483439)**	1 497 557	1373	Plants?	No	Weeks[14]

*Obtained from the PATRIC database (https://www.patricbrc.org/) [15]
**This study

3.11.1.1 Proteome comparison

First, the *M. florum* proteome was compared with four mollicutes species for which GEMs were previously generated (*Mycoplasma genitalium*[7], *Mycoplasma pneumoniae*[8], *Mycoplasma gallisepticum*[9], *Mycoplasma mycoides* JCVI-Syn3.0A [16]) using the PATRIC proteome comparison tool [15] (Figure 3.1A and B). Despite having the lowest total number of proteins (438), JCVI-syn3.0 had the highest number of orthologs (411) followed by *M. gallisepticum* (344), *M. pneumoniae* (326), and finally, *M. genitalium* (318) (Figure S3.1), which is consistent with the phylogenetic distance between *M. florum* and these organisms [17].

Gene names are useful to query public databases for putative functions but were initially scarcely assigned in *M. florum* with 106 occurrences out of 676 predicted proteins common to both RefSeq and PATRIC annotations (Supplementary file 1). This comparison resulted in the association of 366 *M. florum* proteins to a gene name in at least one species (Figure 3.1E). Redundant gene name association across species provided higher confidence in these associations. Among the 281 gene names that were identified in more than one species, 156

were identical, 113 had two possible identifications (similar) and 12 had more than two different attributions. This approach allowed the identification of 72 new gene names in *M. florum*, considering that 149 genes names total are shared across all species surveyed and that 77 of those were already identified in *M. florum*. 15 of the remaining 19 gene names identified in *M. florum* were shared with at least one other species surveyed while 4 gene names (*guaA*, *rpsT*, *rpmC*, *rpsF*) were specific to *M. florum*.

The proteome comparison approach allowed linking proteins to model reaction identifiers, thereby generating an initial draft reconstruction of the *M. florum* metabolic network (Supplementary file 1). The gene-reaction rule of each model allowed to link genes orthologous to *M. florum* to reactions in the models. The gene to reaction mapping obtained for each model was converted to BiGG identifiers [18] using MetaNetX [19]. This draft was used to initiate the metabolic reconstruction process. Identified reactions were added using the SimPheny software [20] and later extracted to be used with the COBRApy toolbox [21].

3.11.1.1 Homology modeling

Next, *M. florum* proteins were investigated from a structural standpoint. Given the unavailability of *M. florum* crystallized protein structures in the Protein Data Bank (PDB) [22], the Structural Systems Biology software (ssbio) [23] was used to map the 680 open reading frames predicted in the *M. florum* L1 genome annotation from RefSeq (NC_006055.1) against known domains in the PDB. Proteins containing successfully mapped domains were then selected for 3D homology modeling using the I-TASSER suite [24] (Figure 3.1C). Initially, 459 (67.5%) mapped to known domains in the PDB (Supplementary file 2). After filtering, 386 domain-supported proteins were chosen for 3D structure prediction (Material and methods).

The quality of the generated reconstruction is evaluated by the quality of the threading alignment and convergence of the assembly refinement performed by I-TASSER [25],

summarized in a C-score (Figure S3.2A). As recommended by the authors of the I-TASSER suite, we used a C-score cutoff of -1.5 to determine structures of higher quality. From the 386 proteins selected for homology modelling, we obtained 361 proteins (95.6%) with a C-score exceeding this threshold. Based on the quality scores provided by I-TASSER, 361 structures were deemed reliable (Figure 3.1E, Figure S3.2) and were further analyzed using COFACTOR, a software providing EC numbers, Gene Ontology (GO) terms, and binding site predictions [26] (Figure 3.1C, Supplementary file 2).

3.11.1.2 EC number identification

EC numbers provide valuable information on the biochemical reactions catalyzed by enzymes [27] and were obtained from both RefSeq and PATRIC genome annotations and compared to the predictions formulated by both COFACTOR and DETECT v2 (Figure 3.1D). Of the 393 proteins associated with at least one EC number, 207 were obtained with more than one method and 186 were identified with a single one (Supplementary file 3). 112 of the 125 high-probability identifications found by DETECT v2 were identical with at least one method. The remaining 13 identifications shared the first three EC digits with at least one other method. Of the 186 identifications found with a single method, 164 were specific to COFACTOR, 20 were specific to PATRIC and 2 were lower quality hits found only by DETECT v2. The consistency between EC number predictions was compared by matching the EC digits obtained with each method, showing that more than 87% have identical (113) or similar (67) EC digits (Figure 3.1E, Material and methods).

Our study is not the first to use *ad hoc* reconstruction of 3D structures for genome-wide identification of protein functions [28–30]. Of all 4 approaches used, COFACTOR identified the most EC numbers. Its predictions were nonetheless frequently different from standard, sequence-based methods. Albeit the potential for false positive identifications, these predictions were useful to formulate contextual hypotheses where the metabolic network

suggested the need for a given function (Figure 3.6). Faced with the great challenge of identifying several molecular functions required for synthetic biology, our study demonstrated the useful application of protein structures for the generation of testable hypotheses. With an increasing reliability of structure prediction algorithm [31–33], this type of approach is likely to gain interest.

3.11.1.3 Consolidated confidence score

Our bioinformatic analysis allowed us to extract extensive information from the genome and attribute a confidence score for each of those genes to function association. While a similar approach using a combination of computational methods was previously used to predict unknown molecular functions in a genome [34], our rationale was that the cross-validations would increase confidence levels and establish a hierarchy in the current annotation (Figure 3.1E). Indeed, most of the proteins for which no gene name was identified also did not map to known structural domains or functional EC number. Overall, between 283 and 315 proteins, ~45% of the total proteins had a lower confidence. Cross-validated top tier confidence proteins were more scarce with 156 identical gene names identified and 113 identical EC numbers identified through different methods which sums up to ~20% of the total proteins. The remaining ~35% of proteins had mid-range confidence. These results show that the identification of molecular functions, even in small genomes, is far from complete and once again, the high proportion of mitigated functions should stimulate the effort for experimental protein characterization [35].

3.11.2 Genome-scale metabolic network reconstruction

The reconstruction of the *M. florum* metabolic network was executed as described by Thiele and Palsson [36]. To increase the reliability of the reconstruction, both GenBank and PATRIC [15]

genome annotations were used as reference annotations. The potential metabolic candidates were extracted based on enzyme commission number and product name. This information was used to query publicly available reaction databases [37–41]. The identified reactions were added using the SimPheny framework to ensure charge balance and conformity with an existing functional nomenclature (Supplementary file 1). Refinement of the initial reconstruction was made by curating each metabolic objective individually and studying literature for biochemical evidence in *M. florum*.

The details of this manual reconstruction process are presented here and divided in 6 sections. Each section corresponds to a greater category named "Module". The utility of dividing the metabolism in such sections is to simplify future engineering tasks, a concept that was brought forth by Danchin and colleagues [42,43]. The 6 modules presented here are: (1) Nucleotide synthesis, (2) Protein synthesis, (3) Lipid and membrane synthesis, (4) Glycans synthesis, (5) Cofactors and coenzymes synthesis and (6) Energy production (Figure 3.3). The detail of every module composition together with the model in a spreadsheet format are available in Supplementary file 4.

The EC numbers and gene names identified through the computational identification of molecular functions were used to attribute reactions to the genes in the network. Along with our reasoning for the reconstruction of the model, the next sections identify genes for which a problem seemed obvious in the curation. These areas of lesser knowledge would require further experimental biochemical characterization.

3.11.2.1 Nucleotide synthesis

The synthesis of nucleotides is fundamental to all life forms. Manual curation of the genome and identification of gene names and EC numbers revealed 39 genes belonging the Nucleotides module (Supplementary file 4).This module is the largest by number of reactions (Figure 3.3A)

but contains a lower number of genes (39) than the Energy module (57) (Figure 3.3B). This module holds the highest number of genes involved in multiple reactions (20). In particular, the pyruvate kinase (Mfl175) is involved in 10 reactions, the largest number for any gene in the model. Like in most mollicutes, a dedicated nucleotide diphosphate kinase is absent in *M. florum* [44], the pyruvate kinase was hypothesized to generate nucleotide triphosphates for all nucleotides.

Ribose. Ribose, which serves as a backbone for nucleotides, has an associated ABC transporter encoded in the *M. florum* genome (Mfl666, Mfl667, Mfl668 and Mfl669). Intracellular ribose is then phosphorylated by a specific ribose kinase (Mfl642). It is noteworthy that, through the proteome comparison process, this gene matched with *fruK* from *M. genitalium*, where this gene was annotated as putative phosphofructokinase. Here, both PATRIC and RefSeq annotations suggest a ribokinase activity with both PATRIC and COFACTOR identifying EC 2.7.1.15 (ribokinase or deoxy-ribokinase) as the EC number at four digits. We corrected the gene name from *fruK* to *rbsK*, to be consistent with the *E. coli* nomenclature (Supplementary file 4).

Phosphoribosyltransferases. Phosphoribose can then be used in the pentose phosphate pathway (discussed later) or in the nucleotide synthesis. To be included into the synthesis of nucleotides, a nucleobase needs to be fixed to ribose. The enzyme that catalyzes this process for the various nucleobases is a phosphoribosyltransferase. Reviewing the annotation of *M. florum* identified a total four phosphoribosyltransferases (Mfl 107, Mfl276, Mfl463, and Mfl588).

Mfl107 (*upp*) encodes a uracil phosphoribosyltransferase for which 3 methods assigned the EC number 2.4.2.9. The confidence is high that this enzyme is specific to uracil. These 3 enzymes suggest that *M. florum* is capable of synthesizing every nucleic acid necessary for the *de novo* synthesis of DNA and RNA from free nucleobases and ribose.

The EC number identification for Mfl276 (*apt*) was consistent in all four methods and converged on EC 2.4.2.7. The reaction catalyzed by this gene is adenine or adenosine phosphoribosyltransferase which fixes the nucleobase adenine to the first carbon of the phosphorylated ribose, generating a nucleotide.

The phosphoribosyltransferase Mfl463 (*hpt*) annotated in RefSeq is "hypoxanthine-guanine phosphoribosyltransferase". Three methods associated the same EC number to this gene (EC 2.4.2.8). The KEGG [37] annotation states that the guanine phosphoribosyltransferase can use hypoxanthine as a substrate for the reaction. Two reactions were therefore associated with this gene in the model: GUAPRT and HXPRT.

The fourth enzyme of this class is the nicotinate phosphoribosyltransferase encoded by Mfl588 and is discussed further in the Vitamins & cofactors synthesis module (6). The current RefSeq annotation identified the EC number 2.4.2.11 which is obsolete according to KEGG [37]. The replacement EC number (6.3.4.21) was correctly identified by both PATRIC and DETECT while COFACTOR also attributed the old EC number 2.4.2.11. Given the high confidence in the EC number attributed to that gene we rename it *pncB* to be consistent with the *Salmonella typhimurium* annotation from which the function was fetched [45].

DNA uptake. Previous studies have suggested that no transporter exists for nucleosides or nucleotides in Mollicutes [46]. Nevertheless, the manual curation of the genome allowed to identify 2 genes that could satisfy the demand for individual nucleobases. Mfl413 and Mfl658 are both identified in RefSeq as Uracil/Xanthine permease. The structure for a uracil permease (*uraA*) was previously generated [47] and the authors suggested a proton symport mechanism. The associated gene in *M. florum* and other Mollicutes is named *pyrP*. While the *E. coli* gene seemed specific to uracil, the current annotation for Mollicutes suggests the import of both a purine (Xanthine) and a pyrimidine (Uracil). Considering that those genes are the only two associated with nucleobases import in *M. florum* and the potential for promiscuity of reactions catalyzed by an organism whose genome has been reduced [48], we initially include the import

of all nucleobases for which a phosphoribosyltransferase reaction is annotated. Hereby, we identified adenine, guanine, xanthine/hypoxanthine and uracil as the first attempt at characterizing essential nucleobases for *M. florum*.

Another possible system worth mentioning is catalyzed by the DNA uptake proteins (Mfl027 and Mfl329). As suggested before [49,50], in the Mollicutes' natural environment, DNA uptake may occur through the direct import of larger fragments of DNA from nearby dying cells [46]. In a laboratory setting the long DNA fragments could come from undefined media components such as yeast extract. These large fragments could then be digested using exonucleases. Mfl055 is such a 5'-3' exonuclease that could be used for this process. While this mechanism remains hypothetical, the possibility that long DNA fragments could be chewed by membrane associated nucleases and incorporated by a competency related protein should be kept in mind for further biochemical characterizations.

Finally, the whole network curation identified four putative essential components necessary in for *M. florum* growth: adenine, guanine, thymidine and ribose. Although an entire ABC transport system is annotated for ribose (Mfl666, Mfl667, Mfl668), the exact nature and function for DNA uptake would require further characterization in a completely defined medium.

Phosphorylation of nucleotides. Monophosphate-nucleotides formed by the combination of the phosphorylated ribose backbone and the imported nucleobases need to be phosphorylated twice before they can be incorporated into macromolecules (DNA and RNA). The first phosphorylation step leads to the formation of diphospho-nucleotides. Adenylate kinase (Mfl144, *adk*), guanylate kinase (Mfl195, *gmk*), cytidylate kinase (Mfl198, *cmk*), uridine kinase (Mfl306, *udk*), and thymidylate kinase (Mfl676, *tmk*) activities are annotated in *M. florum* with consistent EC numbers found across three different methods for each of them (Supplementary files 3 and 4).

In *M. florum,* the phosphorylation of deoxy-nucleotides was not specifically reported by the RefSeq annotation. Nevertheless, by consulting the PATRIC annotation, deoxyadenosine kinase (EC 2.7.1.76) and deoxyguanosine kinase (EC 2.7.1.113) were attributed to Mfl547. The EC 2.7.1.113 was also reported by COFACTOR. Our approach also identified a specificity to deoxythymidine. In fact, a dTMP kinase activity (EC 2.7.4.9) was attributed to Mfl676 by PATRIC, DETECT and COFACTOR.

In Mollicutes, the lack of annotation for a Nucleotide Diphosphate Kinase (NDPK) is common [49,50]. It has been hypothesized that the relaxation of the catalytic site of the glycolytic enzyme pyruvate kinase (Mfl175, *pyk*) would allow it to phosphorylate other nucleotides than ADP [44]. It has been reported that the Mollicute's *pyk* conserves 5 to 21% of its activity when using other substrates than ADP [44]. For modeling purposes, reactions PYK2 to PYK10 (8 reactions) were added to ensure that all nucleotide-diphosphate could be converted into nucleotide-triphosphate, building blocks of DNA and RNA.

Ribonucleoside-diphosphate reductase. The conversion between deoxy- and ribonucleotides is ensured by ribonucleosides-diphosphate reductases. *M. florum* encodes a thioredoxin (Mfl178, *trx*) and a thioredoxin reductase (Mfl064, *ntr)*. This system plays an important role in oxidoreductive balance and is present in Mollicutes [50,51]. The conversion of all 4 nucleosides di-phosphate into nucleotides di-phosphate is likely to be catalyzed by the trimer complex formed by Mfl528 (*nrdA* or *nrdE*), Mfl529 (*nrdI*) and Mfl530 (*nrdF*) in a promiscuous manner.

Synthases. A GMP synthase (Mfl342, *guaA*) activity was identified through the four EC numbers identification methods used (EC 6.3.5.2). This enzyme enables the conversion of L-glutamate to L-glutamine, consuming one Xanthosine 5'-phosphate and producing one GMP. The fact that this enzyme is kept in *M. florum* may indicate either the requirement for an easy conversion between amino acids in case of starvation or the need to utilize non-conventional nucleotides like XMP.

A thymidylate synthase (Mfl419, *thyA*) activity was identified through the four EC numbers identification methods used (EC 2.1.1.45). This enzyme converts dUMP into dTMP using folate as a cofactor (5,10-Methylenetetrahydrofolate to 7,8-Dihydrofolate). This mechanism is likely conserved to ensure that accidental deoxidation of UMP into dUMP can be re-utilized.

A cytidine triphosphate synthetase (Mfl648, *pyrG*) activity (EC 6.3.4.2) was also identified through the four EC numbers identification methods used, namely RefSeq, PATRIC, DETECT v2, and COFACTOR . This enzyme enables the production of CTP from UTP, converting glutamine into glutamate in the process. This process is likely conserved to ensure the availability of cytosine based nucleotides and nucleosides. While the DNA uptake remains unclear given the current annotation of transporters, no specific transporter for cytosine was found.

Other nucleobases. Both Mfl074 and Mlf075 are annotated as adenylosuccinate lyase (EC 6.3.4.4 and EC 4.3.2.2, respectively). The reactions ADSS and ADSL1r convert aspartate to fumarate. In the metabolic network, fumarate is a dead-end metabolite. Further biochemical characterizations could link fumarate to other reactions in the network, hereby explaining the conservation of these enzymes.

An adenosine deaminase (Mfl215, *hit1*) matches both EC numbers 3.5.4.2 and 3.5.4.4. These are both responsible for the deamination of adenosine, producing either inosine or hypoxanthine and releasing ammonium. Both of these products are dead-ends in the current metabolic network. One possible explanation for this reaction would be that un-orthodox nucleobases accumulating in the cell through various processes could be converted back to adenosine if this reaction was reversible.

A nucleotidyl hydrolase/transferase (Mfl245, *hit1*) was associated with two EC numbers 3.6.1.17 using Patric or 3.-.-.- with COFACTOR. The confidence in this annotation is weaker given that PATRIC identifies it as a Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical). The

EC number provided by COFACTOR gives only the first digit, which is not really precise. Further experiments could reveal the true activity of this enzyme.

3.11.2.2 Amino acids synthesis

De novo synthesis of amino acid is generally absent from Mollicutes species [46], a feature that was confirmed through the manual curation of *M. florum*'s genome. Hence, salvage of free amino acids or oligopeptides appears as the only viable solution for *M. florum* to sustain protein production and growth. The possibility that both free amino acids and oligopeptides be imported in Mollicutes was previously discussed [5,46,52]. The Amino acid module is composed of two main transporter systems (single amino acid and peptides) that were suggested to import small peptides directly. These peptides are digested within the cell and the resulting amino acids are used to express proteins, a process that avoids the need for any energy expensive synthesis pathways. The apparent low number of transporters compared to the number of substrates (20 for all amino acids) has been suggested as biochemically possible [53] and suiting the genome reduction history of Mollicutes [46]. Eight different genes are annotated in what could compose 3 different systems for amino acid and oligopeptides transport (Tableau S3.2).

Tableau S3.2 Amino acid transporters gene annotation.

Gene	GenBank	PATRIC
Mfl605	amino acid/amine (lysine) APC transporter	Uncharacterized amino acid permease, GabP family
Mfl094	oligopeptide ABC transporter permease	Oligopeptide transport system permease protein OppB (TC 3.A.1.5.1)
Mfl095	oligopeptide ABC transporter permease	Oligopeptide transport system permease protein OppC (TC 3.A.1.5.1)
Mfl096	oligopeptide ABC transporter ATP-binding protein	Oligopeptide transport ATP-binding protein OppD (TC 3.A.1.5.1)

Tableau S3.2 Amino acid transporters gene annotation. (suite)

Mfl097	oligopeptide ABC transporter ATP-binding protein	Oligopeptide transport ATP-binding protein OppF (TC 3.A.1.5.1)
Mfl098	oligopeptide ABC transporter periplasmic binding component	hypothetical protein
Mfl183	amino acid ABC transporter permease	hypothetical protein
Mfl184	amino acid ABC transporter permease	hypothetical protein

In *M. pneumoniae,* no amino acid can be synthesized *de novo* and the defined growth media contains all amino acids [52]. The decision was made to provide the *M. florum* model with exchange reactions for each amino acid. For the import of oligopeptides, we referred to the solution proposed in the *M.genitalium* model (*i*PS189). Here, 15 dipeptide import reactions simulate the import of oligopeptides through oligopeptide ABC transporter and 14 reactions simulate the cleavage by a protease of these dipeptides into free amino acids that can be incorporated into proteins. These 29 reactions were imported from iPS189 and the gene-reaction rule was changed so that the Mfl094 to Mfl098 are associated with each of the import reactions. 8 proteases/dipeptidases are annotated in *M.florum* based on GenBank (Tableau S3.3).

Tableau S3.3 Other amino acid related genes and their annotation.

Gene	**GenBank**	**PATRIC**
Mfl056	Proline dipeptidase	protein co-occurring with transport systems (COG1739)
Mfl263	Intracellular protease/amidase	Hypothetical protein
Mfl287	Membrane associated Zn-dependent protease	Intramembrane protease RasP/YluC, implicated in cell division based on FtsL cleavage
Mfl379	Xaa-Pro dipeptidase	Aminopeptidase YpdF (MP-, MA-, MS-, AP-, NP- specific)
Mfl402	PFPI-like cysteine protease	ThiJ/PfpI family protein

Tableau S3.3 Other amino acid related genes and their annotation. (suite)

Mfl404	class III heat shock DNA-binding ATP dependent Lon protease	ATP-dependent protease La (EC 3.4.21.53) Type I
Mfl564	CAAX amino terminal membrane bound protease	hypothetical protein
Mfl659	zinc metalloprotease	Zinc metalloprotease

Despite further evidence, all proteases but 2 were linked to these dipeptide cleavage reactions. The two proteases not taking in cleavage of imported oligopeptides are the cell-division associated RasP/YluC (Mfl287), potentially involved in cell division and the DNA-binding Lon protease (Mfl404) which seems to be related to heat-shock response.

The free amino acid import was hypothesized to be mediated via either Mfl605 or the complex between Mfl183 and Mfl184. Since Mfl183 and Mfl184 are annotated as hypothetical proteins in PATRIC no further constraint was added through the addition of an ABC transport system requiring ATP. Instead all free amino acid import reactions were considered to be proton symport.

3.11.2.3 Energy production and carbon sources

The Energy module contains reactions associated with the production of ATP, alternate carbon metabolism reactions, oxydoreduction balance, pyruvate metabolism and an ATP pump. As in most mollicutes [5], the tricarboxylic acid (TCA) cycle is absent from *M. florum* and glycolysis is the only ATP generating pathway, with lactate and acetate being the two possible fermentation by-products.

PTS systems. It was reported before that Mollicutes have the capacity to import and convert monosaccharides and phosphorylate them upon entry [54]. This statement is consistent with the current genome annotation which suggests that carbon sources can be incorporated via PTS

systems. Specificity was found for sucrose (Mfl516, Mfl527), glucose (Mfl187, Mfl214), fructose (Mfl181), trehalose (Mfl426, Mfl431, Mfl500) with the phosphotransferase (Mfl519) and phosphohistidine containing protein (Mfl565). Two gene clusters ensure the transport of glycerol-3-phosphate (Mfl023, Mfl024, Mfl025, Mfl026) and ribose (Mfl666, Mfl667, Mfl668, Mfl669) through ABC transport.

Glycolysis. Although carbon sources utilized in glycolysis are phosphorylated through PTS-associated transport, *M. florum*'s genome entails 2 sugar kinases. Glucose kinase (Mfl497) probably phosphorylates the remaining phosphate free glucose molecule after trehalose is cleaved by trehalose-6-phopshate hydrolase. On the other hand, fructose kinase (Mfl514) most likely phosphorylates fructose after the sucrose molecule is cleaved by sucrose-6-phosphate hydrolase (Mfl515 or Mfl526). Aside from this initial phosphorylation step *M. florum*'s glycolysis differs from some previously modelled Mollicutes at the glyceraldehyde-3-phosphate dehydrogenase step. In *M. florum* this enzyme has 2 versions. Mfl578 is annotated as the standard NAD dependent dehydrogenase converting glyceraldehyde-3-phosphate (g3p) into 3-phospho-glycerol phosphate (13dpg), a reducing reaction that produces NADH. The alternative reaction is catalyzed by gene Mfl259 (both PATRIC and GenBank annotations agree for NADP specificity) and converts g3p into 3-phospho-glycerate (3pg), bypassing phosphoglycerate kinase (Mfl577) reaction. This reaction utilizes NADP and generates NADPH.

ATPase pump. Similarly to other Mollicutes, *M. florum* possesses an ATPase pump. Contrary to previous observations that the ATPase of Mollicutes is composed of 7 genes [55], in *M. florum*, a cluster of 8 genes (Mfl109 to Mfl116 inclusively) is proposed to form this complex. Unlike other bacteria where the F_1F_0 ATPase is used to generate energy from a proton gradient, in Mollicutes the ATPase is believed to be used by the cell to maintain an electro-chemical gradient at the cost of ATP. As previously reported the ATPase pump is also essential in *M. florum* with 6 of the 8 genes untouched by any transposon. The first and the last gene in the

genomic sequence were hit by a transposon only in the terminal part of the gene (the last 20%) which could still allow for the complex to form.

Secretion products. In *M. florum*, the enzyme lactate dehydrogenase (Mfl596) and the pyruvate dehydrogenase complex (PDH, Mfl039, Mfl040, Mfl041 and Mfl042) are annotated and would allow two outcomes for pyruvate. The first path through lactate leads to the production of NAD+ and lactate. NAD+ is used in glycolysis again whereas lactate needs to disappear from the system. No transporter was annotated for lactate, hence the orphan reaction L-LACt and the sink SK_L_LAC were added to the network creating an escape route for lactate. The PDH path leads to the formation of acetate for which no transporter was annotated either. Again two orphan reactions were added to eliminate acetate from the system, a transport (ACtr) and a sink (SK_AC).

3.11.2.4 Lipids

The Lipids module contains the necessary machinery to synthesize the single *M. florum* cell membrane. Whole fatty acids are imported through two lipid transport proteins (Mfl590 and Mfl591). These fatty acids are then fixed to a glycerol backbone in a process dependent on the acyl-carrier protein (ACP, Mfl593). In the model, this generic glycerolipid is used to form the different lipid species previously detected in *M. florum* [56].

Identification of lipid synthesis genes. The lipid synthesis network in *M.florum* was reconstructed using the available annotation and experimental lipidomic experiment (Supplementary file 4). Most Mollicutes do not possess the ability to generate long chain fatty acid, an energy extensive process [50]. Lipid metabolism and requirements in Mollicutes is hard to assess [52]. Despite their genetic simplicity Mollicutes have conserved a rather high level of lipid complexity [50]. Although some studies have shown that *Acholeplasma laidlawii* can execute *de novo* synthesis of fatty acids, the majority of less complex Mollicutes cannot execute such task primarily because they lack the necessary machinery for it and also because

the metabolic cost of fatty acid elongation (32mole ATP/fatty acid) that could be too high for these scavengers [57]. The genes potentially coding for lipid biosynthesis in *M. florum* are listed here (Tableau S3.4).

