Modélisation de voies métaboliques pour la production de molécules

Ophélie Lo-Thong-Viramoutou

THÈSE DE DOCTORAT

UNIVERSITÉ DE LA RÉUNION
ÉCOLE DOCTORALE SCIENCES, TECHNOLOGIES ET SANTÉ

Thèse en Sciences spécialité Bioinformatique

Modélisation de voies métaboliques pour la production de molécules

Par

Ophélie LO-THONG-VIRAMOUTOU

Le 27 Mai 2021

Devant le jury composé de :

Rapporteurs :	
Pr. Aitao LI	Professeur, School of Life Sciences - Hubei University
Dr. Birgit WILTSCHI	Docteur, Austrian Center of Industrial Biotechnology
Pr. Narayanaswamy SRINIVASAN	Professeur, Indian Institute of Science
Examinateurs :	
Pr. Brigitte GRONDIN-PÉREZ	Professeur, Université de la Réunion
Pr. Bernard OFFMANN	Professeur, Université de Nantes
Dr. Emma SAAVEDRA	Docteur, Instituto Nacional de Cardiología Ignacio Chávez
Directeurs de thèse :	
Pr. Frédéric CADET	Professeur, Université de la Réunion *(Directeur)*
Dr. HDR. Cédric DAMOUR	Maître de conférences, Université de la Réunion *(Co-Directeur)*
Membre invité :	
Dr. Philippe CHARTON	Maître de conférences, Université de la Réunion *(Co-Encadrant)*

UR UNIVERSITÉ DE LA RÉUNION SIMB Inserm La science pour la santé From science to health #RÉGIONRÉUNION l'Europe s'engage à La Réunion avec le FEDER

UNIVERSITÉ DE LA RÉUNION
ÉCOLE DOCTORALE SCIENCES, TECHNOLOGIES ET SANTÉ

DOCTORAL THESIS

Thesis in Sciences speciality Computational Biology

Metabolic pathways modeling for the production of molecules

By

Ophélie LO-THONG-VIRAMOUTOU

Defended on the 27th May 2021

Jury composed by :

Rapporteurs :	
Pr. Aitao LI	Professor, School of Life Sciences - Hubei University
Dr. Birgit WILTSCHI	Doctor, Austrian Center of Industrial Biotechnology
Pr. Narayanaswamy SRINIVASAN	Professor, Indian Institute of Science
Examiners :	
Pr. Brigitte GRONDIN-PÉREZ	Professor, Université de la Réunion
Pr. Bernard OFFMANN	Professor, Université de Nantes
Dr. Emma SAAVEDRA	Doctor, Instituto Nacional de Cardiología Ignacio Chávez
Thesis directors :	
Dr. HDR. Cédric DAMOUR	Lecturer, Université de la Réunion *(Co-Director)*
Pr. Frédéric CADET	Professor, Université de la Réunion *(Director)*
Invited member :	
Dr. Philippe CHARTON	Lecturer, Université de la Réunion *(Co-Supervisor)*

À mes parents, mes piliers dans la joie
comme dans l'affliction,
À mon cher époux, qui n'a cessé de me
pousser dans mes retranchements

« Il distribue de merveilleux conseils et augmente les capacités de discernement. »

Remerciements
(Acknowledgements)

My sincere gratitude goes to each members of this thesis jury: Pr. Aitao LI, Dr. Birgit WILTSCHI and Pr. Narayanaswamy SRINIVASAN, the rapporteurs, Pr. Brigitte GRONDIN-PÉREZ, Pr. Bernard OFFMANN and Dr. Emma SAAVEDRA, the examiners. I thank them for having accepted to evaluate this work and to be members of the jury. Thank you for your patience, understanding and dedication in proofreading and editing this thesis manuscript.

Je remercie le Professeur Frédéric CADET, mon Directeur de thèse de m'avoir donné cette opportunité de réaliser ces travaux de recherche. Je le remercie pour cette confiance qu'il m'a accordée depuis mes années de Licence, sa pédagogie, sa rigueur et sa patience. Je le remercie également pour ses encouragements, ses conseils avisés et son suivi tout au long de ces années de thèse riches en rebondissements ! Cette formidable aventure n'aurait jamais été possible sans la présence du Docteur Cédric DAMOUR, mon Co-Directeur, et du Docteur Philippe CHARTON, mon Co-Encadrant. Je les remercie infiniment de m'avoir guidée durant cette thèse. Leurs différents domaines de compétence et leur savoir-faire ont été les moteurs qui me poussaient à progresser et m'ont permis de grandir et de gagner en maturité. Je leur suis reconnaissante pour leur œil critique, quant aux résultats et aux démarches quelques fois discutables; et leur indulgence, face à mes méconnaissances ou mes erreurs.

Je vous remercie chacun en particulier pour votre présence et votre suivi lors de ces travaux, je n'aurais jamais pu imaginer de meilleurs encadrants que vous !

Mes remerciements au Docteur Fabrice GARDEBIEN, directeur du laboratoire DSIMB, de m'avoir accueillie durant cette thèse. Je le remercie d'avoir tout mis en œuvre pour le bon déroulement de ces travaux et pour m'avoir accordé sa confiance pour effectuer les Travaux Pratiques de Bioinformatique.

Je tiens particulièrement à remercier le Docteur Emma SAAVEDRA et son équipe pour avoir répondu positivement au premier mail et d'avoir accepté de collaborer avec notre équipe lors de ces travaux. *I would particularly like to thank Doctor Emma SAAVEDRA and her team for having answered positively to the first mail and for having accepted to collaborate with our team during this work.*

Je témoigne également toute ma gratitude à ma famille. À mon père et ma mère qui ont toujours tout donné, pour m'offrir le meilleur. Je vous remercie pour cet amour sacrificiel, pour

cette persévérance dont vous avez fait preuve, notamment durant ces trois dernières années. À Ingrid - Bruno et Laurianne - Samuel, vous qui ne vous êtes jamais lassés de m'écouter me plaindre dans les mauvais moments et me réjouir dans les bons. À Nolan, mon rayon de soleil, qui m'a encouragée durant le temps de confinement. Merci à toi aussi mon cher époux, Keran, qui a eu le courage de me supporter, surtout durant les dernières semaines de rédaction. Merci pour ton soutien infaillible et pour tes encouragements. À mes beaux-parents, Thierry et Aline, pour votre présence et votre aide.

Ma gratitude va également à chacun de mes collègues qui m'ont accompagnée le long de ce voyage, avec une attention particulière pour des collègues qui sont devenus des amis, Anamya et Guillaume, je comprends à présent ce que vous avez vécu il y a quelques années.

Je ne pourrais clore ces remerciements sans exprimer toute ma reconnaissance à mes amis, qui ont été là à chaque pas