Tableau S3.4 Fatty acid synthesis pathway gene annotation.

Gene ID	GenBank	PATRIC
Mfl230	Glycerol-3-phosphate acyltransferase PlsX	Phosphate:acyl-ACP acyltransferase PlsX (EC 2.3.1.n2)
Mfl286	CDP-diglyceride synthetase	Phosphatidate cytidylyltransferase (EC 2.7.7.41)
Mfl315	Lysophospholipase	Hypothetical prote
Mfl325	Lysophospholipase	Hypothetical protein
Mfl337	Hypothetical protein	Acyl-phosphate:glycerol-3-phosphate O-acyltransferase PlsY (EC 2.3.1.n3)
Mfl382	1-acyl-sn-glycerol-3-phosphate acyltransferase	Acyl-CoA:1-acyl-sn-glycerol-3-phosphate acyltransferase (EC 2.3.1.51)
Mfl384	Holo-ACP-synthase	Holo-[acyl-carrier-protein] synthase (EC 2.7.8.7)
Mfl465	Phosphatidylglycerophosphatase B	Hypothetical protein
Mfl474	Lysophospholipase	Lysophospholipase (EC 3.1.1.5); Monoglyceride lipase (EC 3.1.1.23)
Mfl590	Fatty acid binding/lipid transport protein	DegV family protein
Mfl591	Fatty acid binding/lipid transport protein	DegV family protein
Mfl593	Acyl carrier protein I	Acyl carrier protein
Mfl607	Acyltransferase	N-Acyltransferase ElaA

Experimental identification of lipid species. Extracting the putative lipid biosynthesis genes combined with experimental qualitative lipidomics allowed drawing the pathways leading to lipid production in *M. florum*. *Ad hoc* lipidomics results generated in the complex medium

identified 7 different lipid species (Tableau S3.5, Material and methods). The possibility that these lipids are residual from the undefined growth medium cannot be ruled out, even considering our best efforts to perform adequate wash phases of cells before the experiment and algorithmic noise reduction on the results. Notwithstanding a more precise description of *M. florum*'s membrane composition, the lipid species present in the lipidomics results were considered constituents of *M. florum's* biomass and the metabolic network was reconstructed to ensure their production (adding orphan reactions when necessary).

Tableau S3.5 Conversion of fatty acids classes identified experimentally to BiGG identifiers.

Relative abundance	Lipid name	BiGG identifier
1	Sphingomyelin (SM)	sphmyln_mf_c
2	Phosphatidylcholine (PC)	pc_c
3	Phosphatidylinositol 3,4,5-triphosphate (PIP3)	pail345p_c
4	Phosphatidic acid (PA)	pa180_c
5	Diacylglycerol (DAG)	12dgr180_c
6	Phosphatidylserine (PS)	ps_c
7	Triacylglycerol (TAG)	tag_mf_c

General mechanism. The mechanism for the production of these lipids was assumed to be dependent on the Acyl-Carrier Protein (ACP, Mfl593). This highly conserved protein [58] can fix free fatty acids (FFA). The FFA transport system is potentially executed by Mfl590 and Mfl591, both annotated as "fatty acid binding/lipid transport protein" (GenBank) or a DegV family protein (PATRIC) in Pfam. The decision was made to use a single FFA (Octadecanoate (n-C18:0)) to serve as the fatty acid chain for all lipid classes in the model. The elongation of fatty acids is generally absent in Mollicutes [50] and no gene was identified that could catalyze this process. Since no elongation was modelled, the length of the fatty acid does not add a constraint on the system. If *M.florum* is presented with many different FFA in a complex

growth media, these FFA may be imported in the cell and next included in the cytoplasmic membrane. Upon activation of the ACP (Mfl384), the putative mechanism would involve the fixation of the FFA by an acyltransferase (Mfl607) yielding a FFA bound ACP. Fixation of the fatty acid chain to the glycerol backbone requires the production of a phosphorylated FFA (Mfl230) that can be fixed to the glycerol backbone (Mfl337). This FFA bound glycerol is converted into phosphatidic acid upon fixation of another fatty acid (Mfl382).

Phosphatidic acid derivatives. Previous experimental results show that *M. florum's* lipidic composition entails phosphatidylcholine, phosphatidylinositol-3-phosphate, phosphatidic acid and phosphatidylserine (PC, PIP3, PA and PS, respectively) [56]. The suggested general mechanism for assembly of lipids in *M. florum* would result in the formation of PA. The formation of the 3 other species (PC, PIP3 and PS) would result from the addition of the specific head on PA. Choline kinase was annotated in RefSeq but not in PATRIC. 3 orphan reactions were added to fulfill the gap in choline production by the model (CHOLt, CHLPCTD, DAGCPT_mf). These reactions sequentially catalyze the fixation of choline-phosphate on CDP and further the fixation of choline from CDP-choline onto diacylglycerol (12dgr180_c). A similar modelling decision was taken for the formation of PIP3 where no transporter exists for inositol and the fixation of inositol on CDP requires the addition of an orphan reaction (CDIPT). PS also does not have an annotated gene for its synthesis from PA. While it is possible that the annotated Phosphatidylglycerol synthase (Mfl663) catalyzes this reaction, no evidence can confirm it. An orphan reaction (PSSA_mf) was therefore added for the synthesis of PS for *M. florum*.

Di-acylglycerol (DAG). Diacylglycerol can be formed from PA. The current annotation does not contain any phosphatidate phosphatase that would be required for the synthesis of this metabolite. An orphan reaction (PAPA180) was added to satisfy this need.

Cardiolipin and phosphatidylglycerol. Cardiolipin is a component of cell membrane in all 3 domains of life [59] and a ubiquitous component of the core biomass for prokaryotes as revealed

by Rocha and colleagues [60]. While cardiolipin was not specifically identified in previous lipidomics results, Mfl626 is annotated as a cardiolipin synthase in both RefSeq and PATRIC. Phosphatidylglycerol (PG) was detected slightly in lipidomic and may be produced by *M. florum*. The presence of PG could be associated with cardiolipin due to its structure (also known as di-phosphoglycerol). The entire pathway for the synthesis of cardiolipin is annotated in *M. florum* so the reactions were added to the model.

Sphingomyelin. In relative abundance, sphingomyelin is the first ranked lipid in previously generated lipidomic experiments [56]. Nevertheless, bacteria do not possess the capacity to produce sphingomyelin, an essential component of nerve tissue in mammalian cells [61]. Therefore, if this compound is really present in the *M. florum* membrane it would be the result of a direct salvage from the environment. It has been reported that Mollicutes possess lipid salvage capability [62,63]. Sphingomyelin has also been shown to favor growth in a defined media for some *Spiroplasma* species [64]. Despite these observations, no gene could be attributed to the import of sphingomyelin by *M. florum*. We hypothesized that the favored growth in presence of sphingomyelin was due to the increased lipid solubility which would facilitate the import of free-fatty acids from the medium. Given its high abundance in the lipidomics dataset, sphingomyelin was added to the model and to the BOF. Characterizing the cell membrane again, in a completely defined medium, would help determine the role and importance of sphingomyelin in *M. florum*.

3.11.2.5 Glycans

A similar data-driven approach was used for the reconstruction of the Glycan module, which contains the reactions responsible for the synthesis of the extracellular polysaccharide layer previously described for *M. florum* [56]. For modelling purposes, the synthesis of the capsular polysaccharide (CPS) was assumed to include the conversion of sugars (glucose, galactose, mannose and rhamnose) in a sugar-1-phosphate form and their fixation onto a nucleotide backbone. The only predicted glycosyltransferase (Mfl568) in *M. florum* was assumed to

assemble the CPS directly on a diacylglycerol on the intracellular side of the membrane. The CPS is next transferred on the extracellular milieu by the flippase (Mfl562) (Tableau S3.6).

Many Mollicutes species have been reported to produce a thick layer of polysaccharide composed of neutral sugars, either attached to the membrane (Capsular polysaccharide, CPS) or secreted (Extracellular polysaccharide, EPS) [65–67]. For instance, Blanchard and colleagues identified the thickness of the *Mycoplasma penetrans* CPS at 11-13nm. In *M. florum*, transmission electron microscopy (TEM) showed a membrane thickness (8.18 to 18.35nm) that would exceed the expected width for a single lipid bilayer (4nm) [56]. The composition of the *M. penetrans* CPS assessed by gas liquid-chromatography (GC) revealed that it is composed of four different sugars: mannose, glucose, N-acetylglucosamine and N-acetylgalactosamine [66]. Like *M. penetrans*, GC results for *M. florum* revealed that it is composed of 4 components. Glucose and mannose were also found in *M. florum*, but the amino-sugars N-acetylglucosamine and N-acetylgalactosamine were not found. The two remaining components identified were rhamnose and galactose.

The genome annotation of *M. florum* contains both a glucosyltransferase (Mfl568) and a *Wzx* flippase (Mfl562). In *Escherichia coli* and *Salmonella* strains the pathway for O-antigen synthesis using Wzx/Wzy proteins can be summarized in a sequence of 5 events [68]:

1. The sugars are imported and phosphorylated in the process, usually via a PTS system.
2. The phosphate group is transferred onto the first carbon, hereby labelling the sugar for polysaccharide synthesis.
3. The individual sugars are fixed to a triphosphate nucleotide via a nucleotidyltransferase.
4. The sugars are polymerized into a chain by a glycosyltransferase, using the energy contained in the phosphate bond with the nucleotide diphosphate.
5. The polymerized glycan is flipped on the extracellular side of the membrane by a flippase.

We suggested that a glucose transporter, either Mfl217 or Mfl187, could be promiscuous and allow the entry of sugar molecules composing the CPS. The GC results also revealed the presence of rhamnose. While this sugar is similar to the other two, it lacks a hydroxyl group which is necessary for its phosphorylation upon entry. Therefore, the import of rhamnose was not associated with a gene and is included in the functions in search for a gene (Supplementary file 4).

Sugars imported through the PTS system should be phosphorylated on carbon 6. In order to be included in a polysaccharide, the phosphate group should be transferred on the first carbon. Mfl120 is annotated in RefSeq as a phosphomannomutase while being a D-Ribose 1,5-phosphomutase in PATRIC. Also, our re-annotation process allowed identifying 3 different EC numbers for this gene (Supplementary file 3). Together these observations suggest that this enzyme is promiscuous. The conversion of hexose-6-phosphate to hexose-1-phosphate was therefore assigned to this gene for all sugars. Hexose-1-phosphate sugars are fixed to a nucleotide-triphosphate via a nucleotidyl transferase/hydrolase. We suggest that Mfl245 takes care of that function for all sugars.

Aside from the TEM images, one of the strongest evidence for the presence of CPS in *M. florum* is the annotation of both a glycosyltransferase (Mfl568) and a nearby O-antigen flipase/transporter (Mfl562). The addition of the first sugar to the diacylglycerol backbone and further elongation may be conducted by the only glycosyltransferase (Mfl568). Additional protein characterization may reveal the importance of other yet unannotated proteins in this process. Once the chain reaches a certain length, the instability may trigger the end of the elongation, preventing the attachment of the glycosyltransferase. The fully elongated chain may then be flipped on the other side of the membrane by the flippase (Mfl562). A polymer length of 13.27nm may allow chains of 7 or 8 sugars. The proportions of each sugar identified in *M. penetrans* were 1:6:1:2 (mannose, glucose, N-acetylglucosamine, N-acetylgalactosamine). Here mannose, glucose, rhamnose, galactose were identified through

GC experiments in *M. florum* at a ratio of 1:4:4:11, and the CPS was incorporated in the model with this stoichiometry.

Tableau S3.6 Glycan synthesis pathway gene annotation.

Gene	GenBank	PATRIC
Mfl497	Glucose kinase	Putative sugar kinase
Mfl120	phosphomannomutase	D-Ribose 1,5-phosphomutase (EC 5.4.2.7)
Mfl187	PTS system glucose-specific transporter subunit IIA	PTS system, glucose-specific IIA component (EC 2.7.1.199)
Mfl214	PTS system glucose-specific transporter subunit IIABC	PTS system, N-acetylglucosamine-specific IIA component (EC 2.7.1.193) / PTS system, N-acetylglucosamine-specific IIB component (EC 2.7.1.193) / PTS system, N-acetylglucosamine-specific IIC component
Mfl245	Nucleotidyl transferase	Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical) (EC 3.6.1.17)
Mfl562	Transporter	O-antigen flippase Wzx
Mfl568	Glycosyltransferase	Hypothetical protein

3.11.2.6 Vitamins & cofactors

As for lipids and glycans, the synthesis of vitamins and cofactors in M. florum is very minimal. We describe here the pathways leading to the import and utilization of coenzymes that were identified in the annotation and used in the reconstruction process.

Nicotinamide adenine dinucleotide. Both phosphorylated and unphosphorylated forms of nicotinamide adenine dinucleotide (NADP and NAD) are found in reactions of the metabolic network. Additionally, this coenzyme has a detailed pathway for incorporation in *M. florum*. While no transporter is specifically annotated for its import, NAD is a combination of two

nucleotides joined by their phosphate groups, it is possible that this configuration allows it to be imported through the same transporter as nucleobases (discussed above).

Upon import, four enzymes compose this pathway. Nicotinamide is converted to nicotinate through the nicotinamidase (Mfl340, EC 3.5.1.19). A nicotinate phosphoribosyltransferase is also present in the genome (Mfl588, *pncB*). The current RefSeq annotation identified the EC number 2.4.2.11 which is obsolete according to KEGG [37]. The replacement EC number (EC 6.3.4.21) was correctly identified by both PATRIC and DETECT while COFACTOR also attributed the old EC number 2.4.2.11. An adenylyltransferase (Mfl373, *nadD*) and a NAD synthase (Mfl521, *outB*) catalyze the last two steps of this pathway leading to the formation of NAD. Also, a NAD kinase is annotated (Mfl193, *ppnk*), which supports the presence and utilization of NADP in the metabolic network.

Folate. A major pathway present in *M. florum* is the formation of folate and derivatives. A folate specific transporter is annotated in PATRIC (Mfl061 or Mfl086). A dihydrofolate reductase (Mfl383, *folA*), enables its entry into the folate pathway.

Key metabolic reactions involved in the folate pathway include : the thymidylate synthase (Mfl419), producing dTMP from dUMP; the glycine hydroxymethyltransferase (Mfl106), reversibly producing serine from glycine; and the Methionyl-tRNA formyltransferase (Mfl409), producing formylmethionine, essential to initiate the translation of proteins.

Coenzyme A. In the metabolic network, Coenzyme A is used in the biosynthesis of lipids to activate the apo-Acyl-carrier protein and in the pyruvate dehydrogenase complex. Metabolically speaking, this coenzyme is therefore essential. Nevertheless, its import and synthesis remains under characterized in *M. florum*. Indeed, no transport reaction could be found that imports CoA specifically and a single enzyme, diphosphoCOA kinase (Mfl281) is annotated.

Lipoate. Lipoate is present in the pyruvate dehydrogenase complex where a lipoyl-adenylate protein ligase (Mfl038) is present. The import of that coenzyme is absent as well as a potential pathway to its synthesis.

Thiamine. A thiamine diphosphokinase (Mfl224) is annotated in PATRIC. A consistent EC number (EC 2.7.6.2) was attributed to this gene by both PATRIC and COFACTOR, which is interesting since RefSeq identified it as a "hypothetical protein". The presence of thiamine in the network is therefore supported by this annotation.

Riboflavin. A rather complete pathway leads to the formation of flavin adenine dinucleotide (FAD) in *M. florum*. This pathway includes an annotated transporter for riboflavine (Mfl576). The two following reactions are catalyzed by genes with two different EC numbers. For both Mfl283 and Mfl334 genes, PATRIC and RefSeq annotations suggest two reactions : riboflavin kinase (EC 2.7.1.26) and FAD synthase (EC 2.7.7.2).

Polyamines. Spermidine and putrescine have annotated transporters in *M. florum*. One is a spermidine/putrescine ABC transporter composed of three genes (Mfl509, Mfl 510 and Mfl511). The other is a putrescine/ornithine APC transporter (Mfl664).

Minerals. The import of minerals in the metabolic network was first considered based on the known annotation. The manual curation of the genome identified inorganic phosphate, magnesium, cobalt, zinc, potassium and sodium as potentially imported ions. Some key minerals were not gene-associated but nevertheless included in the model since they represent universally essential cofactors in prokaryotes [60].

The import of inorganic phosphate is also annotated in JCVI-syn3.0A [16] and was associated with three genes in *M. florum* (Mfl233, Mfl234 and Mfl235). Two EC numbers (EC 3.6.3.27 or 3.6.3.33) could be identified for one of these genes (Mfl235), the ATP-binding protein of

the complex. Interestingly, this three-gene cluster has a transcriptional regulator right next to it, suggesting an operon-type regulation and an important feature of *M. florum*'s metabolism.

Magnesium is essential for the polymerization of nucleic acids and a specific ATPase transporter is present to ensure its import (Mfl496). Other genes are also linked to its transport through the cell membrane (Mfl217 and Mfl356). These transporters nevertheless seem to serve a more general purpose of large cation import/export, as revealed by their annotation (Mfl217 : Mg/Co/Ni transporter MgtE, CBS domain-containing; Mfl356 : Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5); Copper-translocating P-type ATPase (EC 3.6.3.4). Other transporters responsible for the evacuation of metals are present in the *M. florum* annotation, namely the energy-coupled factor (ECF) complex (Mfl152, Mfl153, Mfl154), which corresponds to the *cbiO* transport protein of *Salmonella paratyphi A*.

Finally, potassium and sodium may be imported through the three-gene complex (Mfl164, Mfl165 and Mfl166) that are annotated as : K+, Na+ uptake protein integral membrane subunit, and the *trkA* and *trkH* genes.

3.11.3 Medium simplification and growth kinetics

3.11.3.1 i*n vitro* growth medium

The initial culture medium of M. florum was the ATCC1161, a complex mixture of heart infusion broth, horse serum, and yeast extract. Simultaneously removing all three undefined components would have significantly reduced the chance of generating a functional growth media. Instead, we set to reduce the necessary concentration of undefined components by supplementing with a completely defined cell culture medium. Four completely defined cell culture media were tested (CMRL-1066, DMEM/F-12, Medium 199 and CHO-200). When supplementing with these media, the heart infusion broth could be entirely removed. Each defined media was supplemented with a decreasing concentration of both serum and yeast

extract (SYE) and the growth of M. florum monitored. Similar to previous studies interested in obtaining a defined medium for Mollicutes 64,69, the CMRL-1066 based medium yielded better growth on a reduced concentration of SYE (0.313% horse serum, 0.02% yeast extract). The composition of the four media were compared to identify key components allowing M. florum growth. Ten CMRL-1066 specific components were identified (Tableau S3.7).

Tableau S3.7 Comparison of media compositions

	CMRL-1066	**Medium 199**	**CHO-200**	**DMEMF-12**
Total number of components	55	63	51	49
Shared with CMRL-1066	55	44	38	37

A comparison of the components present in the different growth media against CMRL-1066 was performed (Tableau S3.8). Coenzyme A (CoA) stood out as being the only component present in CMRL-1066 systematically missing in the other growth media. As identified in the metabolic reconstruction, CoA is a key coenzyme for the activity of the pyruvate dehydrogenase complex and the activation of the ACP, which is key for the assembly of membrane lipids. On the other hand, putrescine was the only component present in all media but CMRL-1066. Different polyamines may be imported by *M. florum* and it is possible that spermidine is the only required nutriment.

Tableau S3.8 Components specific to CMRL 1066 or specific to the other medium.

	Medium 199	**CHO-200**	**DMEMF-12**
CMRL vs Medium	Coenzyme A, Magnesium, Sodium	Coenzyme A, Magnesium', L-Ascorbate, Acetate	Coenzyme A, L-Ascorbate, Acetate
Medium vs CMRL	Putrescine, Adenine, Guanine, Uracil, Xanthine, D-ribose	Putrescine, Copper, Iron, Zinc	Putrescine, Copper, Iron, Zinc

3.11.3.2 *in silico* growth medium

The composition of the four defined growth media (Supplementary file 7), revealed ten nutrients specific to CMRL 1066. Converting all media components to BiGG identifiers 70 allowed comparing the composition of CMRL 1066 to metabolites in the reconstruction. The majority (43/55) of the CMRL 1066 components were found in the metabolic reconstruction and, more specifically, 31 were found among the 84 metabolites part of the in silico medium. Only two of the ten CMRL 1066 specific nutrients were absent in the model (glucuronate and hydroxyproline). Of the 8 metabolites remaining, coenzyme A is the sole found in the in silico medium (Tableau S3.9). The seven remaining metabolites present in the model but not specifically part of the in silico medium are three deoxynucleosides and four essential cofactors.

All 10 metabolites are interesting for the development of a fully characterized minimal medium. While both glucuronate and hydroxyproline were absent from the metabolic reconstruction, they should be part of the screen for a minimal medium. Indeed, glucuronate is known to take part in the formation of proteoglycan [71] and its absence in the reconstruction may be attributable to the lack of confidence in this part of the metabolic network (Figure 3.3). The 7 components that were identified in the model but not in the model-medium should also be considered, since the metabolic need is surely present but the transporter specificity for the exact form of these nutrients could be broadened or more specifically defined once tested in a completely defined medium.

Tableau S3.9 *In silico* minimal medium composition.

Carbon sources	Amino acids		Vitamins	Minerals
1. D-Fructose 2. Sucrose 3. D-galactose	4. N-L-Alanyl-L-leucine 5. N-L-alanyl-L-threonine	12. L-Histidine 13. L-Isoleucine 14. L-Methionine 15. L-Lysine	22. Coenzyme A 23. Folate 24. Nicotinamide	25. Calcium 26. Co2+ 27. Chloride 28. magnesium

Tableau S3.9 *In silico* minimal medium composition. (suite)

	6. L-Arginine 7. L-Asparagine 8. Cys-Gly 9. L-Glutamate 10. N-glycyl-L-aspartic acid 11. L-Glycylglutamine	16. L-Tryptophan 17. L-Tyrosine 18. L-Valine 19. L-Serine 20. L-Proline 21. L-Phenylalanine		29. Mn2+ 30. Molybdate 31. Sodium 32. potassium 33. Nickel 34. Zinc
Nucleotides	**Lipids**		**Polyamines**	
35. Adenine 36. Guanine 37. Uracil	38. Octadecanoate (n-C18:0) 39. Glycerol 3-phosphate 40. Choline 41. N-acylsphingosine/Ceramide		42. Spermidine 43. Putrescine	

The metabolic reconstruction provided 84 transport reactions and extracellular metabolites. To define a growth media for simulation, the extracellular metabolites were compared with the defined composition of the CMRL-1066 media. Of the 55 components included in CMRL-1066 medium, 36 were present in the original extracellular metabolites and 19 were missing. The missing components are evaluated individually (Tableau S3.10):

Trans-4-hydroxy-proline: hydroxyproline is a component of collagen. This metabolite is present in *S. cerevisiae* where a hydroxyproline reductase is present [18]. This reaction is absent from *M. florum*. This metabolite is not likely to be necessary for *M. florum*'s growth and is therefore not added to the model.

4-aminobenzoate: aminobenzoate can be converted into dihydropteroate (EC 2.5.1.15). This EC number is absent in *M. florum* L1 (see all_ec_numbers supplementary sheet). It is not impossible that 4-aminobenzoate is used to produce folate in *M. florum* but not enough evidence is present to add the compound to the medium.

Biotin: Biotin has been reported to be an essential cofactors in bacteria but in pathways that

are unnecessary for *M. florum* [72]. These pathways include fatty acid biosynthesis, replenishment of the tricarboxylic acid (TCA) cycle and amino acid metabolism. Since *M. florum* does not contain a TCA cycle, nor any elaborate fatty acid or amino acid biosynthesis pathways, this vitamin is therefore not likely to be used.

Thiamin diphosphate: a complete pathway with a specific transporter is annotated in *M. florum* for thiamin. This coenzyme is used by the pyruvate dehydrogenase complex, an enzyme essential for the production of acetate in the presence of oxygen. Thiamin is likely to be used by *M. florum*, it is nevertheless likely that thiamin diphosphate is used not imported directly but rather thiamin itself.

Deoxyadenosine, deoxyguanosine, deoxycytidine: deoxynucleoside present in the media may be imported by *M. florum* but this is an hypothesis to be validated with a completely defined media. The current model allows the import of guanosine, guanine, uracil and thymidine.

Flavin adenosine dinucleotide (FAD): the precursor to FAD, riboflavin, is present in CMRL-1066 and both the transporter and pathways for the production of FAD from riboflavin are present in *M. florum*. This metabolite is not likely to be used by *M. florum*.

Nicotinamide adenine dinucleotide (Phosphate) (NAD, NADP) and nicotinic acid: the precursor of these three components, nicotinamide, is present in CMRL-1066 and both the transporter and pathways for the production of NAD from nicotinamide are present in *M. florum*. These three components are likely to be redundant with the presence of nicotinamide.

Pantothenate: this coenzyme is involved in the synthesis of coenzyme A which is already present in the media and which *M. florum* is incapable of fabricating. This component is likely to be non-essential for *M. florum*.

Pyridoxine and pyridoxal: also known as vitamin B6, pyridoxine was not directly involved as a coenzyme in any metabolic reactions. Whether or not *M. florum* requires this coenzyme for growth should be assessed upon the elaboration of a completely defined medium.

Sulfate: cysteine desulfurase seems to be taking the role of providing the cell with sulfur, an essential metabolite.

Glutathione: as a tripeptide, this medium component could be imported by the peptide importer system. Given its role as an antioxidant, its import could reduce the susceptibility to oxidative stress in *M. florum* when added to the medium. This hypothesis could be tested by growth assays under oxidative stress with or without glutathione.

Glucuronate: this monosaccharide is involved in proteoglycan synthesis in many species. It is likely that its import could be done by one of the sugar importers and it could potentially contribute to the synthesis of the *M. florum* glycans.

Cholesterol: helps to solubilize the free fatty acids and facilitate their import.

Tableau S3.10 Comparison of the ten CMRL 1066 specific nutrients with model metabolites.

BiGG ID	Name	Status
4hpro_LT	Hydroxyproline	Not in model
Glcur	Glucuronate	Not in model
Thmpp	Thiamine diphosphate	In model
Dad__2	Deoxyadenosine	In model
Dgsn	Deoxyguanosine	In model
Dcyt	Deoxycytidine	In model
Fad	Flavin adenosine nucleotide	In model

Tableau S3.10 Comparison of the ten CMRL 1066 specific nutrients with model metabolites. (suite)

Nad	Nicotinamide dinucleotide	In model
Nadp	Nicotinamide dinucleotide phosphate	In model
Coa	Coenzyme A	*In silico* medium

3.11.4 Conversion into a mathematical format

3.11.4.1 Biomass objective function

The biomass objective function (BOF) represents the sum of all metabolic goals of an organism in a given environment. In order to be representative of the cellular state, the biomass function should be derived from experimental measurements. Previous work yielded the detailed composition of *M. florum* [56]. Along this data, the BOFdat software [73] was used to determine the biomass precursors to include to the BOF and their respective stoichiometric coefficients. Genomic (DNA), transcriptomic (RNA) and proteomic (proteins) data along with macromolecular weight fractions (MWF) for each category were used as input to determine stoichiometric coefficients using the Step 1 of BOFdat (Tableau S3.11).

The Step2 of BOFdat identified 16 coenzymes and cofactors to be added to the biomass. ions that are commonly found in bacteria and also identified in the reconstruction. Eleven ions were identified in this step (calcium, manganese, cobalt, molybdate, chloride, sodium, ammonium, zinc, potassium, nickel and magnesium). The only coenzyme identified was nicotinamide and its derivatives : oxidized and reduced versions of the phosphorylated and non phosphorylated forms.

Lipids and glycans were not added to the equation by the first and second step. The decision was made to forgo their addition in these steps since the experimental data required curation. Their inclusion was left to the unbiased genetic algorithm performed in BOFdat Step3.

Accordingly, two lipids were identified (phosphatidylcholine, phosphatidylserine). Supporting the evidence for the presence of phosphatidylcholine in the membrane was the identification of the phosphorylated version of choline. The Acyl-carrier protein was also added given its importance for the synthesis of lipids and its ubiquitous presence in prokaryotes. The capsular polysaccharide metabolite formulated during the reconstruction was also added during this step.

Interestingly, S-adenosyl methionine was identified. This metabolite is a common cosubstrate involved in the transfer of methyl groups and is excessively important in many organisms. Consistent with this identification, methyltetrahydrofolate and sulfur were also identified and added during Step3.

The polyamines spermidine and putrescine were added during Step3. The exact function of these metabolites is not precisely known in prokaryotes but they are found widely across species. As mentioned previously, putrescine was found in three of the four medium tested but not in the favored CMRL-1066. Depriving *M. florum* from either of these polyamines could shed light on their function in prokaryotes.

Finally, cytidine and adenosine were identified by BOFdat Step3, which can probably be attributed to the essentiality of the genes that make these specific metabolites. This likely means that some routes that were proposed in the nucleotide salvage pathway are not actually possible *in vitro*.

Tableau S3.11 Detail of the final biomass composition.

Proteins (46.6%)		**Others (1.2%)**	
1. L-Alanyl-tRNA(Ala) 2. L-Arginyl-tRNA(Arg)	10. L-Histidyl-RNA(His) 11. L-Isoleucyl-tRNA(Ile) 12. L-Leucyl-tRNA(Leu) 13. L-Lysine-tRNA (Lys)	21. S-Adenosyl-L-methionine	31. Ammonium 32. Nickel 33. Putrescine 34. Spermidine

Tableau S3.11 Detail of the final biomass composition. (suite)

3. L-Asparaginyl-tRNA(Asn) 4. L-Aspartyl-tRNA(Asp) 5. L-Cysteinyl-tRNA(Cys) 6. L-Glutaminyl-tRNA(Gln) 7. L-Glutamyl-tRNA(Glu) 8. L-Valyl-tRNA(Val) 9. Glycyl-tRNA(Gly)		14. L-Methionyl-tRNA (Met) 15. L-Phenylalanyl-tRNA(Phe) 16. L-Prolyl-tRNA(Pro) 17. L-Seryl-tRNA(Ser) 18. L-Threonyl-tRNA(Thr) 19. L-Tryptophanyl-tRNA(Trp) 20. L-Tyrosyl-tRNA(Tyr)	22. 5,10-Methylenetetrahydrofolate 23. Nicotinamide adenine dinucleotide 24. Nicotinamide adenine dinucleotide - reduced 25. Nicotinamide adenine dinucleotide phosphate 26. Nicotinamide adenine dinucleotide phosphate – reduced 27. Manganese 28. Molybdate 29. Sodium 30. Magnesium	35. Sulfur 36. Adenosine 37. Cytidine 38. Zinc 39. Calcium 40. Choline phosphate 41. Chloride 42. Potassium 43. Co2 44. H2O 45. Acyl carrier protein
RNA (22.9%)	**DNA (7.7%)**	**Lipids (18.3%)**	**Glycans (4.1%)**	
46. ATP 47. CTP 48. UTP 49. GTP	50. dATP 51. dCTP 52. dGTP 53. dTTP	54. Phosphatidylserine 55. Phosphatidylcholine 56. Phosphatidylglycerophosphate 57. Sphingomyelin betaine	58. Capsular polysaccharide *Mesoplasma florum*	

3.11.4.2 Sensitivity analysis

The main carbohydrate provided in ATCC 1161 medium was sucrose. Hence, when grown in CSY medium, *M. florum* was also provided sucrose and its specific uptake rate was measured with high-performance liquid chromatography (HPLC). The obtained value was -5.61 mmol gDW^{-1} h^{-1}, which is similar to previously published results for *M. pneumoniae* (7.37 mmol gDW^{-1} h^{-1}) and *M. gallisepticum* (16.53 mmol gDW^{-1} h^{-1})[8,9]. Nevertheless, the glucose uptake rate was measured in these species, and, to our knowledge, the sucrose uptake rate calculated here is a first amongst mollicutes. We compared our value to that of *E. coli* where sucrose uptake rates were measured and ranged between 7.01 and 14.10 mmol gDW^{-1} h^{-1} following adaptive laboratory evolution [74].

The secretion rates were obtained for both lactate and acetate, the two possible fermentation products in *M. florum* (Figure 3.2). The cumulative value for both products was 8.69 mmol gDW^{-1} h^{-1}, which is exactly in the range of acetate secretion rates in *E. coli* (4.2 to 15.9 mmol gDW^{-1} h^{-1}), slightly higher than *M. pneumoniae* (6.93 mmol gDW^{-1} h^{-1}) and lower than *M. gallisepticum* (10.2 mmol gDW^{-1} h^{-1}).

While an acetate secretion rate was measured in *M. pneumoniae*, a lactate secretion rate was measured in *M. gallisepticum*. The metabolic reconstruction of *M. florum* shows that both secretion pathways are complete and should be functional. Since our current experimental setup did not allow the direct measurement of each secretion product, the exact secretion rates still remain hypothetical. The model can nevertheless reveal some interesting trade-offs around the production of lactate and acetate. The path to lactate secretion is rather simple with a single enzyme, lactate dehydrogenase (Mfl596, *ldh*), converting pyruvate into lactate. The secretion of acetate is more intricate and involves a complete operon composed of the lipoate-ATP adenylate transferase (Mfl038), the complete pyruvate dehydrogenase complex (Mfl039, *pdhA*; Mfl040, *pdhB*; Mfl041, *pdhC*; and Mfl042, *pdhD*), a phosphotransacetylase (Mfl043, *pta*), and an acetate kinase (Mfl044, *ackA*). Contrary to lactate production, the path to acetate releases CO_2 as a metabolic waste, yields one ATP but reduces one NAD molecule into NADH. While the production of one molecule of ATP seems profitable for the cell, the one NADH molecule must be re-oxidized to make this process sustainable. One key reaction involved in this process is the NADH oxidase (Mfl037, *nox*). This enzyme uses molecular oxygen (O_2) to convert NADH back to NAD. This reaction exists in two forms: H_2O producing or H_2O_2 producing. While producing water molecules is not harmful for the cell, hydrogen peroxide is a toxic waste that needs to be eliminated. While this task could be achieved by the L-methionine S oxide reductase (Mfl050, *msrA*), the specificity to H_2O_2 is not confirmed. The final model therefore uses the NOX2 reaction (BiGG identifier), which produces water instead of hydrogen peroxide. Detecting the production of hydrogen peroxide by *M. florum* could shed light on this process.

The acetate production acetate can be probed using *i*JL208. In the final version of the model, the lactate secretion was favored since the expression of the lactate dehydrogenase was much higher than the pyruvate dehydrogenase complex. To favor the production of lactate, restrictive bounds were applied to key reactions. The upper bound to the NOX reaction was fixed at 5 mmol gDW^{-1} h^{-1}. This limits the amount of oxygen that can be used to oxidize NADH back to NAD which can be used in the glycolysis. Since *M. florum* is a facultative aerobe, the logical decision was to limit the impact of oxygen on its growth phenotype. Considering that the NADH oxidase does not have an unlimited capacity was one option, the other was to reduce the possibility for oxygen import. This could be done by reducing the lower bound of its exchange reaction (EX_o2_e).

To probe the production envelope of acetate, the bounds can be changed on these critical reactions. Providing equal lower and upper bounds on the secretion of lactate and acetate, here 0 and 10, releases the experimental constraints. With these bounds applied, increasing the limit oxygen uptake rate and the upper bound on the NADH oxidase reaction eventually results in a favorable utilization of the acetate secretion pathway. In this small case study, this was observed when the bounds were at 25, which is >1.5 times the upper bound on the secretion rates or approximately > 5 times the sucrose uptake rate.

These model predictions stating that a very high amount of oxygen is required to efficiently produce acetate are consistent with the expression levels of *ldh* and *pdh* genes which suggest a higher production of lactate. The settings implemented in the final version of *i*JL208 ensured that the ATP synthase pump was essential as observed experimentally, which also supported the final choice of constraints.

3.11.5 Validation of model phenotypic predictions

3.11.5.1 Carbohydrates utilization

Reducing the concentration of rich undefined components in the medium revealed a clear difference between sucrose supplemented medium and a no-sugar control, further enabling to validate the assimilation of 14 different carbohydrates by *M. florum.* Upon comparison with model predictions, 8 no-growth and 4 growth phenotypes were correctly predicted. Two additional sugars, mannose and maltose, were found to be utilized by *M. florum* but had not been predicted by the model.

To recover those phenotypes, the alternate carbon metabolism of *M. florum* was studied, seeking enzymes that would likely carry a promiscuous activity. The three-dimensional structures reconstructed with I-TASSER were leveraged for that task. While the specificity of transporters could not be addressed with this method, downstream enzymes allowing the catabolism of mannose and maltose could be compared with the protein databank (PDB) using the FATCAT 2.0 server [75]. Specifically, the annotation of three enzymes (Mfl120, Mfl254 and Mfl499) involved in the assimilation of glucose and trehalose were considered. Using the same approach, the specificity of an aldolase (Mfl121, *deoC*) was assessed to recover the expression phenotype of enzymes of the pentose phosphate pathway (Tableau S3.12).

Tableau S3.12 Main candidates following structural comparison using FATCAT.

Locus tag	Original RefSeq annotation	PDB match	FATCAT p-value	Suggested promiscuous reaction catalyzed
Mfl499	Trehalose-6-phosphate hydrolase (TRE6PH)	5zcbA	0*	Hydrolysis of maltose-6-phosphate to produce glucose and glucose-6-phosphate
Mfl254	Glucose-6-phosphate isomerase (PGI)	1tzbA	8.68e-12	Conversion of mannose-6-phosphate to fructose-6-phosphate

Tableau S3.12 Main candidates following structural comparison using FATCAT. (suite)

Mfl120	Phosphomannomutase (PMANM)	1k2yX	0*	Conversion of glucose-6-phosphate to glucose-1-phosphate
Mfl121	2-deoxyribose-5-phosphate aldolase (DRPA)	3igx	4.86e-5	Pentose phosphate pathway transaldolase

*p-values of 0 mean that the PDB match was used as template by I-TASSER for reconstruction of the 3D structure.

The structural similarity between maltose and trehalose suggested they could use the same route into glycolysis. While the promiscuity of the transporter used to import maltose could not be tested *in silico*, it was hypothesized that the trehalose hydrolase (Mfl499) could also hydrolyze maltose. To generate a 3D structure for Mfl499, I-TASSER used the *Bacillus subtilis* α-glucosidase BspAG13_31A (PDB: 5zcc) as a template given the similarity of both sequences. This template structure was shown to have a high-specificity to α-(1-4)-glucosidic linkage [76]. The reconstructed structure was compared to the template used by I-TASSER with FATCAT (p = 0.00, Figure S3.9). The sequence and structural similarity with an enzyme capable of acting on both maltose and trehalose supports the hypothesis that Mfl499 is involved in maltose assimilation. The addition of both the promiscuous transport and digestion reactions were sufficient to provide a growth prediction on maltose.

The capacity of *M. florum* to metabolize mannose could be explained if the glucose-6-phosphate (G6P) isomerase, PGI, (Mfl254) was able to convert mannose-6-phosphate (M6P) into fructose-6-phosphate (F6P), hereby entering glycolysis. The reconstructed structure of the *M. florum* PGI was compared to that of *Pyrobaculum aerophilum* [77] (PDB:1TZB), known for its capability of converting either G6P or M6P into F6P. The structural similarity between these enzymes (p = 8e-12, Figure S3.9) was consistent with this hypothesis and the model was modified accordingly.

The utilization of mannose was also studied in the context of glycan synthesis. Gas chromatography previously revealed the presence of both glucose and mannose in the CPS of *M. florum* [56]. The presence of a phosphomannomutase, PMM, (Mfl120) in the annotation suggested the conversion of M6P in mannose-1-phosphate (M1P), a necessary precursor for glycan synthesis [78]. The template used by I-TASSER for the reconstruction of the 3D structure of Mfl120 was the PMM/PGM structure from *Pseudomonas aeruginosa* (PDB:1K35). In this organism, the enzyme is necessary for the production of exopolysaccharides [79] with G6P and M6P entering the glycan synthesis pathway through the same enzyme. Given the structural similarity of the *M. florum* and *P. aeruginosa* enzymes (Figure S3.11), the promiscuous mutase reaction catalyzed by Mfl120 was added to the model and was sufficient to formulate a positive growth prediction for mannose.

3.11.5.2 Validation with proteomic and transcriptomic data

Gene expression. Flux balance analysis enables the prediction of other phenotypes such as flux states and gene essentiality. These predictions can be used along with experimental data to improve the model's quality [36]. Genome-wide expression [56] and transposon insertion (Tn-seq) [80] datasets available for *M. florum* were used as a reference for the validation of model predictions. Gene expression was compared to the model predicted flux states by converting both datasets to binary "on" or "off" values. The set of expressed genes was defined by finding the thresholds that would provide the best match between transcriptomic and proteomic data while maximizing the number of expressed genes (Figure S3.10A and B, Supplementary file 9). At the selected thresholds, 531 genes had a consistent signal in both proteomic and transcriptomic data while 145 had mixed signals (e.g. proteomic "on" and transcriptomic "off"). Only the genes for which datasets were consistent with each other were used for comparison with the model. This set contains 423 expressed and 108 silent genes (Figure S3.10C). Weiner and colleagues [81] previously reported that, using mRNA expression in *M. pneumoniae*, 564 of the 676 genes considered were expressed. While this number is higher than the 423 we report in *M. florum*, it is quite comparable to the 525 expressed genes out of

677 when considering only transcriptomic data (mean FPKM = 1221.0) at the selected threshold (FPKM = 168.0). Combining both datasets hence provides a more conservative observation.

The flux state through the metabolic network was obtained by optimizing the production of biomass using parsimonious flux balance analysis (pFBA), a version of FBA that allows the generation of a unique flux state prediction through minimization of enzyme usage [82]. This method is best suited for the comparison of predicted fluxes to gene expression [83]. A reaction flux was defined as active when the predicted value exceeded the numerical error (1e-8) and the flux was attributed to every gene that could catalyze the reaction via the gene-reaction rule. The comparison of binary flux predictions and observed expression was performed on the subset of model genes (173/208) for which proteomic and transcriptomic data showed no discrepancy.

Gene essentiality. Single gene essentiality was reported previously [80] where ~290 *M. florum* genes were proposed to be essential, which was considerably inferior to the 382/482 essential genes reported by Glass *et al.* [84] in *M. genitalium* and the 473 genes in the minimal cell JCVI-syn3.0 [10]. This number was revisited by including both growth data for mutants with transposon insertion (Supplementary file 9) as well as the relative position of the insertion within the gene (Figure S3.11). The latter assumes that an insertion in the final 20% section of the gene would not hamper its activity, a method that proved useful in the design of JCVI-syn3.0 [10]. The number of essential genes was therefore raised to 362 (Figure S3.11). The fact that the single gene essentiality is not equal to the number of genes in JCVI-syn3.0 indicates that there is still redundancy in *M. florum* that could be removed by genome reduction. Single-gene knockout growth simulations were performed on every gene in the model, hereby generating a prediction of essentiality that can be compared to the experimental transposon insertion dataset.

Comparing with model predictions. Comparing the model predictions with experimental data initially revealed erroneous predictions of the model that were manually addressed. True false negatives (TFN) were defined as genes simultaneously expressed and essential while no flux or essentiality was predicted. Eight TFN were identified. A single true false positives (TFP) was found, which had both flux and essentiality prediction but no observed expression nor essentiality. Curating these genes allowed increasing the model accuracy in the prediction of gene expression from 74.25% to 78.03% and essentiality from 74.52% to 76.92%. It is noteworthy that solving more than the TFP and TFN would mean fitting the model specifically to either essentiality of expression datasets.

The accuracy of the *i*JL208 model is lower than other mollicutes' models (*M. genitalium* [7] 79% initial and 87% after GrowMatch [85] and *M. pneumoniae* 86% [8]). Two factors may be responsible for that reduced accuracy. First, the number of genes included in the *M. florum* model is higher and some of these genes have a presumptive annotation. The choice of including more genes in the model was compensated by the attribution of a level of confidence for the added genes. Second, *M. florum* has not been as characterized from a biochemical standpoint than these other mollicutes. The development of this model provides a systematic approach to target the gaps in the current knowledge and should bolster the efforts towards characterization of *M. florum*'s molecular functions.

Curating True False Negatives (TFN). Two cases could be solved by the addition of a forced flux (DHAK, Mfl229 and dUPTase, Mfl257), two by the addition of new components to the biomass (RBFK, Mfl334 and ACPS, Mfl384) while the other 4 had missing information. It was the case for Mfl558, which is annotated as a chitin deacetylase. Chitin is likely to be absent from the growth medium under which the experimental data was generated, hence the reported expression and essentiality likely indicates that this enzyme was mis-annotated. The specificity of the inosine-5-monophosphate dehydrogenase (Mfl343) should also be addressed as it could not be explained in the current metabolic network. Finally, the necessity of the CTP synthase (Mfl648) suggested that the exchange of amino groups within *M. florum* occurred primarily

through exchanges in the glutamine/glutamate pool rather than through the import of ammonium from the medium but insufficient information was found in the literature to support that hypothesis.

Solving false negatives required the addition of specific constraint(s). The dUTPase (Mfl257) and dihydroxyacetone kinase (Mfl229) are simple examples of such cases (Figure S3.12). In the first case, the accidental production of dUTP by the cell was mimicked by forcing a flux through the PYK10 reaction which produces it from dUDP. In turn, this forces the activity of Mfl257 to produce dUMP and a pyrophosphate (Figure S3.12A). The second case represents a similar cellular situation where the highly reactive molecule dihydroxyacetone phosphate spontaneously loses its phosphate yielding dihydroxyacetone, a toxic molecule for the cell. A forced flux through this spontaneous reaction resolved the discrepancy, making Mfl229 carrying flux and being essential (Figure 3.6F).

A more complicated case is observed for the ribulose-5-phosphate epimerase (Mfl223). Activating it in the model requires forcing flux through the pentose phosphate pathway (PPP). In many mollicutes, this pathway is incomplete [5,7,8,16], often because no gene can be attributed to the transaldolase reaction (TALA). To address this apparent lack of annotation in *M. florum*, the reconstructed 3D structure for Mfl121 was compared to the structure the transaldolase B from *Francisella tularensis* (PDB identifier: 3IGX). The shared Beta barrel secondary structure between the two enzymes and the FATCAT p-value of 4.86e-5 suggests that Mfl121 might have been mis-annotated. Adding the TALA enables flux through the PPP but does not force it since active uptake of ribose circumvents its need. The non-essentiality/non-expression of both ribose kinase and ribose ABC transporter suggests that ribose was altogether absent from the ATCC1161 growth medium in which the datasets were generated. Ribose was therefore removed from the *in silico* medium, resulting in flux through the PPP and increased prediction accuracy for expression.

Adding the TALA reaction enables flux through the PPP but does not force it since active uptake of ribose circumvents its need. The non-essentiality/non-expression of both ribose kinase (Mfl642, *rbsK*) and ribose ABC transporter (Mfl666, Mfl667, Mfl668, Mfl669) suggests that ribose was altogether absent from the ATCC 1161 growth medium in which the datasets were generated. Ribose was therefore removed from the *in silico* medium, resulting in flux through the PPP and increased prediction accuracy for expression.

Curating True False Positive (TFP). The path for synthesis of nicotinamide dinucleotide in *M. florum* was discussed above. The presence of a nicotinamidase (Mfl340) suggested the import of nicotinamide from the media. Nevertheless, experimental data revealed that this enzyme is both non-essential and non-expressed, suggesting that the downstream metabolite, nicotinate, may be imported instead. Adding this metabolite to the *in silico* media as well as an import reaction avoids the need for Mfl340, recapitulating experimental observations (Figure S3.12C).

3.11.6 Model-driven prediction of a minimal genome

3.11.6.1 Varying the growth rate results in different genome reduction scenarios

Together with experimental gene essentiality and the transcription unit architecture, *i*JL208 was used to formulate a minimal genome prediction. Using the MinGenome algorithm [86], genome reduction scenarios were generated at different growth rates (Figure S3.13). This was made possible because minGenome attempts to find the largest possible deletion in the genome without breaking the established constraints. The constraints imposed include the impossibility to delete an essential or its associated promoter, and ensuring the feasibility of the genome-scale model. The model's objective and value are therefore fixed. If a gene deletion prevents the model from solving at this specific growth rate, then the deletion is not possible.

Varying the minimum growth rate imposed as a constraint on the optimization problem formulated with minGenome enables the deletion of genes that could hamper the growth rate without being completely lethal. While an array of growth rates were tested, only three different genome reduction scenarios were obtained. The similarity of the resulting genomes to JCVI-syn3.0 were assessed for each growth rate constraint imposed. The genomes formulated with a lower minimum growth rate constraint were more similar to JCVI-syn3.0 (Figure S3.13). The final genome size of the lower growth rate constraint scenarios were also smaller than higher ones.

In all cases, no more genes could be deleted after 75 deletions. A minimal size was reached at 60% of the optimal growth rate, yielding a 552kbp genome-size containing, a 30% reduction from the initial *M. florum* L1 genome (Tableau S3.1). In size, this minimal genome scenario lies between the JCVI-syn3.0 inspired (470kbp, 409 genes) and the core genome of *M. florum* (644kbp, 585 genes) suggested by Baby *et al.* [80]. While this model-driven prediction does not reduce the number of genes beyond that of JCVI-syn3.0, a reduction of 30% in genome-size is a level that has been reached experimentally in different species. In fact, the *Bacillus subtilis* genome was trimmed by 36% and yielded a functional genome [87] with growth rates similar to the wild-type strain. To our knowledge, the smallest *Escherichia coli* genome allowing robust growth is that of DGF-298 [88]. At 2.98 Mbp, this genome represents a 34.4% reduction compared to the original K-12 substr. W3110 [89]. *B. subtilis* and *E. coli* reduced genome strains yielded robust cells with growth rates similar to their parental strain. This was not the case for JCVI-syn3.0, which was originally reported to have a lower growth rate and an altered morphology [10]. In order to produce a more robust and functional cell usable in the laboratory, some 12 genes needed to be added back into the original JCVI-syn3.0, a 485 genes bacterium called JCVI-syn3.0A [16]. Our current prediction of a minimal gene set is 63 genes above JCVI-syn3.0A.

3.11.6.2 Functional analysis of the reduced genome

The functional categories in which the deleted/conserved proteins were analysed using the KEGG ontology (Figure S3.14). Interestingly, the largest portion of the candidates *loci* for deletion belonged to the unmapped category. Reducing the number of unknown components is a key argument justifying research efforts in minimal cells. Identifying these 80 non-essential hypothetical proteins was therefore crucial for further experimental efforts to reduce the genome of *M. florum*.

Next, we compared the number of genes in each KEGG functional category identified as deletion targets to the genes kept in the reduced genome scenario. Interestingly, the main category affected by deletions were uncharacterized proteins ("Not mapped") with 81 proteins (~56% of all deleted proteins), and a small fraction of those (16) had homologs in JCVI-syn3.0 (Figure S3.14A). With 191 proteins out of 535, the proportion of uncharacterized proteins retained in the *M. florum* reduced genome scenario (~36%) is also very similar to the reported proportion in JCVI-syn3.0 (149/438, 34%).

The second KEGG category with the highest number of deletions was "Metabolism", with 34 proteins removed (Figure S3.14A). 155 proteins of this category remained in the reduced genome and 54 of those were not homologous to JCVI-syn3.0 (Figure S3.14B). Specifically, proteins deleted in *M. florum* but present in JCVI-syn3.0 were mostly found in the transport sub-category (Mfl019, Mfl234, Mfl533, Mfl534) and are annotated as ABC transporters. In accordance with our experimental data and *i*JL208, both the glutamine ABC transporter (Mfl019) and the phosphate ABC transporter (Mfl234) were not essential in *M. florum* since other routes exist for their import. Mfl533 and Mfl534 were annotated as lipid A export proteins (*msbA*) but no evidence supports the presence of this metabolite in *M. florum*. Since the lipid module was amongst the least well characterized and given its status in our prediction, these genes represent top priorities for further biochemical characterization. Comparing the deleted proteins with the remaining ones in *i*JL208 revealed that gene redundancy in the

sucrose phosphotransferase system (PTS) importer allowed to keep this function in the reduced genome. Contrarily, the trehalose PTS system was completely removed suggesting that alternate versions of a minimal gene set for *M. florum* may have different auxotrophies.

Finally, the "Genetic information processing" (GIP) category was the least affected by deletions (Figure S3.14A) and contained the highest number of proteins in the reduced genome scenario (Figure S3.14B). This category also contains the highest proportion of proteins shared with JCVI-syn3.0 (~89%).

We provide here a detailed analysis of the composition of the reduced genome by functional category, detailed information is available in Supplementary file 10:

Not mapped: 191 not mapped proteins conserved in the reduced genome with 106 proteins specific to *M. florum* (absent from either JCVI-syn1.0 and JCVI-syn3.0).

Metabolism: The metabolism category is composed of 12 sub-categories, three of which exclusively contain genes that have homologs in JCVI-syn3.0: ATP synthase, amino acid metabolism, and secretion system. To evaluate the possibility for further reduction or potential alternative genome reduction scenarios, we detail the composition of the remaining nine metabolism sub-categories.

- **Transport:** This sub-category contains 33 proteins with eight *M. florum* specific proteins, 18 homologs to JCVI-syn3.0, and seven JCVI-syn1.0 homologs. The main features shared with JCVI-syn3.0 in this category are the following: oligopeptide ABC transporter, cobalt transporter, K+/NA+ uptake protein, phosphate ABC transporter, spermidine/putrescine ABC transporter, phosphonate ABC transporter (half conserved) and ribose ABC transporter.

The transport proteins that were not conserved in JCVI-syn3.0 are the following: Amino acid transporter permease (Mfl184), Mg/Co/Ni transporter, formate/nitrate transporter, Mg2+ transport, and xanthine/uracil permease.

- **PTS system:** This sub-category contains 5 proteins, with one specific to *M. florum*, three having homologs in JCVI-syn3.0, and one homolog to JCVI-syn1.0. The conserved functions in JCVI-syn3.0 are phosphoenolpyruvate-protein phosphotransferase, phosphotransferase system phosphohistidine, and the glucose specific PTS transporter subunit.

  The PTS system related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: sucrose specific PTS component, and fructose specific PTS component.

- **Glycolysis and carbohydrate metabolism:** This sub-category contains 27 proteins, with eight specific to *M. florum*, 15 having homologs in JCVI-syn3.0, and four homologs in JCVI-syn1.0. The conserved functions in JCVI-syn3.0 are: pyruvate dehydrogenase, phosphomannomutase, pyruvate kinase, as well as all glycolysis enzymes and secretion routes for both lactate and acetate.

  Glycolysis and carbohydrate metabolism related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: all beta-glucosidases, both E1 subunits of the pyruvate dehydrogenase, a sucrose-6-phosphate hydrolase, a fructokinase and a 1-phosphofructokinase.

- **Pentose phosphate metabolism:** This sub-category contains 8 proteins, with three specific to *M. florum*, five having homologs in JCVI-syn3.0, and no homologs in JCVI-syn1.0. The conserved functions in JCVI-syn3.0 are: phosphoribosylpyrophosphate synthetase, ribose-5-phosphate isomerase, transketolase, one of the two 2-deoxyribose-

5-phosphate aldolase and one of the two pentose-5-phosphate epimerase.

Pentose phosphate metabolism related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: one of the two 2-deoxyribose-5-phosphate aldolase, one of the two pentose-5-phosphate epimerase, as well as the ribokinase.

- **Cofactor biosynthesis:** This sub-category contains 16 proteins, with two specific to *M. florum*, seven having homologs in JCVI-syn3.0, and seven homologs in JCVI-syn1.0. The conserved functions in JCVI-syn3.0 are: cytosol aminopeptidase (duplication or complex), riboflavin kinase (duplication), nicotinate-nucleotide adenylyltransferase, nicotinate phosphoribosyltransferase, NAD kinase, holo-ACP synthase, folyl-polyglutamtate synthetase.

  Cofactor biosynthesis related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: lipoate protein ligase, dephospho-CoA kinase, nicotinamidase/pyrazinamidase.

- **Purine and pyrimidine metabolism:** This sub-category contains 20 proteins, with two specific to *M. florum*, 14 having homologs in JCVI-syn3.0, and four homologs in JCVI-syn1.0. The conserved functions in JCVI-syn3.0 are: thioredoxin reductase, phosphoribosyl transferases for uracil, adenine, and hypoxanthine-guanine, as well as kinases for adenylate, guanylate, cytidylate, thymidylate, and finally a purine-nucleoside phosphorylase.

  Purine and pyrimidine metabolism related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: adenylosuccinate lyase, thymidine phosphorylase, cytidine deaminase, guanosine-5-monophosphate oxidoreductase, Inositol-5-monophosphate dehydrogenase, xanthosine triphosphate pyrophosphatase.

- **Lipid metabolism:** This sub-category contains eight proteins, with none specific to *M. florum*, seven having homologs in JCVI-syn3.0, and one homolog to JCVI-syn1.0. The conserved functions in JCVI-syn3.0 are: glycerol-3-phosphate acetyltransferase, CDP-diglyceride synthetase, 1-acyl-sn-glycerol-3-phosphate acyltransferase, cardiolipin synthase, acyl carrier protein, phosphatidylglycerophosphate synthase.

  The only Lipid metabolism related proteins conserved in the reduced genome but not shared with JCVI-syn3.0 is the NAD-dependent-glycerol-3-phosphate dehydrogenase.

Genetic Information Processing: The Genetic Information Processing general category contains 13 sub-categories, five of which contain exclusively genes that have homologs in JCVI-syn3.0: RNA polymerase, Translation factors, Sulfur relay system, Protein export, and tRNA loading and maturation. To evaluate the possibility for further reduction or potential alternative genome reduction, we detail the composition of the remaining eight Genetic Information Processing sub-categories. Since most of the proteins within this general category have homologs in JCVI-syn3.0, we only provide the detail of the discrepancies between the reduced *M. florum* genome and JCVI-syn3.0.

- **DNA repair:** This sub-category contains 18 proteins, with one specific to *M. florum*, 11 having homologs in JCVI-syn3.0, and six homologs in JCVI-syn1.0. DNA repair related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: Recombination proteins (U and A, Mfl267 DNA repair/recombination protein is conserved), Uracil-DNA glycosylase, DNA-3-methyladenine glycosidase, Exodeoxyribonuclease V (but VII is conserved), DNA polymerase IV, and formamidopyrimidine-DNA glycosylase.

- **DNA replication and partition:** This sub-category contains 26 proteins, with one specific to *M. florum*, 23 having homologs in JCVI-syn3.0, and two homologs in JCVI-

syn1.0. The three proteins DNA replication and partition related proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: DNA cytosine methyltransferase (Mfl308), SepF/FtsZ-interacting protein related to cell division (Mfl391), Ribonuclease HII (EC 3.1.26.4) (Mfl537).

- **Transcription factors:** This sub-category contains five proteins, with none specific to *M. florum*, three having homologs in JCVI-syn3.0, and two homologs in JCVI-syn1.0. The two transcription factors that were conserved in the reduced genome but not shared with JCVI-syn3.0 are: an unknown transcriptional regulator and a transcriptional repressor of the fructose operon, DeoR family (consistent with metabolism).

- **Ribosome:** The Ribosome contains 50 proteins, with a single one specific to *M. florum*, and the 49 others having homologs in JCVI-syn3.0. The single ribosomal protein that is conserved in the reduced genome but not shared with JCVI-syn3.0 is the 50S ribosomal protein L33. Interestingly, this protein was also absent from JCVI-syn1.0 but was added when generating JCVI-syn3.0A.

- **Ribosome biogenesis:** This sub-category contains 26 proteins, with one specific to M. florum, 23 having homologs in JCVI-syn3.0, and two homologs in JCVI-syn1.0. The three proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are all methyltransferases: 16S rRNA (cytosine(967)-C(5))-methyltransferase, 16S rRNA (uracil(1498)-N(3))-methyltransferase (EC 2.1.1.193), and RNA binding methyltransferase FtsJ like.

- **Nucleases:** This sub-category contains eight proteins, with none specific to M. florum, seven having homologs in JCVI-syn3.0, and one homolog in JCVI-syn1.0. The single protein that was conserved in the reduced genome but not shared with JCVI-syn3.0 is the Mg2+ dependent DNAse.

- **Chaperones:** This sub-category contains five proteins, with none specific to M. florum, three having homologs in JCVI-syn3.0, and two homologs in JCVI-syn1.0. The two proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are the hsp33 redox-regulated chaperone and the cell division trigger factor (EC 5.2.1.8).

- **Peptidases:** This sub-category contains five proteins, with two specific to M. florum, three having homologs in JCVI-syn3.0, and no homologs in JCVI-syn1.0. The two proteins that were conserved in the reduced genome but not shared with JCVI-syn3.0 are the intramembrane protease RasP/YluC, implicated in cell division based on FtsL cleavage and the GMP synthase [glutamine-hydrolyzing], amidotransferase subunit (EC 6.3.5.2) / GMP synthase [glutamine-hydrolyzing], ATP pyrophosphatase subunit (EC 6.3.5.2).

3.11.7 Supplementary references

1. Mc COY, R. E. et al. *Acholeplasma florum*, a New Species Isolated from Plants?
2. Morowitz, H. J. Special guest lecture: The completeness of molecular biology. Isr. J. Med. Sci. 2, (1984).
3. Fraser, C. M. et al. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–403 (1995).
4. Lazarev, V. N. et al. Complete genome and proteome of *Acholeplasma laidlawii*. J. Bacteriol. 193, 4943–4953 (2011).
5. Miles, R. J. Catabolism in mollicutes. J. Gen. Microbiol. 138, 1773–1783 (1992).
6. Dennis Pollack, J. 11. Carbohydrate Metabolism and Energy Conservation.
7. Suthers, P. F. et al. A Genome-Scale Metabolic Reconstruction of *Mycoplasma genitalium*, iPS189. (2009) doi:10.1371/journal.pcbi.1000285.
8. Wodke, J. A. H. et al. Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol. Syst. Biol. 9, 653 (2013).

9. Bautista, E. J. et al. Semi-automated Curation of Metabolic Models via Flux Balance Analysis: A Case Study with *Mycoplasma gallisepticum*. (2013) doi:10.1371/journal.pcbi.1003208.
10. Hutchison, C. A., 3rd et al. Design and synthesis of a minimal bacterial genome. Science 351, aad6253 (2016).
11. Gibson, D. G. et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329, 52–56 (2010).
12. Peterson, S. N. & Fraser, C. M. The complexity of simplicity. Genome Biol. 2, COMMENT2002 (2001).
13. Quinlan, D. C. & Maniloff, J. Deoxyribonucleic acid synthesis in synchronously growing *Mycoplasma gallisepticum*. J. Bacteriol. 115, 117–120 (1973).
14. Windsor, H. M., Windsor, G. D. & Noordergraaf, J. H. The growth and long term survival of *Acholeplasma laidlawii* in media products used in biopharmaceutical manufacturing. Biologicals 38, 204–210 (2010).
15. Wattam, A. R. et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 45, D535–D542 (2017).
16. Breuer, M. et al. Essential metabolism for a minimal cell. Elife 8, (2019).
17. Barré, A., de Daruvar, A. & Blanchard, A. MolliGen, a database dedicated to the comparative genomics of Mollicutes. Nucleic Acids Res. 32, D307–10 (2004).
18. King, Z. A. et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 44, D515–22 (2016).
19. Moretti, S. et al. MetaNetX/MNXref--reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 44, D523–6 (2016).
20. Schilling, C. H., Thakar, R., Travnik, E., Van Dien, S. & Wiback, S. SimPheny^{TM}: A Computational Infrastructure for Systems Biology.
21. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst. Biol. 7, 74 (2013).

22. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
23. Mih, N. et al. ssbio: a Python framework for structural systems biology. Bioinformatics 34, 2155–2157 (2018).
24. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. Nat. Methods 12, 7–8 (2015).
25. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5, 725–738 (2010).
26. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res. 45, W291–W299 (2017).
27. International Union of Biochemistry and Molecular Biology. Nomenclature Committee & Webb, E. C. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. (Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992).
28. Yang, Z. & Tsui, S. K.-W. Functional Annotation of Proteins Encoded by the Minimal Bacterial Genome Based on Secondary Structure Element Alignment. J. Proteome Res. 17, 2511–2520 (2018).
29. Yang, Z., Zeng, X. & Tsui, S. K.-W. Investigating function roles of hypothetical proteins encoded by the *Mycobacterium tuberculosis* H37Rv genome. BMC Genomics 20, 394 (2019).
30. Antczak, M., Michaelis, M. & Wass, M. N. Environmental conditions shape the nature of a minimal bacterial genome. Nat. Commun. 10, 3100 (2019).
31. Billings, W. M., Hedelius, B., Millecam, T., Wingate, D. & Della Corte, D. ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. doi:10.1101/830273.
32. AlQuraishi, M. AlphaFold at CASP13. Bioinformatics 35, 4862–4865 (2019).
33. Senior, A. W. et al. Protein structure prediction using multiple deep neural networks in CASP13. Proteins: Struct. Funct. Bioinf. (2019).

34. Ghatak, S., King, Z. A., Sastry, A. & Palsson, B. O. The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. Nucleic Acids Res. 47, 2446–2454 (2019).

35. Glass, J. I., Merryman, C., Wise, K. S., Hutchison, C. A. & Smith, H. O. Minimal Cells—Real and Imagined. Cold Spring Harb. Perspect. Biol. (2017) doi:10.1101/cshperspect.a023861.

36. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. 5, 93–121 (2010).

37. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361 (2017).

38. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–62 (2016).

39. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000).

40. Placzek, S. et al. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res. 45, D380–D388 (2017).

41. Artimo, P. et al. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 40, W597–603 (2012).

42. Acevedo-Rocha, C. G., Fang, G., Schmidt, M., Ussery, D. W. & Danchin, A. From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet. 29, 273–279 (2013).

43. Danchin, A. & Fang, G. Unknown unknowns: essential genes in quest for function. Microb. Biotechnol. 9, 530–540 (2016).

44. Dennis Pollack, J., Myers, M. A., Dandekar, T. & Herrmann, R. Suspected Utility of Enzymes with Multiple Activities in the Small Genome Mycoplasma Species: The Replacement of the Missing ‘ Household’ Nucleoside Diphosphate Kinase Gene and

Activity by Glycolytic Kinases. OMICS A Journal of Integrative Biology 6, 247–258 (2002).

45. Vinitsky, A. & Grubmeyer, C. A new paradigm for biochemical energy coupling. *Salmonella typhimurium* nicotinate phosphoribosyltransferase. J. Biol. Chem. 268, 26004–26010 (1993).
46. Dennis Pollack, J. The necessity of combining genomic and enzymatic data to infer metabolic function and pathways in the smallest bacteria: amino acid, purine and pyrimidine metabolism in mollicutes. Frontiers in Bioscience 7, 1762–1781 (2002).
47. Lu, F. et al. Structure and mechanism of the uracil transporter *uraA*. Nature 472, 243–246 (2011).
48. Seelig, B. Multifunctional enzymes from reduced genomes - model proteins for simple primordial metabolism? Mol. Microbiol. 105, 505–507 (2017).
49. Bizarro, C. V. & Schuck, D. C. Purine and pyrimidine nucleotide metabolism in Mollicutes. Genet. Mol. Biol. 30, 190–201 (2007).
50. Pollack, J. D., Williams, M. V. & McElhaney, R. N. The comparative metabolism of the mollicutes (Mycoplasmas): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. Crit. Rev. Microbiol. 23, 269–354 (1997).
51. Ben-Menachem, G., Himmelreich, R., Herrmann, R., Aharonowitz, Y. & Rottem, S. The thioredoxin reductase system of mycoplasmas. Microbiology 143 (Pt 6), 1933–1940 (1997).
52. Yus, E. et al. Impact of genome reduction on bacterial metabolism and its regulation. Science 326, 1263–1268 (2009).
53. Hosie, A. H. & Poole, P. S. Bacterial ABC transporters of amino acids. Res. Microbiol. 152, 259–270 (2001).
54. Pollack, J. D., Tryon, V. V. & Beaman, K. D. The metabolic pathways of Acholeplasma and Mycoplasma: an overview. Yale J. Biol. Med. 56, 709–716 (1983).
55. Béven, L. et al. Specific evolution of F1-like ATPases in mycoplasmas. PLoS One 7, e38793 (2012).

56. Matteau D, Lachance J-C, Grenier F, Gauthier S, Daubenspeck JM, Dybvig K, et al. Integrative characterization of the near-minimal bacterium *Mesoplasma florum*. Mol Syst Biol. 2020;16: e9844.
57. Heath, R. J., Jackowski, S. & Rock, C. O. Chapter 3 Fatty acid and phospholipid metabolism in prokaryotes. New Compr. Biochem. 36, 55–92 (2002).
58. Byers, D. M. & Gong, H. Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family. Biochem. Cell Biol. 85, 649–662 (2007).
59. Schlame, M. Cardiolipin synthesis for the assembly of bacterial and mitochondrial membranes. J. Lipid Res. 49, 1607–1620 (2008).
60. Xavier, J. C., Patil, K. R. & Rocha, I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. Metab. Eng. 39, 200–208 (2017).
61. Oshida, K. et al. Effects of Dietary Sphingomyelin on Central Nervous System Myelination in Developing Rats. Pediatric Research vol. 53 589–593 (2003).
62. Salman, M. & Rottem, S. The cell membrane of *Mycoplasma penetrans*: lipid composition and phospholipase A1 activity. Biochim. Biophys. Acta 1235, 369–377 (1995).
63. Saito, Y., Silvius, J. R. & McElhaney, R. N. Membrane lipid biosynthesis in *Acholeplasma laidlawii* b: elongation of medium- and long-chain exogenous fatty acids in growing cells. J. Bacteriol. 133, 66–74 (1978).
64. Hackett, K. J., Ginsberg, A. S., Rottem, S., Henegar, R. B. & Whitcomb, R. F. A defined medium for a fastidious Spiroplasma. Science 237, 525–527 (1987).
65. Browning, G. & Citti, C. Mollicutes: Molecular Biology and Pathogenesis. (Horizon Scientific Press, 2014).
66. Neyrolles, O. et al. Identification of two glycosylated components of *Mycoplasma penetrans*: a surface-exposed capsular polysaccharide and a glycolipid fraction. Microbiology 144 (Pt 5), 1247–1255 (1998).
67. Bertin, C. et al. Characterization of free exopolysaccharides secreted by *Mycoplasma mycoides subsp. mycoides.* PLoS One 8, e68373 (2013).

68. Hong, Y. & Reeves, P. R. Diversity of o-antigen repeat unit structures can account for the substantial sequence variation of wzx translocases. J. Bacteriol. 196, 1713–1722 (2014).
69. Keçeli, S. A. & Miles, R. J. Differential inhibition of mollicute growth: an approach to development of selective media for specific mollicutes. Appl. Environ. Microbiol. 68, 5012–5016 (2002).
70. Norsigian, C. J. et al. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Res. (2019) doi:10.1093/nar/gkz1054.
71. Moriarity, J. L. et al. UDP-glucuronate decarboxylase, a key enzyme in proteoglycan synthesis: cloning, characterization, and localization. J. Biol. Chem. 277, 16968–16975 (2002).
72. Salaemae, W., Booker, G. W. & Polyak, S. W. The Role of Biotin in Bacterial Physiology and Virulence: a Novel Antibiotic Target for *Mycobacterium tuberculosis*. Microbiol Spectr 4, (2016).
73. Lachance, J.-C. et al. BOFdat: generating biomass objective function stoichiometric coefficients from experimental data. bioRxiv 243881 (2018) doi:10.1101/243881.
74. Mohamed, E. T. et al. Generation of an E. coli platform strain for improved sucrose utilization using adaptive laboratory evolution. Microb. Cell Fact. 18, 116 (2019).
75. Li, Z., Jaroszewski, L., Iyer, M., Sedova, M. & Godzik, A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. Nucleic Acids Research vol. 48 W60–W64 (2020).
76. Auiewiriyanukul, W., Saburi, W., Kato, K., Yao, M. & Mori, H. Function and structure of GH13_31 α-glucosidase with high α-(1→4)-glucosidic linkage specificity and transglucosylation activity. FEBS Letters vol. 592 2268–2281 (2018).
77. Swan, M. K., Hansen, T., Schönheit, P. & Davies, C. A Novel Phosphoglucose Isomerase (PGI)/Phosphomannose Isomerase from the Crenarchaeon *Pyrobaculum aerophilum* Is a Member of the PGI Superfamily Structural evidence at 1.16-Å resolution. J. Biol. Chem. 279, 39838–39845 (2004).

78. Bertin, C. et al. Highly dynamic genomic loci drive the synthesis of two types of capsular or secreted polysaccharides within the *Mycoplasma mycoides* cluster. Appl. Environ. Microbiol. 81, 676–687 (2015).
79. Regni, C., Tipton, P. A. & Beamer, L. J. Crystal structure of PMM/PGM: an enzyme in the biosynthetic pathway of *P. aeruginosa* virulence factors. Structure 10, 269–279 (2002).
80. Baby, V. et al. Inferring the Minimal Genome of *Mesoplasma florum* by Comparative Genomics and Transposon Mutagenesis. mSystems 3, (2018).
81. Weiner, J., 3rd, Zimmerman, C.-U., Göhlmann, H. W. H. & Herrmann, R. Transcription profiles of the bacterium Mycoplasma pneumoniae grown at different temperatures. Nucleic Acids Res. 31, 6306–6320 (2003).
82. Lewis, N. E. et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. 6, 390 (2010).
83. Machado, D. & Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput. Biol. 10, e1003580 (2014).
84. Glass, J. I. et al. Essential genes of a minimal bacterium. Proc. Natl. Acad. Sci. U. S. A. 103, 425–430 (2006).
85. Kumar, V. S. & Maranas, C. D. *GrowMatch: an automated method for reconciling in silico/in vivo* growth predictions. PLoS Comput. Biol. 5, e1000308 (2009).
86. Wang, L. & Maranas, C. D. MinGenome: An *In Silico* Top-Down Approach for the Synthesis of Minimized Genomes. ACS Synth. Biol. 7, 462–473 (2018).
87. Reuß, D. R. et al. Large-scale reduction of the *Bacillus subtilis* genome: consequences for the transcriptional network, resource allocation, and metabolism. Genome Res. 27, 289–299 (2017).
88. Hirokawa, Y. et al. Genetic manipulations restored the growth fitness of reduced-genome *Escherichia coli*. J. Biosci. Bioeng. 116, 52–58 (2013).

89. Westphal, L. L., Sauvey, P., Champion, M. M., Ehrenreich, I. M. & Finkel, S. E. Genome wide Dam Methylation in *Escherichia coli* during Long-Term Stationary Phase. mSystems 1, (2016).

3.11 SUPPLEMENTARY FIGURES

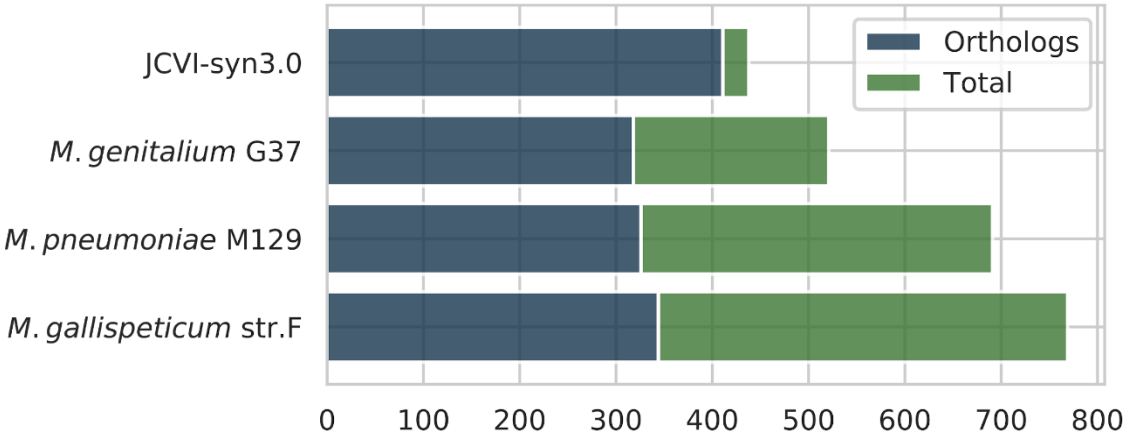

Figure S3.1 Orthologous proteins in other mollicute species with an existing metabolic model. Orthologous proteins were identified using the PATRIC proteome comparison tool. The total number of genes is presented for each species, and the number of orthologs is indicated in blue. The synthetic bacterium JCVI-syn3.0 has both the smallest protein count and the highest number of orthologs with *M. florum* L1.

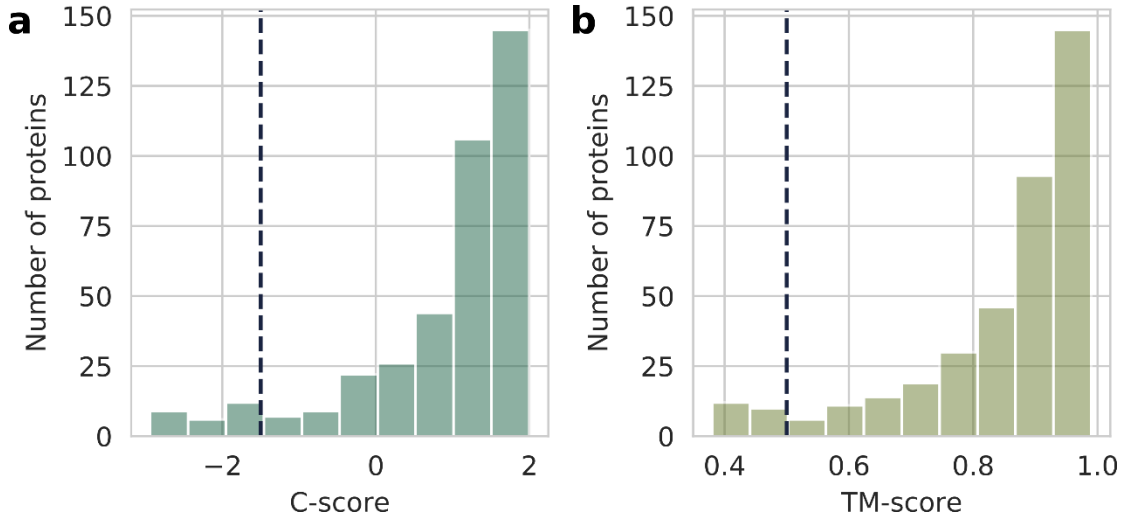

Figure S3.2 Distribution of the scores from the 3D protein reconstructions obtained with I-TASSER. Based on the documentation, a model structure is considered reliable if its C-score (**a**) andTM-score (**b**) are above 1.5 and 0.5, respectively. The majority of the modelled proteins (361/386) had scores higher than these thresholds (dark blue dotted lines).

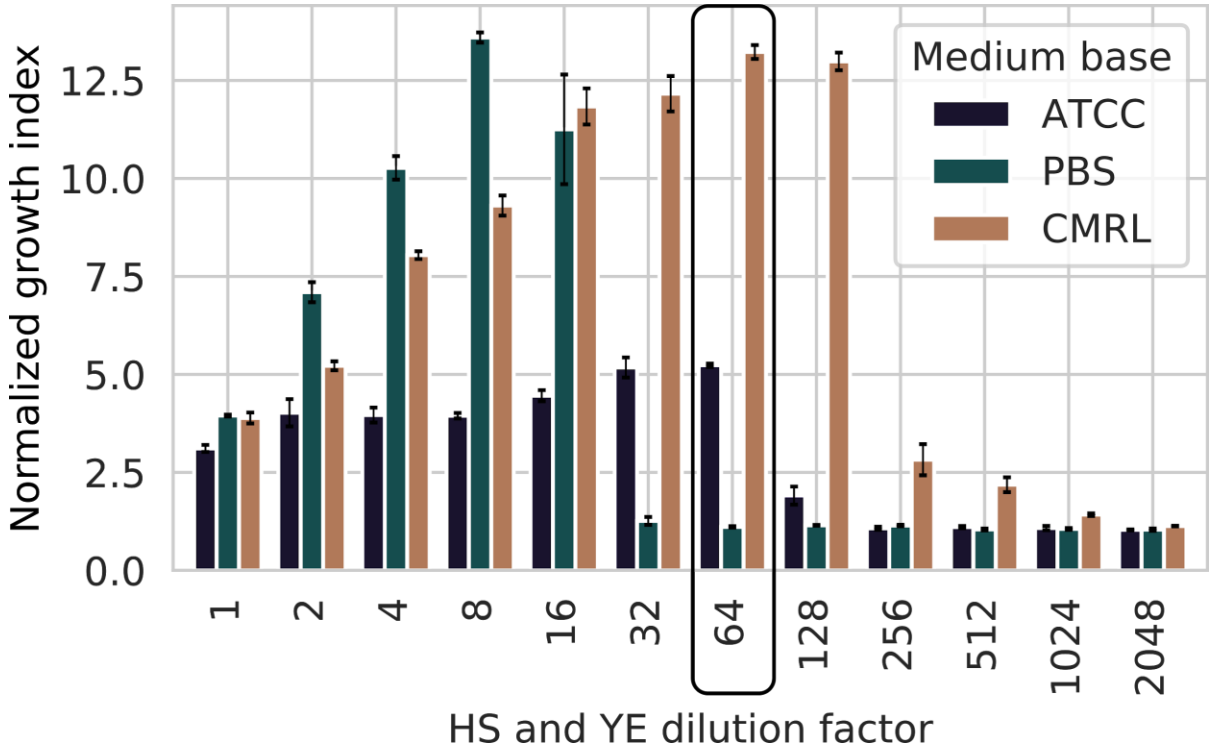

Figure S3.3 Medium simplification to maximize the difference in apparent growth between sugar supplemented and non-supplemented media. The normalized growth index (color fold change after 24 hours over a no sugar control) determined in different medium bases (ATCC 1161, CMRL 1066, and PBS 1X) supplemented with decreasing concentrations

of horse serum (HS) and yeast extract (YE). The black box shows maximum difference in color fold change for the highest depletion in HS and YE (0.313% HS and 0.02% YE) when using a CMRL 1066 medium base (CSY medium), corresponding to a 64 fold dilution factor over the concentrations found in ATCC 1161 (20% HS/1.35% YE). Bars and error bars indicate the mean and standard deviation calculated from technical triplicate.

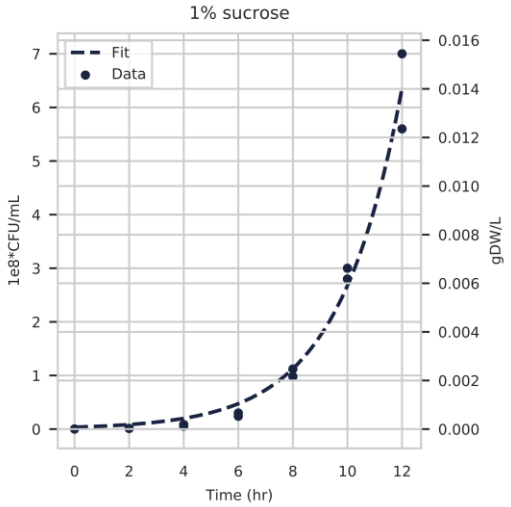

Figure S3.4 Biomass concentration over time of a *M. florum* culture growing in CSY medium with 1% sucrose. Biomass was measured using colony forming units (CFU/ml; left axis) and converted to grams of dry weight (gDW/L; right axis). A simple exponential growth fit is shown.

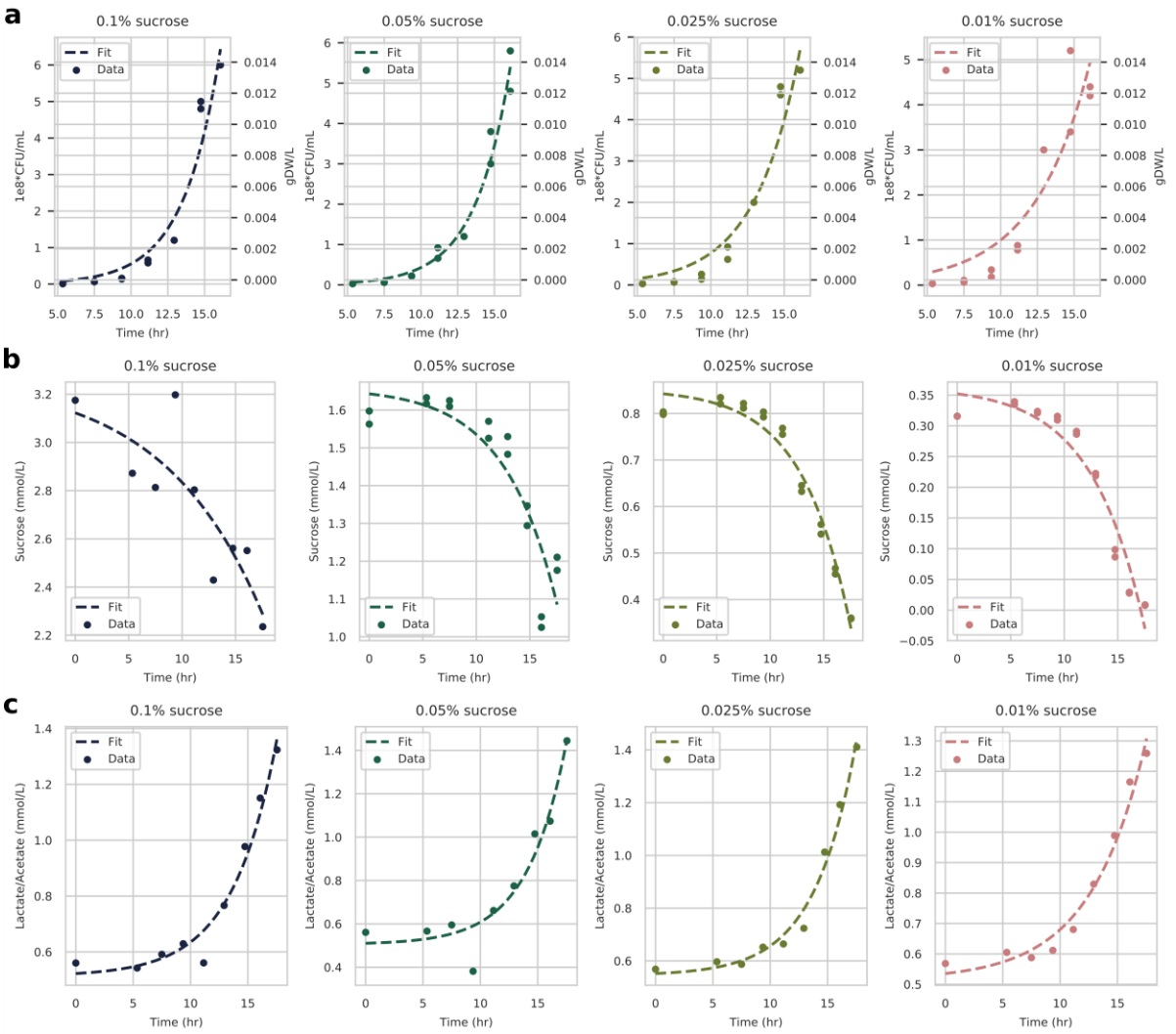

Figure S3.5 Raw data used to infer growth rates, sucrose uptake and lactate/acetate secretion rates. a Biomass concentration over time of *M. florum* cultures growing in CSY medium with varying initial concentration of sucrose. Biomass was measured using colony forming units (CFU/ml; left axis), and converted to grams of dry weight (gDW/L; right axis). For each graph, a simple exponential growth fit is shown (dotted line). **b** Same as panel a, but for sucrose and **c** for lactate/acetate (indiscernible peaks) concentrations measured by high performance liquid chromatography.

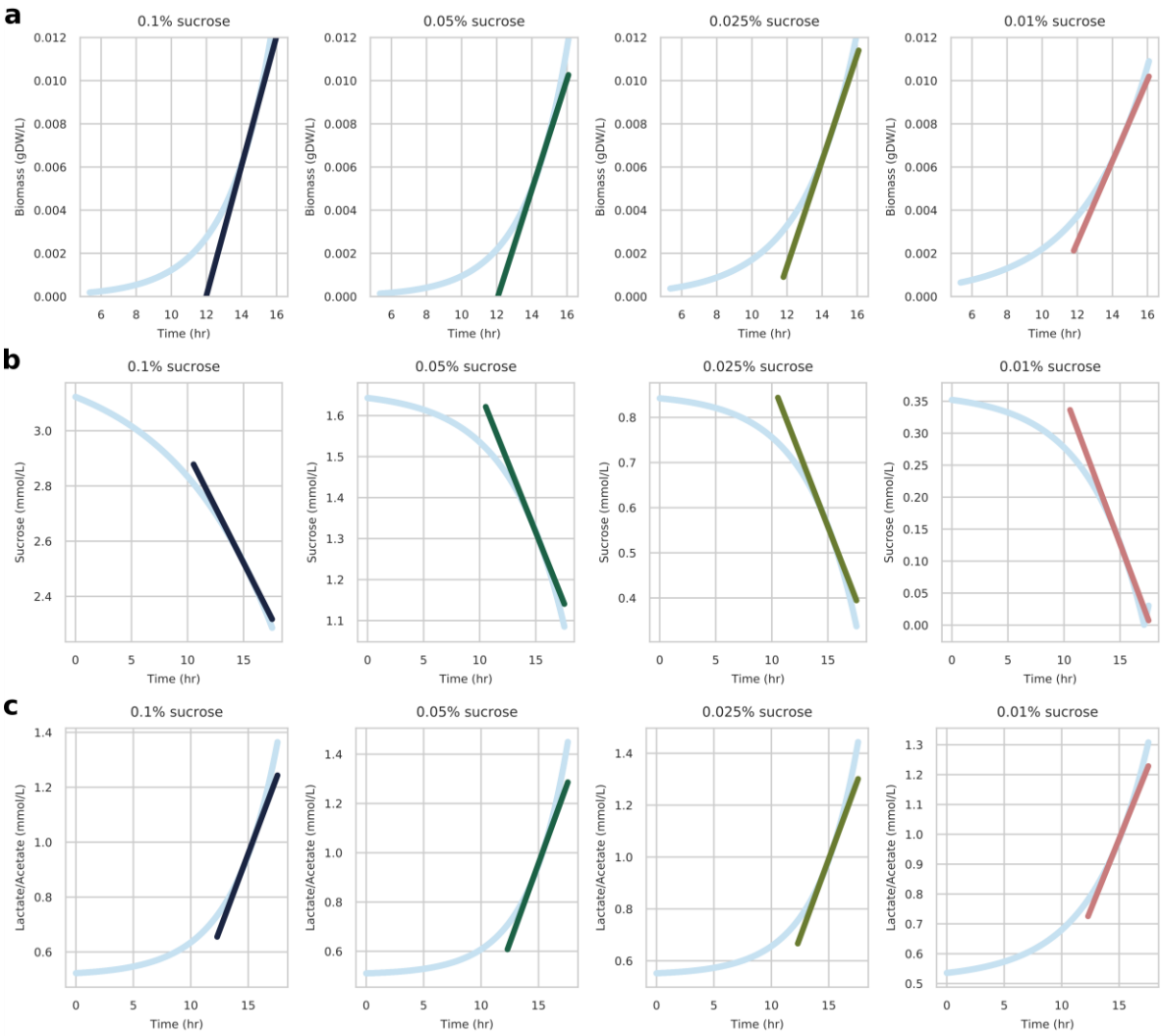

Figure S3.6 Linear regression in *M. florum* exponential growth phase (14 to 16 hours) for **a** biomass, **b** sucrose, and **c** combined acetate/lactate concentrations. The slopes associated with these linear regressions were used to calculate the substrate (sucrose) and product (lactate/acetate mixture) specific rates (see Material and methods). The linear regression plot was extrapolated to facilitate visualization.

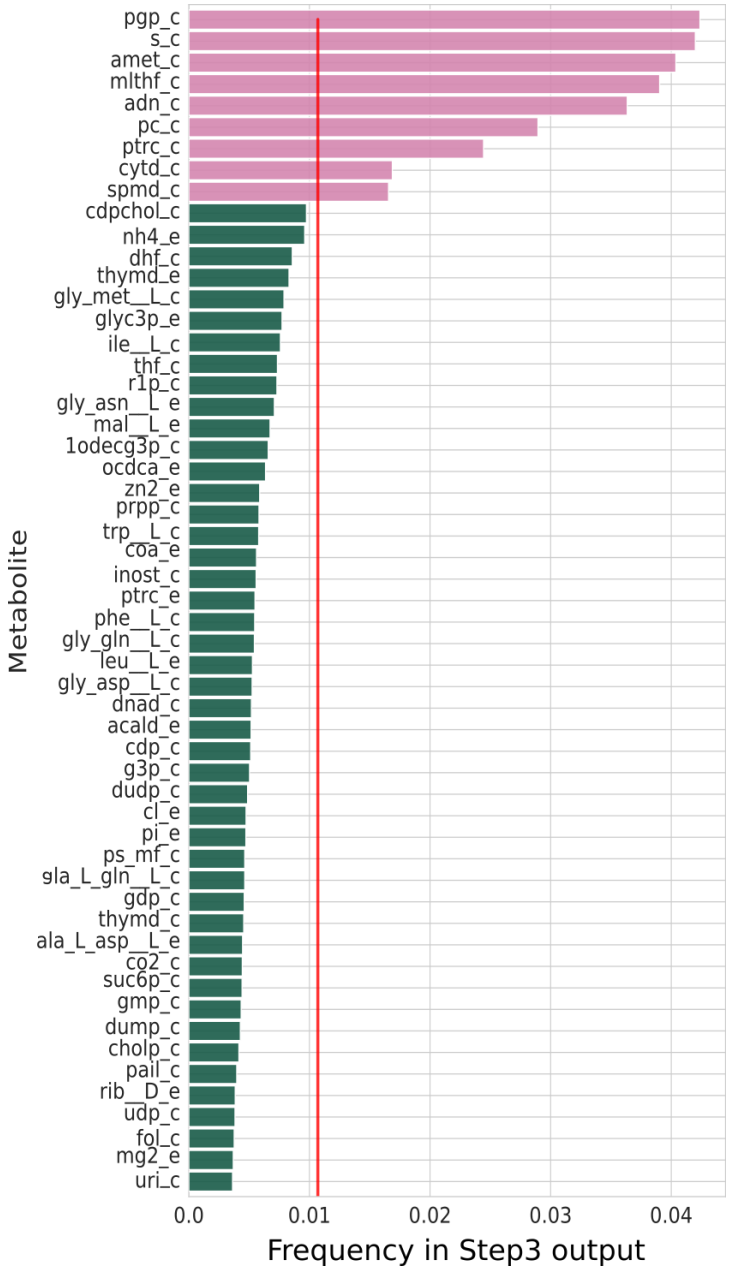

Figure S3.7 Metabolite apparition frequencies from the genetic algorithm output of BOFdat Step3. The nine metabolites above average (red line) shown in pink were included in the BOF (pgp: phosphatidylglycerol phosphate, s: sulfur, amet: adenosyl methionine, mlthf: 5,10-Methylenetetrahydrofolate, adn: adenosine, pc: phosphatidylcholine, ptrc: putrescine, cytd: cytidine, spmd: spermidine). _c and _e suffixes indicate cytoplasm and extracellular localization, respectively. For all identifiers please refer to the BiGG database [23].

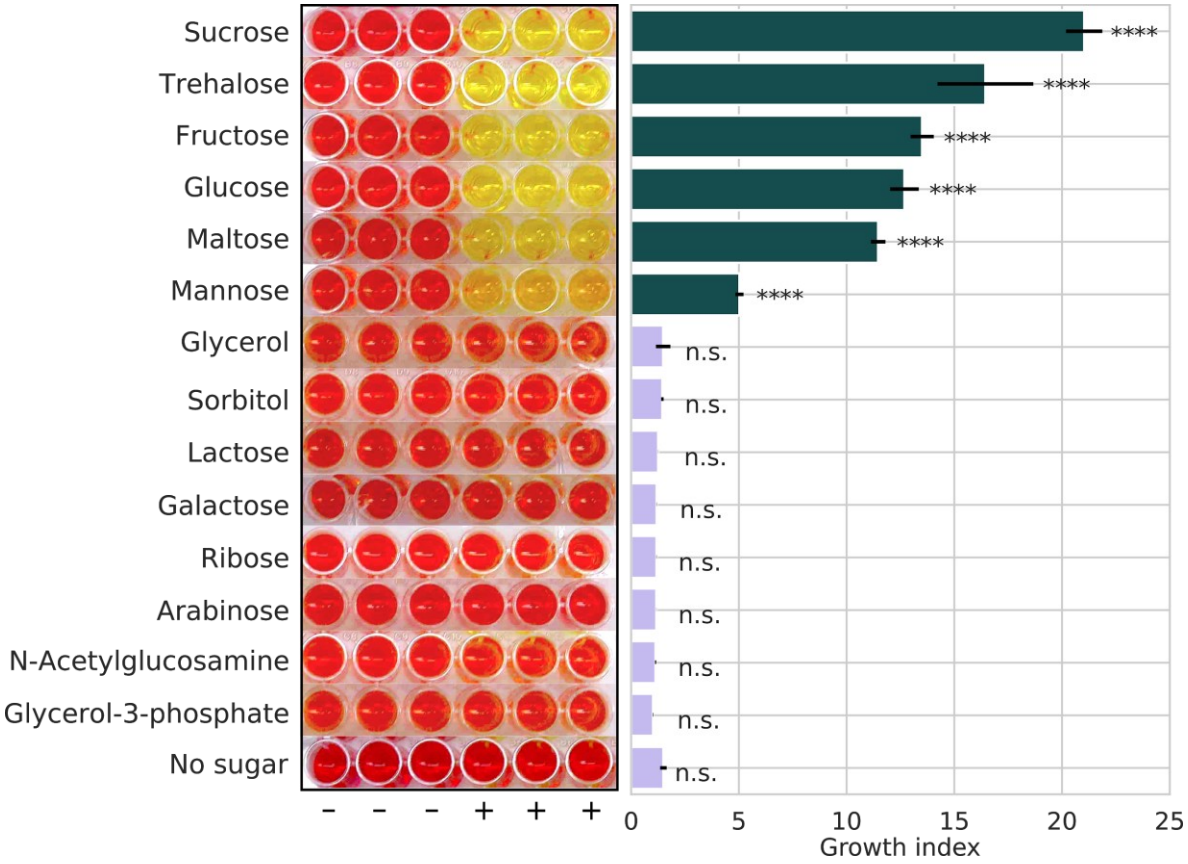

Figure S3.8 Experimental evaluation of *M. florum* growth on different carbohydrates. The phenol red, a pH indicator present in the CSY medium, changes color upon metabolic activity. The medium color (OD_{560nm}) observed after a 24 hour incubation period and normalized over a non-inoculated control, corresponding to the growth index, is reported for each carbohydrate tested in CSY. Carbohydrates were supplemented at 1% (w/v) final concentration. Bars and error bars indicate the mean and standard deviation calculated from technical triplicate, respectively (One-way ANOVA, '****' = $p < 0.0001$).

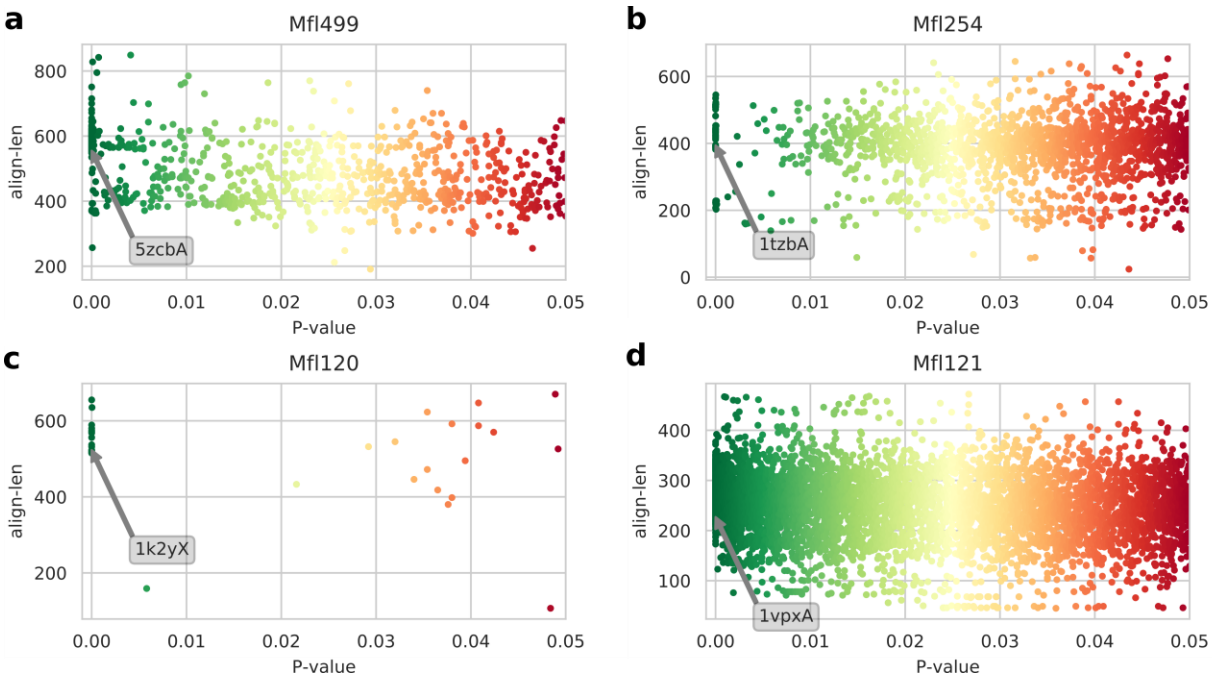

Figure S3.9 FATCAT 2.0 database alignment results. On each plot, significant alignments (P-value < 0.05) between a selected *M. florum* candidate protein and the Protein DataBank are shown. Alignments are separated according to their respective length (align-len) and associated P-value (coloured from red to green). **a** Alignment results for the Mfl499 protein. The arrow indicates a positive match with the A chain of the α-glucosidase of *Bacillus* sp. AHU2216 (5zcbA; *Bsp*AG13_31A) specific to α-(1-4)-glucosidic linkage. **b-d** Same as panel **a** but for **b** the Mfl254 protein and the A chain of the phosphoglucose/phosphomannose isomerase of *Pyrobaculum aerophilum* (1tzbA; PaPGI/PMI), **c** Mfl120 protein and the X chain of the phosphomannomutase/phosphoglucomutase of *Pseudomonas aeruginosa* (1k2yX; PMM/PGM), and **d** Mfl121 protein and the A chain of the transaldolase of *Thermotoga maritima* (1vpxA; TM0295).

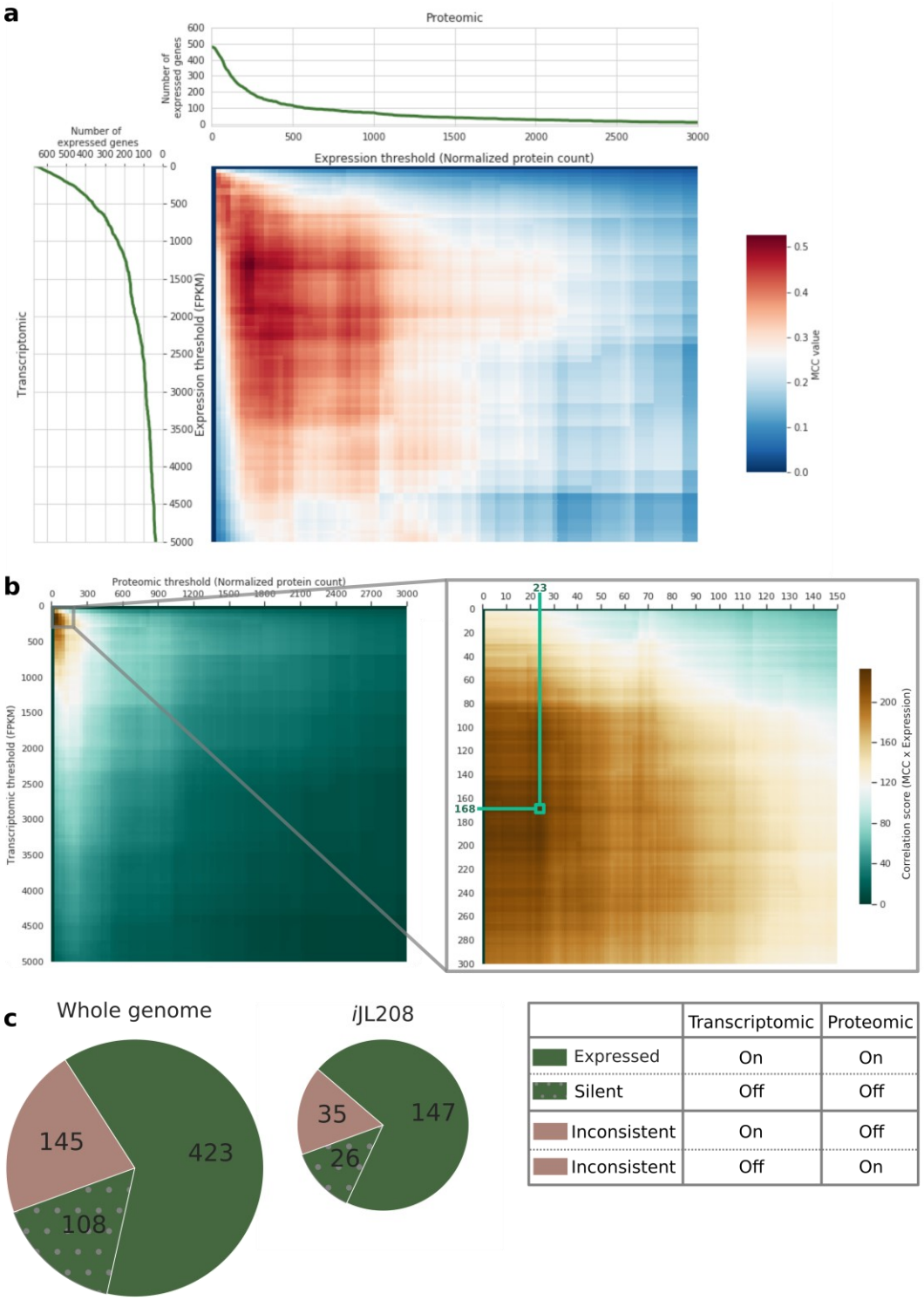

Figure S3.10 Determining the optimal expression thresholds for transcriptomic and proteomic datasets. **a** Applying thresholds to recently published transcriptomic (FPKM) and proteomic (protein molecules per cell) datasets impacts the resulting number of expressed

genes (green curves). For each pair of thresholds, the Matthews correlation coefficient (MCC) comparing binary vectors of gene expression status is shown in the matrix, where high (red) and low (blue) values indicate similarity between expression status from the two experimental datasets. **b** To identify the thresholds maximizing both the expression status and the number of genes consistently expressed, the correlation values (defined by the MCC, see panel A) were multiplied by the average number of expressed genes for each threshold pair, resulting in a correlation score matrix (left). The expression thresholds selected for comparison with the predicted metabolic fluxes were 23 proteins per cell for the proteomic experiment and an FPKM of 168 for the transcriptomic experiment (right), for a total of 423 genes considered expressed in both datasets. **c** Proportion of expressed genes in the entire genome (left) and in *i*JL208 (middle) according to both proteomic and transcriptomic data and selected thresholds**.** In *i*JL208, a total of 173 genes had a consistent expression status in both datasets.

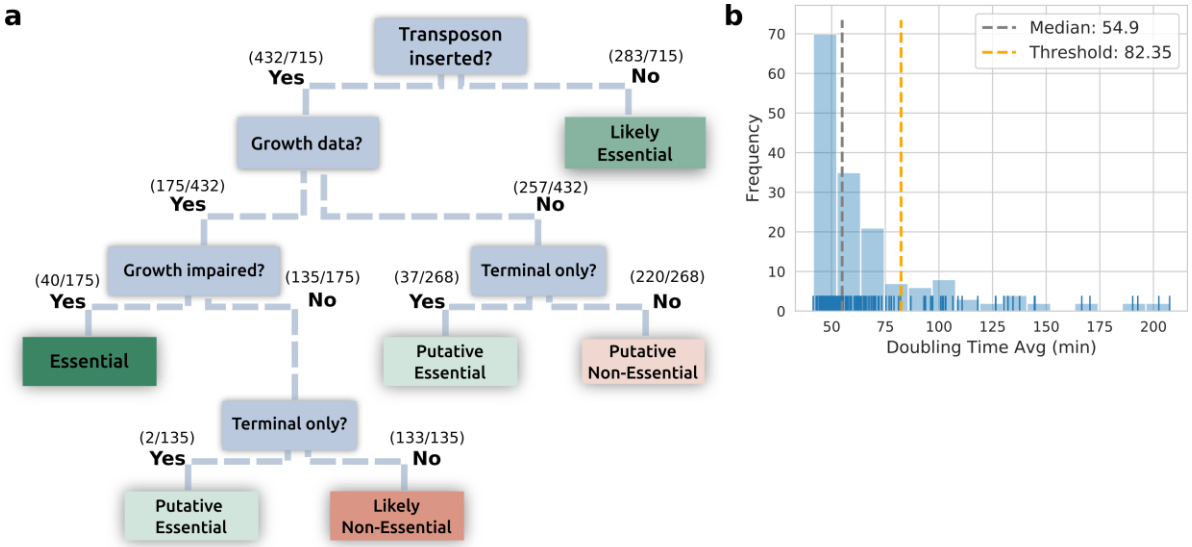

Figure S3.11 Revisiting the *M. florum* genome-wide essentiality data. a Transposon mutagenesis experiments were previously published for *M. florum* L1, but considered only the presence or absence of a transposon insertion into a gene to determine its essentiality [1]. This data was re-analyzed using the presented decision tree and now accounts for the relative position of the insertion site within the interrupted gene as well as growth data of the isolated mutants. For insertions not impairing the growth of *M. florum*, interrupted genes were considered essential only if the transposons were strictly restricted to the terminal region of genes, defined as the last 20% of the gene length. In downstream analyses, all likely and putative essentials were considered as essential genes, whereas likely and putative non-essentials were considered as non-essentials. **b** Defining growth impairment threshold. Histogram (blue bars) showing the distribution of doubling time for the 175 insertion mutants (blue ticks) with reliable growth data (Supplementary file 9). Mutants showing a doubling time higher than the sum of the median and the median absolute deviation were considered non-viable.

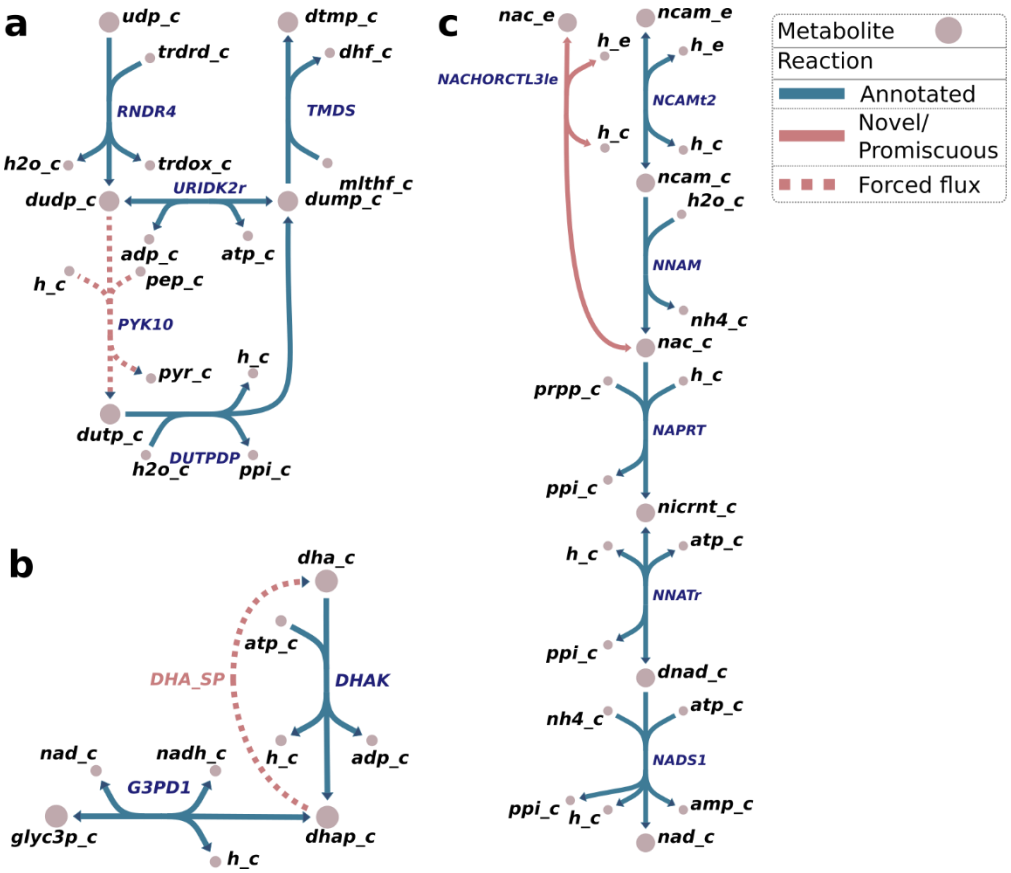

Figure S3.12 Resolving false negative and false positive predictions. a Forcing a flux through the PYK10 reaction simulated the production of deoxyuridine triphosphate (dutp), ensuring that the dUTPase carried flux and was essential. **b** Forcing the spontaneous production of dihydroxyacetone (dha) ensured that the DHAK carried flux and was essential. **c** Solving the only true false positive identified, the nicotinamidase Mfl340. Adding a transport reaction for nicotinate removes the need for nicotinamidase while keeping the necessity for the rest of the pathway, suggesting that nicotinate is readily available in the growth media.

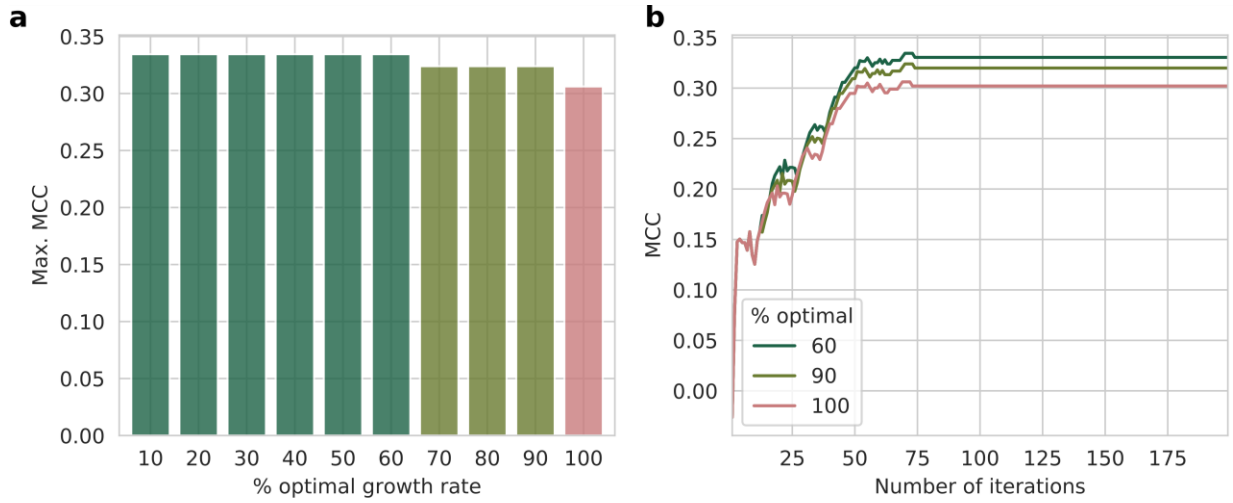

Figure S3.13 Impact of growth rate on reduced genome similarity with JCVI-syn3.0. a Maximal Matthews Correlation Coefficient (MCC) observed between JCVI-syn3.0 and different *M. florum* genome reduction possibilities generated by the minGenome algorithm with increasing growth rate constraints. Three different scenarios were identified based on their impact on the similarity with JCVI-syn3.0 (dark green: Low, 60% growth rate; light green: Intermediate, 90% growth rate; red: Optimal, 100% growth rate). **b** MCC calculated after each minGenome iteration which removes the largest possible stretch of genes in the genome. The three scenarios identified in A are presented.

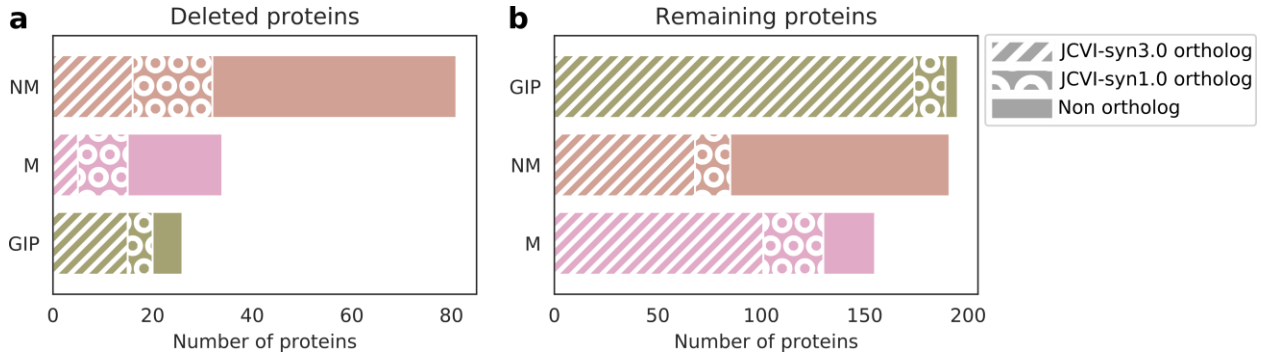

Figure S3.14 Distribution of deleted (a) and conserved (b) proteins from the minimal genome prediction in the KEGG functional categories presented in Figure 7C. The proteins homologous to JCVI-syn3.0 are represented by white hatched bars, while the genes that were absent from JCVI-syn3.0 but homologous to JCVI-syn1.0 are represented by white circles. The plain color represents the proteins present only in *M. florum*.

3.12 SUPPLEMENTARY FILES

All supplementary files are available on the *i*JL208 GitHub:
https://github.com/jclachance/iJL208

Supplementary file 1: Proteome comparison
Supplementary file 2: I-TASSER homology modeling information
Supplementary file 3: EC numbers
Supplementary file 4: Reconstruction manual curation
Supplementary file 5: Complete metabolic map in JSON format for visualization with Escher
Supplementary file 6: The *i*JL208 model in JSON format
Supplementary file 7: Media information
Supplementary file 8: FATCAT 2.0 structural comparison output
Supplementary file 9: Gene expression and essentiality
Supplementary file 10: Genome reduction scenarios

CHAPITRE 4
DISCUSSION ET CONCLUSION

4.1 RÉSUMÉ DU PROJET DE RECHERCHE

Mon projet de doctorat visait à développer des outils de biologie des systèmes afin de participer à la caractérisation exhaustive d'une cellule quasi-minimale qui pourra servir de châssis pour la biologie synthétique. Le but du développement d'un tel châssis est l'établissement de règles de conception pour la construction de cellules artificielles. Ces standards sont le nerf de la guerre des disciplines d'ingénierie et permettent l'avancement rapide de projets complexes tout en réduisant la probabilité que des erreurs majeures se produisent.

En ce sens, le premier objectif de mon projet de doctorat était de produire un logiciel capable de déterminer et de standardiser la définition de la BOF pour les modèles métaboliques. Le logiciel BOFdat a été produit et publié dans la revue *PloS Computational Biology* (Lachance et al., 2019b). Pour définir la BOF, BOFdat prend en entrée des ensembles de données omiques qui sont maintenant générés de manière routinière et encourage aussi l'incorporation de la composition macromoléculaire de la cellule. De plus, BOFdat permet une sélection des métabolites qui composent la biomasse d'une manière algorithmique non-biaisée. Ensemble, l'utilisation de données expérimentales pour la définition de l'équation de biomasse et la sélection non-biaisée assurent que la définition d'une équation de biomasse soit spécifique à l'espèce modélisée et reflète la réalité expérimentale. Compte tenu de l'importance de la fonction objective de biomasse pour la qualité des prédictions phénotypiques, ce genre de standards est nécessaire pour l'avancement de la biologie synthétique.

Le second objectif de mon projet de doctorat était de réaliser un modèle métabolique pour la bactérie quasi-minimale *Mesoplasma florum*. Plusieurs avancées ont contribué à rendre cette bactérie apte à devenir un châssis cellulaire pour la biologie synthétique (Baby et al., 2018a, 2018b; Labroussaa et al., 2016; Matteau et al., 2017) mais aucun effort de modélisation n'y avait encore été dédié. Le modèle métabolique *i*JL208 a donc été généré, d'abord en s'appuyant sur une révision de la qualité de l'annotation du génome entier par des approches bio-informatiques, puis en définissant l'équation de biomasse à partir de la composition macromoléculaire de la cellule et de données omiques provenant d'une étude approfondie de caractérisation intégrative à laquelle j'ai aussi contribué (Matteau et al., 2020). Ensuite, la formulation d'un milieu de croissance semi-défini a permis de distinguer la croissance sur différents sucres, ce qui a permis de définir les taux d'absorptions et de sécrétions de métabolites clés à la croissance de *M. florum*. Ces paramètres ont été appliqués comme contraintes expérimentales sur le modèle. Il a été possible d'analyser la sensibilité du modèle par rapport à celles-ci et puis de les fixer de manière définitive pour bien refléter la croissance en laboratoire de cet organisme quasi-minimal. Cette analyse a entre autres permis de suggérer que la production d'acétate par *M. florum* est dépendante de l'apport en oxygène chez cette bactérie anaérobie facultative. Une fois contraint par des données expérimentales, les prédictions de *i*JL208 ont pu être testées afin de valider les capacités métaboliques attendues, l'utilisation des réactions métaboliques et l'essentialité des gènes. Les différences entre les prédictions du modèle et la réalité expérimentale ont été utilisées afin d'améliorer la qualité du modèle dans un processus itératif.

Finalement, le dernier objectif de mon doctorat était d'utiliser un modèle *in silico* afin de formuler une prédiction de génome minimal pour *M. florum*. Le modèle contraint et validé *i*JL208 a donc été utilisé afin de définir un génome de 552 kpb contenant 535 séquences codant pour des protéines. Pour arriver à cette fin, j'ai utilisé l'algorithme MinGenome (Wang and Maranas, 2018), développé par le laboratoire du Pr. Maranas. Formulé comme un problème d'optimisation, cet algorithme trouve la plus grande délétion qui ne touche pas de gènes

essentiels et permet au modèle de se résoudre au taux de croissance imposé. En l'appliquant itérativement, il devient éventuellement impossible de retirer davantage de gènes dans le génome de l'organisme.

Ensemble, ces accomplissements remplissent les objectifs fixés au début de mon doctorat. Cependant, aucun projet de science n'est parfait et les méthodes utilisées présentent certaines lacunes qui peuvent être améliorées. La section suivante discute de la qualité des méthodes déployées et des perspectives d'avenir.

4.2 BOFdat 2.0: COMMENT MIEUX DÉTERMINER LA BIOMASSE OU LES OBJECTIFS CELLULAIRES?

Le développement de BOFdat était novateur car il s'agissait du premier logiciel dédié entièrement à la définition de la fonction objective de biomasse pour les modèles métaboliques. L'implémentation du logiciel en trois étapes indépendantes permet une certaine flexibilité pour les utilisateurs qui peuvent dès lors choisir la ou les étapes qui correspondent le mieux à leur projet.

La première étape de BOFdat consiste à calculer les coefficients stoechiométriques (CS) pour les composantes des trois macromolécules principales de la cellule (ADN, ARN et protéines). La méthode de calcul suggère d'utiliser les données omiques (génomique, transcriptomique et protéomique) avec la composition macromoléculaire de la cellule afin de calculer les CS pour tous les nucléotides et acides aminés (Figure 2.1). Pour ce calcul, il a été possible de montrer que l'impact des données omiques était moindre que celui de la composition macromoléculaire (Figure S2.7). Il s'agit d'un résultat intéressant car il démontre que les différents acides aminés utilisés dans les protéines sont uniformément distribués. Il est possible d'imaginer qu'un tel scénario possède un avantage évolutif en ne forçant pas l'assimilation ou la production d'un

acide aminé précis, ce qui pourrait rapidement devenir critique en cas de carence. Un des objectifs initiaux de cette approche était aussi de fournir la possibilité de calculer les CS pour d'autres molécules importantes de la cellule tels que les lipides et les glycans. Cependant, la diversité de formats des ensembles de données et des types de molécules engendraient un problème au niveau de la standardisation de ce calcul. La décision a donc été prise de reporter le calcul des CS pour ces objectifs cellulaires précis aux étapes subséquentes de BOFdat qui proposent des approches non biaisées.

L'objectif de la seconde étape est de trouver les coenzymes et cofacteurs qui devraient faire partie de l'équation de biomasse. Pour ce faire, BOFdat utilise le niveau de connectivité des métabolites dans le réseau. Basée sur l'idée que les cofacteurs et coenzymes sont recyclés et utilisés dans plusieurs réactions, ils devraient pouvoir être identifiés en comptant le nombre de réactions auxquelles un métabolite contribue. En fixant un seuil, il est possible de déterminer que certains métabolites se retrouvent dans beaucoup plus de réactions que la majorité des métabolites du réseau (Figure S2.1). Un inconvénient de cette approche est que plusieurs métabolites qui ne sont pas généralement considérés comme des coenzymes sont alors inclus. Il a donc fallu introduire une liste de métabolites fortement connectés à ne pas inclure dans l'équation à cette étape (ex.: H^+, ATP, etc…). Ce biais est moins intéressant pour la découverte de coenzymes et a donc justifié l'utilisation conjointe de la liste de coenzymes formulée par Xavier *et al.* (Xavier et al., 2017). Ceci s'explique par le fait que la définition d'un coenzyme est basée sur l'accumulation de savoir biologique qui permet de distinguer un coenzyme des métabolites réguliers. Dans ce contexte, une définition algorithmique devient difficile et l'utilisation du savoir existant devrait être privilégiée.

La troisième étape utilise un algorithme génétique (Fortin et al., 2012) afin d'identifier les métabolites composants la BOF. Cette approche novatrice a permis d'améliorer légèrement la qualité des prédictions d'essentialité de gènes formulée par *i*ML1515, le modèle de *E. coli* le plus récent (Monk et al., 2017). Le choix d'un algorithme génétique semblait approprié pour

ce problème puisqu'en permettant une sélection aléatoire des combinaisons de métabolites faisant partie de la biomasse, il n'introduisait pas de biais initial et pouvait potentiellement améliorer la qualité des prédictions formulées. Le choix d'optimiser la composition de l'équation de biomasse pour l'essentialité des gènes était justifié par la qualité des prédictions d'essentialité pouvant être faites avec le modèle qui sont reconnues pour être très élevées. Cependant, cette prédiction est limitée par la linéarité des voies de synthèse des métabolites qui sont choisis. Dans une voie linéaire, il n'existe qu'un seul moyen de produire un métabolite essentiel, alors que dans une voie branchée, des réactions alternatives permettent de le produire. Cette réalité combinée au fait qu'un algorithme génétique est un heuristique qui ne génère pas de solution unique, rendait l'interprétation des résultats produits par BOFdat difficile. Pour mieux les intégrer, un algorithme de regroupement, dans ce cas DBSCAN (Ester et al., 1996), a été utilisé afin de grouper les métabolites provenant des meilleures compositions obtenues par l'algorithme génétique. Le nombre de réactions entre les métabolites, mesuré par l'algorithme de Dijkstra (Dijkstra and Others, 1959), fournit une mesure de distance que DBSCAN utilise pour grouper les métabolites ensembles. Ces regroupements sont intéressants car ils représentent des ensembles de métabolites d'une même voie de synthèse reliés à un même objectif cellulaire. Un usager du modèle peut alors utiliser ces groupes afin de mieux comprendre, à travers les données d'essentialité, les métabolites importants que la bactérie tente de synthétiser.

Une avenue intéressante utilisant BOFdat qui n'a pas été poursuivie au cours de mon doctorat consistait à utiliser la troisième étape de BOFdat avec des données d'essentialité provenant de différentes conditions de croissance. Comme les gènes essentiels sont dépendants des conditions de culture, il est possible d'identifier des gènes conditionnellement essentiels à partir des données expérimentales (Zhao et al., 2017). L'identification de ces gènes est intéressante car elle révèle l'adaptation d'un organisme aux différentes conditions. Il est aussi possible de grouper ces gènes par fonctions cellulaires et ainsi d'obtenir une idée globale des changements opérés par la cellule pour s'adapter. En utilisant un modèle métabolique et l'étape

trois de BOFdat pour analyser ce genre de données, il aurait été possible d'identifier la variation de la composition de l'équation de biomasse en fonction des conditions. Les métabolites différentiellement identifiés révéleraient donc des objectifs cellulaires spécifiques à certaines conditions. L'identification de ces métabolites ou des voies de synthèses qui leurs sont reliées permettrait, par exemple, de mieux diriger le développement de nouveaux antibiotiques car les objectifs cellulaires spécifiques à certaines conditions sont critiques à la survie des bactéries.

Le développement de BOFdat a donc permis de fournir à la communauté scientifique utilisant la modélisation métabolique un logiciel structuré qui permet d'assurer un standard pour la définition de la composition de biomasse. Fait intéressant, moins de deux ans après sa publication, BOFdat a été cité 25 fois. L'approche non-biaisée pour définir les métabolites importants pour la cellule à partir des données d'essentialité ouvre la voie à de nouvelles découvertes. Une limitation de cette approche est la nécessité d'un réseau métabolique bien connu pour arriver à correctement identifier les objectifs métaboliques. Finalement, la définition des objectifs cellulaires reste un aspect fort intéressant de la modélisation cellulaire lorsque formulé comme un problème d'optimisation. Ce sujet peut être particulièrement intéressant pour les industries pour arriver à comprendre les limitations de croissance dans un ensemble de conditions.

En guise d'ouverture, il convient d'observer d'autres approches qui ont été développées afin de définir les objectifs cellulaires. Ces approches utilisent d'autres ensembles de données que l'essentialité des gènes et pourraient aider à mieux caractériser les objectifs cellulaires lorsque ces données ne sont pas disponibles. Par exemple, les données d'AFM ont été utilisées afin de retracer les flux métaboliques dans une approche qui se nomme BOSS (Gianchandani et al., 2008). Contrairement à d'autres approches telles que ObjFind (Burgard and Maranas, 2003) et invFBA (Zhao et al., 2016) qui identifient des coefficients d'importance sur certaines réactions en fonction des données de fluxomique, BOSS permet de reconstruire une fonction objective *de novo*. Pour y arriver, BOSS implémente une méthode d'optimisation à deux niveaux qui: 1)

réduit la somme du carré des erreurs entre les flux *mesurés* expérimentalement et les flux prédits *in silico* qui eux sont soumis au FBA standard, et 2) maximise une fonction objective hypothétique initiale, nécessaire à l'optimisation. Dans le manuscrit original, BOSS a été validé sur le réseau métabolique central de la levure, un réseau composé de 60 métabolites participant à 62 réactions. Une alternative intéressante nommée BIG-BOSS sur laquelle j'ai eu la chance de travailler (Yang et al., 2019a), réussi à appliquer BOSS à un plus grand ensemble de réactions et est donc applicable à un réseau métabolique complet. En plus d'être applicable à un réseau métabolique complet, cette approche a aussi montré la capacité de combiner des données de fluxomique et de protéomique pour déterminer l'objectif cellulaire. Encore plus intéressant, la combinaison de ces ensembles de données conduit à une augmentation de la qualité des flux prédits par la fonction objective générée. Pour inclure les quantités de protéines dans le modèle, une modification au problème de FBA standard est appliquée tel qu'utilisé par Beg *et al*. (Beg et al., 2007). L'utilisation de la concentration de protéines est utile car les données de transcriptomiques ou de protéomiques utilisées pour les estimer sont générées de manière plus courante que les données de fluxomique et, en outre, sont en mesure de couvrir l'ensemble du protéome d'un organisme, ce qui n'est pas le cas pour les flux métaboliques.

4.3 ÉTENDRE LES AVENUES DE MODÉLISATION

Le modèle *i*JL208 représente environ 30% des gènes codant pour une protéine chez *M. florum* (208/682), un pourcentage similaire à celui du modèle de *E. coli* (Monk et al., 2017). Ce modèle basé sur la méthode FBA représente le métabolisme de cette bactérie quasi-minimale. Comme sa réalisation ne requérait pas la connaissance des constantes enzymatiques, il s'agissait d'un bon point de départ pour le développement d'outils de biologie des systèmes pour cet organisme. Les avenues de modélisations actuellement disponibles ne sont toutefois pas limitées à la méthode FBA et au métabolisme des organismes. Il est possible d'affiner les prédictions formulées par le FBA en incluant certaines contraintes supplémentaires sur ce

cadre. Cette section discute des avenues de modélisation subséquentes qui pourront être implémentées chez *M. florum* pour atteindre un niveau de connaissance exhaustif.

Le premier pas vers un raffinement de *i*JL208 serait l'inclusion de l'encombrement moléculaire par une méthode nommée FBAwMC (Beg et al., 2007). Cette méthode suppose que, puisque le volume cytoplasmique total d'une cellule est limité, il ne peut contenir qu'un nombre fini d'enzymes. La distribution de ces enzymes à l'intérieur de la cellule impose donc un compromis sur les enzymes préférentiellement exprimées. La concentration d'une enzyme est liée à son flux par les paramètres de la réaction et après quelques manipulations, la formulation du FBAwMC devient:

$$\sum_{i=1}^{N} a_i f_i \leq 1$$
$$\text{où, } a_i = C v_i / b_i \qquad \text{(équation 4.1)}$$
$$\text{et, } f_i = b_i E_i$$

où ***C*** représente la densité cytoplasmique donnée par la masse de la cellule et son volume (*M/V*) et le paramètre ***b*** est donné par le mécanisme de la réaction, les paramètres cinétiques et la concentration des métabolites. On voit que, dans cette formulation mathématique, le paramètre ***b*** est simplifié et n'est donc plus requis. Dans notre récente caractérisation exhaustive de la cellule, les paramètres ***M***, ***V*** et ***C***, ont été clairement identifiés pour *M. florum* (Matteau et al., 2020). Ainsi, la densité cytoplasmique varie entre 1.05 à 1.08g/mL, permettant de fixer la contrainte ***C***. Dans cette formulation, ***E*** représente la concentration des protéines qui peut être inférée à partir des données d'expression obtenue par transcriptomique et par protéomique qui sont elles-aussi disponibles.

En utilisant cette approche, les auteurs ont démontré qu'il était possible de récapituler adéquatement l'utilisation différentielle de substrats par *E. coli* (Beg et al., 2007). Notre étude a montré que l'élaboration d'un milieu de culture semi-défini permettait de distinguer la

croissance sur différents sucres. En poursuivant le développement de ce milieu vers une définition complète, il sera fort probablement possible de déterminer l'utilisation séquentielle des différents sucres chez *M. florum* par HPLC. Le déploiement d'un modèle FBAwMC basé sur les données disponibles permettra donc d'établir une comparaison entre les prédictions et ces observations expérimentales. Il sera intéressant de vérifier si une cellule quasi-minimale est soumise à un ensemble de contraintes similaires à celles de *E. coli*. Compte tenu de la réduction de complexité du réseau métabolique de *M. florum*, il est plus facile d'imaginer les restrictions qui s'appliqueront ici. Les sucres plus complexes comme le sucrose, le tréhalose et le maltose sont probablement utilisés en dernier puisqu'ils requièrent une étape de digestion additionnelle. Le fructose pourrait aussi être consommé avant le glucose et le mannose puisqu'il épargne une étape de la glycolyse. Cependant, ces prédictions pourraient être erronées puisque l'histoire évolutive de *M. florum* joue probablement un rôle dans l'expression préférentielle des substrats. Par exemple, la souche de laboratoire *M. florum* L1 est systématiquement cultivée en ATCC 1161 supplémenté avec une haute concentration de sucrose. Il est donc possible que la disponibilité de ce substrat ait poussé la souche vers une adaptation qui favorise ce substrat. La combinaison des prédictions d'un modèle FBAwMC et des données de croissance en milieu défini permettront de faire la lumière sur ce sujet.

Il serait aussi possible d'inclure spécifiquement les contraintes enzymatiques connues en utilisant la boîte à outils GECKO (https://github.com/SysBioChalmers/GECKO) développée par (Sánchez et al., 2017). Cette approche impose une limite sur la capacité des enzymes à catalyser des réactions qui tiennent compte des paramètres cinétiques et de leur concentration telle que mesurée par l'expression des gènes. Pour ce faire, la matrice stoechiométrique est modifiée afin d'inclure les enzymes en tant que métabolites dans les rangées et leur utilisation est inclus dans les colonnes. Les coefficients stoechiométriques pour les nouvelles colonnes et rangées sont donnés par la constante catalytique (k_{cat}). Formellement, la concentration des protéines obtenue par données d'expression est utilisée comme borne supérieure sur les réactions. Les constantes catalytiques pour le model Yeast7 (Aung et al., 2013) utilisées dans l'étude originale ont été tirées de la base de données BRENDA (Schomburg et al., 2013). Il

serait possible d'utiliser une approche semblable pour obtenir les constantes catalytiques pour *M. florum*. Évidemment, le manque d'information spécifique à l'organisme pourrait causer un biais, mais ce modèle pourrait être amélioré au fil du temps lorsque davantage de caractérisation biochimique sera effectuée ou encore en utilisant des méthodes informatiques de déterminations des constantes catalytiques basées sur l'apprentissage machine (Heckmann et al., 2018).

Dans leur étude, Sanchez *et al.* ont démontré que le modèle contraint par les constantes enzymatiques et les abondances de protéines arrivait à récapituler le phénomène de fermentation en conditions aérobies, aussi appelé surplus métabolique (*overflow metabolism*). Compte tenu de l'absence d'une voie de respiration chez *M. florum*, il ne sera pas possible d'étudier ce phénomène en implémentant ce modèle. Ceci dit, notre étude a révélé que le choix métabolique de sécréter du lactate ou de l'acétate était vraisemblablement corrélé avec le niveau d'expression des enzymes responsables de leur production. Nous avons aussi spécifiquement contraint le flux maximal de la NADH oxidase afin de refléter ce phénomène. Cette contrainte unique a eu un impact métabolique important en découplant le taux de croissance prédit par le modèle de l'import d'oxygène (Figure 3.5). Il est donc fort possible que l'application de contraintes supplémentaires basées sur l'abondance des protéines et des constantes catalytiques à l'échelle du réseau métabolique résulte en une description plus précise des limitations de *M. florum*.

Dans notre étude, nous avons comparé les données d'expressions aux flux métaboliques en utilisant la méthode de pFBA (Lewis et al., 2010). Il s'agit d'une approche permettant d'augmenter la qualité des prédictions de flux par rapport aux données d'expression qui ne requiert pas d'ajuster les limites de flux métabolique ni la disponibilité de données d'expression sur des conditions différentes, ou encore de données de fluxomique (Tian and Reed, 2018). Cette comparaison a révélé une précision de ~78%. Bien qu'il s'agisse d'une précision initiale satisfaisante, il est possible de l'améliorer. Une cause d'erreur non négligeable est la relation entre les données de transcriptomique et de protéomique qui, même

au seuil optimal, n'était pas parfaite (Figure S3.10). Mise à part cette limitation expérimentale, l'utilisation des réactions prédites par le FBA est basée sur la règle d'association entre les réactions et les gènes. Il est possible que des erreurs se soient glissées là aussi, potentiellement parce que l'annotation du génome n'est pas parfaite et peut-être aussi parce que des erreurs ont été introduites par le modélisateur, moi-même, au moment de la reconstruction. Cependant, il pourrait être intéressant de constater ce qu'il se produirait en générant un modèle capable de produire ses enzymes de la même manière que le FBA arrive à simuler la production de métabolites.

Nous avons discuté de ce type de modèle à la section 1.6 et avions mentionné comment la reconstruction d'une matrice d'expression par Thiele *et al.* (Thiele et al., 2009) avait mené à la reconstruction de modèles du métabolisme et de l'expression (ME) (Lerman et al., 2012; Lloyd et al., 2018; O'brien et al., 2013; Thiele et al., 2012). Ces modèles sont en mesure de reproduire les niveaux d'expression d'une très grande partie des gènes du génome (Salvy and Hatzimanikatis, 2020). En utilisant la boîte à outils COBRAme (Lloyd et al., 2018), il serait possible de reconstruire un tel modèle pour *M. florum*. Comme l'a montré notre étude, environ les deux tiers des protéines du génome sont comprises dans la combinaison du métabolisme ou des mécanismes d'expression génique (Figure 3.7). Ce nombre de gènes est aussi appuyé par une étude phylogénétique qui a montré que le noyau de l'appareil de traduction des mollicutes conserve 129 gènes (Grosjean et al., 2014). Parmi les 39 mollicutes étudiés dans cette étude, *M. florum* partage le second plus grand nombre de protéines associées à la traduction (>160) avec *B. subtilis* et *E. coli*. La réalisation d'un modèle ME permettrait donc de couvrir plus de la moitié des gènes de *M. florum*. Cette couverture rapprocherait *M. florum* d'un niveau nécessaire afin de le définir comme châssis cellulaire pour la biologie synthétique et permettrait aussi une augmentation de la qualité des prédictions d'expression génique par rapport au modèle métabolique. La production d'un modèle ME permettrait aussi de modéliser la résistance de *M. florum* à différents stress physico-chimiques tels que le stress oxydatif (Yang et al., 2019b), de pH (Du et al., 2019) et de choc thermique (Chen et al., 2017). Ces stress

peuvent facilement être imposés en laboratoire et des données d'expression pourraient être générées afin de valider efficacement les prédictions du modèle.

La promesse d'un modèle représentant l'ensemble des gènes codant pour des protéines est cependant dépendante de la capacité de caractériser l'ensemble des fonctions encodées dans le génome de *M. florum.* Comme notre étude l'a montré, une portion significative des protéines de *M. florum* (258/676, ~38%) ont une fonction biologique moins bien caractérisée ou inconnue. Comme démontré dans notre étude, la structure tridimensionnelle des protéines peut aussi être reconstruite *in silico*. Nous avons réussi à fournir 361 nouvelles structures en utilisant l'algorithme I-TASSER (Yang et al., 2015a) qui, combinées à l'information contextuelle provenant des expériences de croissance sur différents sucres, ont permis de suggérer des réactions alternatives catalysées par des enzymes dont l'annotation était connue. Bien que la prédiction de fonctions protéiques à partir de la séquence d'acides aminés soit encore hors de portée, l'avancement de méthodes d'intelligence artificielle tel que l'apprentissage profond pourraient faire changer ce paradigme. En effet, ces méthodes ont déjà démontré des applications en bio-informatique et particulièrement dans l'élucidation des numéros EC (Ryu et al., 2019) et de la reconstruction de structures tridimensionnelles (Senior et al., 2020). Le développement d'algorithmes basés sur les ordinateurs quantiques pourraient eux-aussi devenir un joueur important dans ce domaine (Robert et al., 2019). Il est donc possible que la caractérisation biochimique soit prochainement remplacée ou complémentée par des approches informatiques à haut-débit permettant de prédire les fonctions des protéines, ce qui permettrait des avancées majeures en modélisation.

4.4 PERSPECTIVES SUR L'UTILISATION DES MODÈLES POUR LA CONCEPTION DE GÉNOMES

La reconstruction de *i*JL208 a permis de concevoir des scénarios de réduction de génome chez l'organisme quasi-minimal qu'est *M. florum*. Cette réduction s'est effectuée en utilisant

l'algorithme MinGenome tel que publié par Wang et Maranas (Wang and Maranas, 2018). La proximité phylogénétique de *M. florum* avec JCVI-syn3.0 (Hutchison et al., 2016) a permis d'établir une comparaison détaillée des fonctions conservées dans les génomes minimaux. D'une part, notre prédiction a identifié huit fonctions cellulaires non touchées par des délétions (ATP synthase, facteurs de traduction, ARN polymérase, adressage de protéines, biosynthèse des cofacteurs, système de relais du soufre, métabolisme des lipides et voie des pentoses phosphates). Aussi, trois fonctions cellulaires métaboliques étaient entièrement partagées avec JCVI-syn3.0 (ATP synthase, métabolisme des acides aminés et système de sécrétion). La majorité des gènes contenus dans la catégorie de traitement de l'information génétique étaient partagés entre JCVI-syn3.0 et *M. florum* (174/195). Cinq sous-catégories fonctionnelles du traitement de l'information génétique étaient entièrement partagées avec JCVI-syn3.0 (facteurs de traduction, chargement et maturation des ARNt, ARN polymérase, système de relais du soufre, adressage des protéines). Cet ensemble de sous-catégories partagées avec JCVI-syn3.0 représente les fonctions clés d'une cellule minimale qui ont pu être révélées par notre approche de réduction basée sur *i*JL208.

De manière importante, notre étude a aussi permis de démontrer que de varier les contraintes imposées sur le problème d'optimisation formulé par MinGenome avait un impact sur le nombre de gènes inclus dans les scénarios de réductions. Ici, varier le taux de croissance imposé a permis de générer trois scénarios de réduction possible. En réduisant le taux de croissance maximal permis (augmentation du temps de doublement), davantage de gènes ont pu être retirés dans la prédiction de génome minimal, ce qui est consistant avec l'augmentation du temps de doublement observé entre JCVI-syn1.0 (60 min, 1138 gènes (Gibson et al., 2010)), JCVI-syn3.0A (120 min, 498 gènes (Breuer et al., 2019)) et JCVI-syn3.0 (180 min, 473 gènes (Hutchison et al., 2016)). Cette constatation est intéressante car elle montre l'importance d'une définition adéquate des contraintes de modélisation pour la construction de génomes. On peut donc poser l'hypothèse que ce principe est généralisable et que les méthodes de modélisation incluant davantage de contraintes énumérées à la section précédente résulteront en des conceptions de génome différentes, potentiellement plus précises.

En comparant notre approche à d'autres algorithmes développés pour la prédiction de génomes minimaux (Rees-Garbutt et al., 2020), nous avons constaté que l'imposition de contraintes résulte en un nombre de gènes plus élevé que les approches de modélisation basées sur des équations différentielles ordinaires (ODE). Ce format est utilisé par le modèle cellulaire complet de *Mycoplasma genitalium* (Karr et al., 2012). Le nombre de gènes contenus dans ces prédictions (360 et 380) propose des génome contenant ~25% moins de gènes que JCVI-syn3.0 et est dangereusement inférieur au nombre de gènes essentiels (382) obtenus par insertion de transposons dans le génome de *M. genitalium* (Glass et al., 2006). Il est connu que le nombre de gènes essentiels obtenus par cette méthode sous-estime le nombre réel de gènes essentiels car il ne tient pas compte du phénomène de létalité synthétique (Hutchison et al., 2016). Une prédiction contenant moins de gènes que le nombre de gènes essentiels est donc assurément non fonctionnelle. De plus, les génomes minimaux générés au cours de cette étude ne permettaient d'obtenir qu'un seul doublement cellulaire et permettait aussi d'enlever des gènes reconnus comme essentiels. Il est fort probable qu'un format de modélisation basé sur des ODE ne soit pas compatible avec la conception de génomes *in silico*.

La conception de génomes ne se limite cependant pas à la définition de génomes minimaux. Il deviendra fort intéressant d'intégrer de nouvelles voies de synthèse dans *M. florum* afin d'y tester des conceptions innovantes, en couplant par exemple la croissance à la production d'un métabolite d'intérêt. Le logiciel Cameo (Cardoso et al., 2018) fournit une boîte à outils facilitant le choix de voies de synthèses hétérologues, le choix de knock-outs de gènes et de modulations d'expression qui peuvent aider à concevoir la souche voulue. Suivant la conception d'une telle souche, une approche nommée TFA (Henry et al., 2007) permet d'incorporer les contraintes thermodynamiques dans les modèles et peut donc assurer la faisabilité des voies de synthèse proposées. Une boîte à outils permet son incorporation facile dans les modèles (Salvy et al., 2019). Contrairement aux constantes enzymatiques, les contraintes thermodynamiques peuvent être inférées par des méthodes bio-informatique telles

que eQuilibrator (Flamholz et al., 2011). L'implémentation de ces contraintes assurerait que le métabolisme des génomes modifiés soit thermodynamiquement faisable.

Au cours de notre étude, le développement d'un milieu semi-défini et la reconstruction du réseau métabolique de *M. florum* ont permis de paver la voie vers un milieu complètement défini. Lorsque ce milieu sera disponible, il deviendra possible de tester la croissance de *M. florum* en modulant les composantes du milieu. Comme nous l'avons montré, ce type de validation expérimentale permettra de confirmer les prédictions du modèle par rapport aux capacités métaboliques de *M. florum* (Figure 3.6). En outre, générer des données d'expression pour l'ensemble des conditions testées pourrait permettre de reconstruire l'architecture de régulation de la bactérie en appliquant la méthode d'analyse indépendante des composantes (ICA). Appliquée à un ensemble de données de transcriptomique, cette méthode a montré qu'il est possible de décomposer l'ensemble de ces données en modules régulés indépendamment (Sastry et al., 2019). L'identification de ces modules chez *M. florum* permettra d'organiser son génome relativement à ses capacités adaptatives. Il s'agit d'une couche supplémentaire d'information qui sera définitivement utile lors de la conception de génomes.

La réalisation expérimentale d'un génome désiré et conçu à l'aide d'un modèle basé sur les contraintes devrait cependant pouvoir être réalisée par synthèse chimique d'ADN. Il a été montré qu'avec les méthodes actuelles, certaines régions du génome peuvent être impossibles à synthétiser. Par exemple, les sections de pourcentage GC trop élevé ou trop bas peuvent nécessiter le recodage de certains codons (Venetz et al., 2019). Ce processus de refactorisation nécessite une connaissance approfondie de l'ensemble des fonctions de chacune des paires de base du génome. Heureusement, un cadre bio-informatique récemment développé pourrait permettre d'incorporer l'ensemble de ces informations chez *M. florum*. Le Bitome (Lamoureux et al., 2020) est une matrice dont chacune des colonnes représente une paire de base du génome et chaque rangée représente une fonctionnalité différente. Les *bits* sont codés dans chaque case de cette matrice et représentent la présence (1) ou l'absence (0) d'une fonctionnalité à une paire de base donnée. La réalisation d'un Bitome spécifique à *M. florum* est envisageable compte

tenu de la grande quantité d'information rendue disponible par l'étude de caractérisation intégrative réalisée par notre laboratoire (Matteau et al., 2020) et devrait faciliter la conception de génomes.

4.5 CONCLUSION

La complétion de *i*JL208 marque la première étape importante dans le développement d'outils de biologie des systèmes pour *M. florum*. En lien avec les hypothèses du projet de recherche, il a été possible d'utiliser ce modèle afin de faire une prédiction de génome minimal pour cette bactérie, démontrant le potentiel de l'utilisation de modèles cellulaires *in silico* à cette fin. Cette plateforme servira de base pour le raffinement de l'approche FBA par des contraintes plus spécifiques et devrait permettre l'élaboration de conceptions de génome plus détaillées chez *M. florum*. En combinant les connaissances acquises par l'expérience et le développement de cadres bio-informatiques tels que mentionnés précédemment, il sera possible de faire de *M. florum* un châssis cellulaire pour la biologie synthétique (Danchin, 2012). De cette réalisation émaneront sans doute des règles de conceptions de génomes desquelles les ingénieurs pourront s'inspirer pour systématiquement générer des organismes viables aux propriétés souhaitées.

"What I cannot create, I do not understand"
-Richard Feynman

ANNEXE

AUTRES PUBLICATIONS PERTINENTES

1. The Use of *In Silico* Genome-Scale Models for the Rational Design of Minimal Cells

Abstract: Organism-specific genome-scale metabolic models (GEMs) can be reconstructed using genome annotation and biochemical data available in literature. The systematic inclusion of biochemical reactions into a coherent metabolic network combined with the formulation of appropriate constraints reveals the set of metabolic capabilities harbored by an organism, hereby allowing the computation of growth phenotypes from genotype information. GEMs have been used thoroughly to assess growth capabilities under varying conditions and determine gene essentiality. This simulation process can rapidly generate testable hypotheses that can be applied for the systematic evaluation of growth capabilities in genome reduction efforts and the definition of a minimal cell. Here we review the most recent computational methods and protocols available for the reconstruction of genome-scale models, the formulation of objective functions, and the applications of models in the prediction of gene essentiality. These methods and applications are suited to the emerging field of genome reduction and the development of minimal cells as biological factories

Référence bibliographique: Lachance, J.-C., Rodrigue, S., and Palsson, B.O. (2020). The Use of *In Silico* Genome-Scale Models for the Rational Design of Minimal Cells. In Minimal Cells: Design, Construction, Biotechnological Applications, A.R. Lara, and G. Gosset, eds. (Cham: Springer International Publishing), pp. 141–175.

Lien URL: https://link.springer.com/chapter/10.1007/978-3-030-31897-0_6

2. Synthetic Biology: Minimal cells, maximal knowledge

Abstract: Modeling all the chemical reactions that take place in a minimal cell will help us understand the fundamental interactions that power life.

Référence bibliographique: Lachance, J.-C., Rodrigue, S., and Palsson, B.O. (2019a). Minimal cells, maximal knowledge. Elife *8*.

Lien URL: https://elifesciences.org/articles/45379

BIBLIOGRAPHIE

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput. Biol. *9*, e1002980.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. *9*, 3015–3027.

Andrianantoandro, E., Basu, S., Karig, D.K., and Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. Mol. Syst. Biol. *2*, 2006.0028.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. *40*, W597–W603.

Ataman, M., and Hatzimanikatis, V. (2017). lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. PLoS Comput. Biol. *13*, e1005513.

Aung, H.W., Henry, S.A., and Walker, L.P. (2013). Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. Ind. Biotechnol. *9*, 215–228.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. BMC Genomics *9*, 75.

Baby, V., Lachance, J.-C., Gagnon, J., Lucier, J.-F., Matteau, D., Knight, T., and Rodrigue, S. (2018a). Inferring the Minimal Genome of *Mesoplasma florum* by Comparative Genomics and Transposon Mutagenesis. mSystems *3*.

Baby, V., Labroussaa, F., Brodeur, J., Matteau, D., Gourgues, G., Lartigue, C., and Rodrigue, S. (2018b). Cloning and Transplantation of the *Mesoplasma florum* Genome. ACS Synth. Biol. *7*, 209–217.

Bassalo, M.C., Garst, A.D., Halweg-Edwards, A.L., Grau, W.C., Domaille, D.W., Mutalik, V.K., Arkin, A.P., and Gill, R.T. (2016). Rapid and Efficient One-Step Metabolic Pathway

Integration in E. coli. ACS Synth. Biol. *5*, 561–568.

Bates, M., Lachoff, J., Meech, D., Zulkower, V., Moisy, A., Luo, Y., Tekotte, H., Franziska Scheitz, C.J., Khilari, R., Mazzoldi, F., et al. (2017). Genetic Constructor: An Online DNA Design Platform. ACS Synth. Biol. *6*, 2362–2365.

Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. Nat. Rev. Microbiol. *12*, 841–850.

Bautista, E.J., Zinski, J., Szczepanek, S.M., Johnson, E.L., Tulman, E.R., Ching, W.-M., Geary, S.J., and Srivastava, R. (2013). Semi-automated Curation of Metabolic Models via Flux Balance Analysis: A Case Study with *Mycoplasma gallisepticum*.

Beaucage, S.L., and Caruthers, M.H. (1981). Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. Tetrahedron Lett. *22*, 1859–1862.

Beck, A.E., Hunt, K.A., and Carlson, R.P. (2018). Measuring Cellular Biomass Composition for Computational Biology Applications. Processes *6*, 38.

Beg, Q.K., Vazquez, A., Ernst, J., de Menezes, M.A., Bar-Joseph, Z., Barabási, A.-L., and Oltvai, Z.N. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc. Natl. Acad. Sci. U. S. A. *104*, 12663–12668.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science *277*, 1453–1462.

Bordbar, A., Jamshidi, N., and Palsson, B.O. (2011). iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. BMC Syst. Biol. *5*, 110.

Bordbar, A., Monk, J.M., King, Z.A., and Palsson, B.O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. *15*, 107–120.

Bordbar, A., Yurkovich, J.T., Paglia, G., Rolfsson, O., Sigurjónsson, Ó.E., and Palsson, B.O. (2017). Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. Sci. Rep. *7*, 46249.

Breuer, M., Earnest, T.M., Merryman, C., Wise, K.S., Sun, L., Lynott, M.R., Hutchison, C.A., Smith, H.O., Lapek, J.D., Gonzalez, D.J., et al. (2019). Essential metabolism for a minimal cell. Elife *8*.

Bridges, C.B. (1922). The Origin of Variations in Sexual and Sex-Limited Characters. Am. Nat. *56*, 51–63.

Brower, V. (2001). Proteomics: biology in the post-genomic era. Companies all over the world rush to lead the way in the new post-genomics race. EMBO Rep. *2*, 558–560.

Burgard, A.P., and Maranas, C.D. (2003). Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol. Bioeng. *82*, 670–677.

Cardoso, J.G.R., Jensen, K., Lieven, C., Lærke Hansen, A.S., Galkina, S., Beber, M., Özdemir, E., Herrgård, M.J., Redestig, H., and Sonnenschein, N. (2018). Cameo: A Python Library for Computer Aided Metabolic Engineering and Optimization of Cell Factories. ACS Synth. Biol. *7*, 1163–1166.

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., et al. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. *36*, D623–D631.

Chen, K., Gao, Y., Mih, N., O'Brien, E.J., Yang, L., and Palsson, B.O. (2017). Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc. Natl. Acad. Sci. U. S. A. *114*, 11548–11553.

Choe, D., Cho, S., Kim, S.C., and Cho, B.-K. (2016). Minimal genome: Worthwhile or worthless efforts toward being smaller? Biotechnol. J. *11*, 199–211.

Clark, T.B., Tully, J.G., Rose, D.L., Henegar, R., and Whitcomb, R.F. (1986). Acholeplasmas and similar nonsterol-requiring mollicutes from insects: missing link in microbial ecology. Curr. Microbiol. *13*, 11–16.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Dajani, A.S., Clyde, W.A., Jr, and Denny, F.W. (1965). Experimental infection with *Mycoplasma pneumoniae* (Eaton's agent). J. Exp. Med. *121*, 1071–1086.

Danchin, A. (2012). Scaling up synthetic biology: Do not forget the chassis. FEBS Lett. *586*, 2129–2137.

Danchin, A., and Fang, G. (2016). Unknown unknowns: essential genes in quest for function. Microb. Biotechnol. *9*, 530–540.

Dandekar, T., Huynen, M., Regula, J.T., Ueberle, B., Zimmermann, C.U., Andrade, M.A., Doerks, T., Sánchez-Pulido, L., Snel, B., Suyama, M., et al. (2000). Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. Nucleic Acids Res. *28*, 3278–3288.

Danna, K., and Nathans, D. (1971). Specific cleavage of simian virus 40 DNA by restriction endonuclease of *Hemophilus influenzae*. Proc. Natl. Acad. Sci. U. S. A. *68*, 2913–2917.

Deutscher, D., Meilijson, I., Kupiec, M., and Ruppin, E. (2006). Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nat. Genet. *38*, 993–998.

Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A.A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol. Biol. *985*, 17–45.

Dias, O., Rocha, M., Ferreira, E.C., and Rocha, I. (2015). Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Res. *43*, 3899–3910.

Dijkstra, E.W., and Others (1959). A note on two problems in connexion with graphs. Numer. Math. *1*, 269–271.

Dikicioglu, D., Kırdar, B., and Oliver, S.G. (2015). Biomass composition: the "elephant in the room" of metabolic modelling. Metabolomics *11*, 1690–1701.

Matteau D, Lachance J-C, Grenier F, Gauthier S, Daubenspeck JM, Dybvig K, et al. Integrative characterization of the near-minimal bacterium *Mesoplasma florum*. Mol Syst Biol. 2020;16: e9844.

Du, B., Yang, L., Lloyd, C.J., Fang, X., and Palsson, B.O. (2019). Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in *Escherichia coli*. PLoS Comput. Biol. *15*, e1007525.

Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst. Biol. *7*, 74.

Ebrahim, A., Brunk, E., Tan, J., O'Brien, E.J., Kim, D., Szubin, R., Lerman, J.A., Lechner, A., Sastry, A., Bordbar, A., et al. (2016). Multi-omic data integration enables discovery of hidden biological regularities. Nat. Commun. *7*, 13091.

Edwards, J.S., and Palsson, B.O. (1999). Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype. J. Biol. Chem. *274*, 17410–17416.

Edwards, J.S., and Palsson, B.O. (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. U. S. A. *97*, 5528–5533.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., and Others (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, pp. 226–231.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2017). The Reactome Pathway Knowledgebase. Nucleic Acids Res.

Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Curr. Opin. Microbiol.

13, 344–349.

Flamholz, A., Noor, E., Bar-Even, A., and Milo, R. (2011). eQuilibrator—the biochemical thermodynamics calculator. Nucleic Acids Res. *40*, D770–D775.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science *269*, 496–512.

Fortin, F.-A., Rainville, F.-M.D., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). DEAP: Evolutionary Algorithms Made Easy. J. Mach. Learn. Res. *13*, 2171–2175.

Francois Jacob, David Perrin, Carmen Sanchez, Jacques Monod (1960). The Operon: A Group of Genes Whose Expression is Coordinated by an Operator. Comptes-Rendus Des Seances de L'academie Des Sciences *250*, 1727–1729.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. Science *270*, 397–403.

Gelius-Dietrich, G., Desouki, A.A., Fritzemeier, C.J., and Lercher, M.J. (2013). Sybil--efficient constraint-based modelling in R. BMC Syst. Biol. *7*, 125.

Gianchandani, E.P., Oberhardt, M.A., Burgard, A.P., Maranas, C.D., and Papin, J.A. (2008). Predicting biological system objectives de novo from internal state measurements. BMC Bioinformatics *9*, 43.

Gibson, D.G. (2009). Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. Nucleic Acids Res. *37*, 6984–6990.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., et al. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. Science *319*, 1215–1220.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. Science *329*, 52–56.

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., 3rd, Smith, H.O., and Venter, J.C. (2006). Essential genes of a minimal bacterium. Proc. Natl. Acad. Sci. U. S. A. *103*, 425–430.

Glass, J.I., Merryman, C., Wise, K.S., Hutchison, C.A., and Smith, H.O. (2017). Minimal Cells—Real and Imagined. Cold Spring Harb. Perspect. Biol.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Goodson, H.V., Anderson, B.L., Warrick, H.M., Pon, L.A., and Spudich, J.A. (1996). Synthetic lethality screen identifies a novel yeast myosin I gene (MYO5): myosin I proteins are required for polarization of the actin cytoskeleton. J. Cell Biol. *133*, 1277–1291.

Grosjean, H., Breton, M., Sirand-Pugnet, P., Tardy, F., Thiaucourt, F., Citti, C., Barré, A., Yoshizawa, S., Fourmy, D., de Crécy-Lagard, V., et al. (2014). Predicting the minimal translation apparatus: lessons from the reductive evolution of mollicutes. PLoS Genet. *10*, e1004363.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., et al. (2009). Transcriptome complexity in a genome-reduced bacterium. Science *326*, 1268–1271.

Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., and Ralser, M. (2017). Designing and interpreting “multi-omic” experiments that may change our understanding of biology. Current Opinion in Systems Biology *6*, 37–45.

Hartleb, D., Jarre, F., and Lercher, M.J. (2016). Improved Metabolic Models for *E. coli* and *Mycoplasma genitalium* from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. PLoS Comput. Biol. *12*, e1005036.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1–8.

Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., and Palsson, B.O. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nat. Commun. *9*, 5252.

Heinemann, M., and Panke, S. (2006). Synthetic biology—putting engineering into biology. Bioinformatics *22*, 2790–2799.

Heirendt, L., Thiele, I., and Fleming, R.M.T. (2017). DistributedFBA.jl: high-level, high-performance flux balance analysis in Julia. Bioinformatics *33*, 1421–1423.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. *14*, 639–

702.

Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. Biophys. J. *92*, 1792–1805.

Hirokawa, Y., Kawano, H., Tanaka-Masuda, K., Nakamura, N., Nakagawa, A., Ito, M., Mori, H., Oshima, T., and Ogasawara, N. (2013). Genetic manipulations restored the growth fitness of reduced-genome *Escherichia coli*. J. Biosci. Bioeng. *116*, 52–58.

Holley, R.W. (1965). Structure of an alanine transfer ribonucleic acid. JAMA *194*, 868–871.

Hughes, R.A., and Ellington, A.D. (2017). Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harb. Perspect. Biol. *9*.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., and Venter, J.C. (1999). Global transposon mutagenesis and a minimal Mycoplasma genome. Science *286*, 2165–2169.

Hutchison, C.A., 3rd, Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., et al. (2016). Design and synthesis of a minimal bacterial genome. Science *351*, aad6253.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Joyce, A.R., and Palsson, B.Ø. (2008). Predicting gene essentiality using genome-scale *in silico* models. Methods Mol. Biol. *416*, 433–457.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr, Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

Kauffman, K.J., Prakash, P., and Edwards, J.S. (2003). Advances in flux balance analysis. Curr. Opin. Biotechnol. *14*, 491–496.

King, B., Farrah, T., Richards, M.A., Mundy, M., Simeonidis, E., and Price, N.D. (2018). ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions. Bioinformatics *34*, 1594–1596.

King, Z.A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N.E., and Palsson, B.O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. PLoS Comput. Biol. *11*, e1004321.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., and Lewis, N.E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. *44*, D515–D522.

Koonin, E.V. (2000). How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. *1*, 99–116.

Koonin, E.V., and Mushegian, A.R. (1996). Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. Curr. Opin. Genet. Dev. *6*, 757–762.

Koonin, E.V., Mushegian, A.R., and Bork, P. (1996). Non-orthologous gene displacement. Trends Genet. *12*, 334–336.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. *5*, R7.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235–1240.

Labroussaa, F., Lebaudy, A., Baby, V., Gourgues, G., Matteau, D., Vashee, S., Sirand-Pugnet, P., Rodrigue, S., and Lartigue, C. (2016). Impact of donor–recipient phylogenetic distance on bacterial genome transplantation. Nucleic Acids Res. *44*, 8501–8511.

Lachance, J.-C., Rodrigue, S., and Palsson, B.O. (2019a). Minimal cells, maximal knowledge. Elife *8*.

Lachance, J.-C., Lloyd, C.J., Monk, J.M., Yang, L., Sastry, A.V., Seif, Y., Palsson, B.O., Rodrigue, S., Feist, A.M., King, Z.A., et al. (2019b). BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. PLOS Computational Biology *15*, e1006971.

Lachance, J.-C., Rodrigue, S., and Palsson, B.O. (2020). The Use of *In Silico* Genome-Scale Models for the Rational Design of Minimal Cells. In Minimal Cells: Design, Construction, Biotechnological Applications, A.R. Lara, and G. Gosset, eds. (Cham: Springer International Publishing), pp. 141–175.

Lahner, B., Gong, J., Mahmoudian, M., Smith, E.L., Abid, K.B., Rogers, E.E., Guerinot, M.L., Harper, J.F., Ward, J.M., McIntyre, L., et al. (2003). Genomic scale profiling of nutrient and

trace elements in Arabidopsis thaliana. Nat. Biotechnol. *21*, 1215–1221.

Lamoureux, C.R., Choudhary, K.S., King, Z.A., Sandberg, T.E., Gao, Y., Sastry, A.V., Phaneuf, P.V., Choe, D., Cho, B.-K., and Palsson, B.O. (2020). The Bitome: digitized genomic features reveal fundamental genome organization. Nucleic Acids Res. *48*, 10157–10163.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison, C.A., 3rd, Smith, H.O., and Venter, J.C. (2007). Genome transplantation in bacteria: changing one species to another. Science *317*, 632–638.

LeProust, E.M. (2016). Rewriting DNA synthesis. Chem. Eng. Prog. *2016*, 30–35.

Lerman, J.A., Hyduke, D.R., Latif, H., Portnoy, V.A., Lewis, N.E., Orth, J.D., Schrimpe-Rutledge, A.C., Smith, R.D., Adkins, J.N., Zengler, K., et al. (2012). *In silico* method for modelling metabolism and gene product expression at genome scale. Nat. Commun. *3*, 929.

Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. *6*, 390.

Lewis, N.E., Nagarajan, H., and Palsson, B.O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. Nat. Rev. Microbiol. *10*, 291–305.

Lind, K. (1966). Isolation of *Mycoplasma pneumoniae* (Eaton agent) from patients with primary atypical pneumonia. Acta Pathol. Microbiol. Scand. *66*, 124–134.

Lloyd, C.J., Ebrahim, A., Yang, L., King, Z.A., Catoiu, E., O'Brien, E.J., Liu, J.K., and Palsson, B.O. (2018). COBRAme: A computational framework for genome-scale models of metabolism and gene expression. PLoS Comput. Biol. *14*, e1006302.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics *14*, 265–279.

Majewski, R.A., and Domach, M.M. (1990). Simple constrained-optimization view of acetate overflow in *E. coli*. Biotechnol. Bioeng. *35*, 732–738.

Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J.M., Corrêa, I.R., Jr, and Romesberg, F.E. (2014). A semi-synthetic organism with an expanded genetic alphabet. Nature *509*, 385–388.

Matteau, D., Pepin, M.-E., Baby, V., Gauthier, S., Arango Giraldo, M., Knight, T.F., and Rodrigue, S. (2017). Development of oriC-based plasmids for *Mesoplasma florum*. Appl. Environ. Microbiol.

Matteucci, M.D., and Caruthers, M.H. (1981). Synthesis of deoxyoligonucleotides on a polymer support. J. Am. Chem. Soc. *103*, 3185–3191.

McCoy, R.E., Basham, H.G., Tully, J.G., Rose, D.L., Carle, P., and Bové, J.M. (1984). *Acholeplasma florum*, a new species isolated from plants. Int. J. Syst. Evol. Microbiol. *34*, 11–15.

McGuire, A.L., Colgrove, J., Whitney, S.N., Diaz, C.M., Bustillos, D., and Versalovic, J. (2008). Ethical, legal, and social considerations in conducting the Human Microbiome Project. Genome Res. *18*, 1861–1864.

Miles, R.J. (1992). Catabolism in mollicutes. J. Gen. Microbiol. *138*, 1773–1783.

Monk, J., Nogales, J., and Palsson, B.O. (2014). Optimizing genome-scale network reconstructions. Nat. Biotechnol. *32*, 447–452.

Monk, J.M., Lloyd, C.J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., et al. (2017). *i*ML1515, a knowledgebase that computes *Escherichia coli* traits. Nat. Biotechnol. *35*, 904–908.

Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). MetaNetX/MNXref--reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. *44*, D523–D526.

Morowitz, H.J. (1984). Special guest lecture: The completeness of molecular biology. Isr. J. Med. Sci. *2*.

Morowitz, H.J., and Tourtellotte, M.E. (1962). The smallest living cells. Sci. Am. *206*, 117–126.

Mørtz, E., O'Connor, P.B., Roepstorff, P., Kelleher, N.L., Wood, T.D., McLafferty, F.W., and Mann, M. (1996). Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. Proc. Natl. Acad. Sci. U. S. A. *93*, 8264–8267.

Mushegian, A.R., and Koonin, E.V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. U. S. A. *93*, 10268–10273.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., and O'Neal,

C. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc. Natl. Acad. Sci. U. S. A. *53*, 1161–1168.

Nirenberg, M.W., Jones, O.W., Leder, P., Clark, B.F.C., Sly, W.S., and Pestka, S. (1963). On the Coding of Genetic Information. Cold Spring Harb. Symp. Quant. Biol. *28*, 549–557.

Nursimulu, N., Xu, L.L., Wasmuth, J.D., Krukov, I., and Parkinson, J. (2018). Improved enzyme annotation with EC-specific cutoffs using DETECT v2. Bioinformatics *34*, 3393–3395.

O'brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., and Palsson, B.Ø. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. *9*, 693.

Orth, J.D., and Palsson, B.Ø. (2010). Systematizing the generation of missing metabolic knowledge. Biotechnol. Bioeng. *107*, 403–412.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Palsson, B. (2015). Systems Biology (Cambridge University Press).

Pan, S., and Reed, J.L. (2018). Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. Curr. Opin. Biotechnol. *51*, 103–108.

Papoutsakis, E.T. (1984). Equations and calculations for fermentations of butyric acid bacteria. Biotechnol. Bioeng. *26*, 174–187.

Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., and Schomburg, D. (2017). BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res. *45*, D380–D388.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Rees-Garbutt, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., and Grierson, C. (2020). Designing minimal genomes using whole-cell models. Nat. Commun. *11*, 836.

Reuß, D.R., Altenbuchner, J., Mäder, U., Rath, H., Ischebeck, T., Sappa, P.K., Thürmer, A., Guérin, C., Nicolas, P., Steil, L., et al. (2017). Large-scale reduction of the *Bacillus subtilis* genome: consequences for the transcriptional network, resource allocation, and metabolism. Genome Res. *27*, 289–299.

Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., DiCarlo, J.E.,

Lee, D., Huang, C.L.V., Chandrasegaran, S., Cai, Y., et al. (2017). Design of a synthetic yeast genome. Science *355*, 1040–1044.

Riekeberg, E., and Powers, R. (2017). New frontiers in metabolomics: from measurement to insight. F1000Res. *6*, 1148.

Robert, A., Barkoutsos, P.K., Woerner, S., and Tavernelli, I. (2019). Resource-Efficient Quantum Algorithm for Protein Folding.

Roberts, R.J. (2005). How restriction enzymes became the workhorses of molecular biology. Proc. Natl. Acad. Sci. U. S. A. *102*, 5905–5908.

Ryu, J.Y., Kim, H.U., and Lee, S.Y. (2019). Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc. Natl. Acad. Sci. U. S. A. *116*, 13996–14001.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science *230*, 1350–1354.

Salvy, P., and Hatzimanikatis, V. (2020). The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models. Nat. Commun. *11*, 30.

Salvy, P., Fengos, G., Ataman, M., Pathier, T., Soh, K.C., and Hatzimanikatis, V. (2019). pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. Bioinformatics *35*, 167–169.

Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E.J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol. Syst. Biol. *13*, 935.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977a). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., III, Slocombe, P.M., and Smith, M. (1977b). Nucleotide sequence of bacteriophage φX174 DNA. Nature *265*, 687.

Sastry, A.V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K.S., Yang, L., King, Z.A., and Palsson, B.O. (2019). The *Escherichia coli* transcriptome mostly consists of independently regulated modules. Nat. Commun. *10*, 5536.

Satish Kumar, V., Dasika, M.S., and Maranas, C.D. (2007). Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics *8*, 212.

Savinell, J.M., and Palsson, B.O. (1992a). Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. J. Theor. Biol. *155*, 201–214.

Savinell, J.M., and Palsson, B.O. (1992b). Optimal selection of metabolic fluxes for in vivo measurement. II. Application to *Escherichia coli* and hybridoma cell metabolism. J. Theor. Biol. *155*, 215–242.

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., et al. (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. *41*, D764–D772.

Schrodinger, E. (1967). What is Life?: The Physical Aspect of the Living Cell and Mind and Matter; Mind and Matter (Cambridge University Press).

Segrè, D., Vitkup, D., and Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. U. S. A. *99*, 15112–15117.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710.

Shlomi, T., Berkman, O., and Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc. Natl. Acad. Sci. U. S. A. *102*, 7695–7700.

Sinsheimer, R.L. (1989). The Santa Cruz Workshop—May 1985. Genomics *5*, 954–956.

Sleator, R.D. (2010). The story of *Mycoplasma mycoides* JCVI-syn1.0: the forty million dollar microbe. Bioeng. Bugs *1*, 229–230.

Smith, H.O., and Wilcox, K.W. (1970). A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. J. Mol. Biol. *51*, 379–391.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*, 674–679.

Smolke, C., Lee, S.Y., Nielsen, J., and Stephanopoulos, G. (2018). Synthetic Biology: Parts, Devices and Applications (John Wiley & Sons).

Spencer, G. (2008). International consortium announces the 1000 Genomes project. See Http://www.1000genomes.org/bcms/1000_genomes/Documents/1000Genomes-NewsRelease.Pdf.

Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M., and Heyneker, H.L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides.

Gene *164*, 49–53.

Suthers, P.F., Zomorrodi, A., and Maranas, C.D. (2009a). Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol. Syst. Biol. *5*, 301.

Suthers, P.F., Dasika, M.S., Kumar, V.S., Denisov, G., Glass, J.I., and Maranas, C.D. (2009b). A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. PLoS Comput. Biol. *5*, e1000285.

Thiele, I., and Palsson, B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. *5*, 93–121.

Thiele, I., Jamshidi, N., Fleming, R.M.T., and Palsson, B.Ø. (2009). Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput. Biol. *5*, e1000312.

Thiele, I., Fleming, R.M.T., Que, R., Bordbar, A., Diep, D., and Palsson, B.O. (2012). Multiscale modeling of metabolism and macromolecular synthesis in *E. coli* and its application to the evolution of codon usage. PLoS One *7*, e45635.

Tian, M., and Reed, J.L. (2018). Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics *34*, 3882–3888.

Tully, J.G. (1983). The Emmy Klieneberger-Nobel Award lecture. Reflections on recovery of some fastidious mollicutes with implications of the changing host patterns of these organisms. Yale J. Biol. Med. *56*, 799–813.

Tully, J.G., Bové, J.M., Laigret, F., and Whitcomb, R.F. (1993). Revised Taxonomy of the Class Mollicutes: Proposed Elevation of a Monophyletic Cluster of Arthropod-Associated Mollicutes to Ordinal Rank (Entomoplasmatales ord. nov.), with Provision for Familial Rank To Separate Species with Nonhelical Morphology (Entomoplasmataceae fam. nov.) from Helical Species (Spiroplasmataceae), and Emended Descriptions of the Order Mycoplasmatales, Family Mycoplasmataceae. Int. J. Syst. Evol. Microbiol. *43*, 630–630.

Varma, A., and Palsson, B.O. (1993). Metabolic capabilities of *Escherichia coli*: I. synthesis of biosynthetic precursors and cofactors. J. Theor. Biol. *165*, 477–502.

Venetz, J.E., Del Medico, L., Wölfle, A., Schächle, P., Bucher, Y., Appert, D., Tschan, F., Flores-Tinoco, C.E., van Kooten, M., Guennoun, R., et al. (2019). Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. Proc. Natl. Acad. Sci. U. S. A. *116*, 8070–8079.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome.

Science *291*, 1304–1351.

Waddington, C.H. (1961). Molecular biology or ultrastructural biology? Nature *190*, 184.

Waites, W., Mısırlı, G., Cavaliere, M., Danos, V., and Wipat, A. (2018). A Genetic Circuit Compiler: Generating Combinatorial Genetic Circuits with Web Semantics and Inference. ACS Synth. Biol.

Wang, L., and Maranas, C.D. (2018). MinGenome: An *In Silico* Top-Down Approach for the Synthesis of Minimized Genomes. ACS Synth. Biol. *7*, 462–473.

Watson, J.D., Crick, F.H.C., and Others (1953). Molecular structure of nucleic acids. Nature *171*, 737–738.

Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E.M., Disz, T., Gabbard, J.L., et al. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. *45*, D535–D542.

Wodke, J.A.H., Puchałka, J., Lluch-Senar, M., Marcos, J., Yus, E., Godinho, M., Gutiérrez-Gallego, R., dos Santos, V.A.P.M., Serrano, L., Klipp, E., et al. (2013). Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. Mol. Syst. Biol. *9*, 653.

Xavier, J.C., Patil, K.R., and Rocha, I. (2017). Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. Metab. Eng. *39*, 200–208.

Yang, K., and Han, X. (2016). Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. Trends Biochem. Sci. *41*, 954–969.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015a). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8.

Yang, L., Tan, J., O'Brien, E.J., Monk, J.M., Kim, D., Li, H.J., Charusanti, P., Ebrahim, A., Lloyd, C.J., Yurkovich, J.T., et al. (2015b). Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. Proc. Natl. Acad. Sci. U. S. A. *112*, 10810–10815.

Yang, L., Ma, D., Ebrahim, A., Lloyd, C.J., Saunders, M.A., and Palsson, B.O. (2016). solveME: fast and reliable solution of nonlinear ME models. BMC Bioinformatics *17*, 391.

Yang, L., Saunders, M.A., Lachance, J.-C., Palsson, B.O., and Bento, J. (2019a). Estimating Cellular Goals from High-Dimensional Biological Data. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, (New York, NY, USA: ACM), pp. 2202–2211.

Yang, L., Mih, N., Anand, A., Park, J.H., Tan, J., Yurkovich, J.T., Monk, J.M., Lloyd, C.J., Sandberg, T.E., Seo, S.W., et al. (2019b). Cellular responses to reactive oxygen species are predicted from molecular mechanisms. Proc. Natl. Acad. Sci. U. S. A. *116*, 14368–14373.

Yurkovich, J.T., Yang, L., and Palsson, B.O. (2017). Biomarkers are used to predict quantitative metabolite concentration profiles in human red blood cells. PLoS Comput. Biol. *13*, e1005424.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., Wodke, J.A.H., Güell, M., Martínez, S., Bourgeois, R., et al. (2009). Impact of genome reduction on bacterial metabolism and its regulation. Science *326*, 1263–1268.

Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). 13C-based metabolic flux analysis. Nat. Protoc. *4*, 878.

Zhao, L., Anderson, M.T., Wu, W., T Mobley, H.L., and Bachman, M.A. (2017). TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. BMC Bioinformatics *18*, 326.

Zhao, Q., Stettner, A.I., Reznik, E., Paschalidis, I.C., and Segrè, D. (2016). Mapping the landscape of metabolic goals of a cell. Genome Biol. *17*, 109.

Zomorrodi, A.R., and Maranas, C.D. (2010). Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data. BMC Syst. Biol. *4*, 178.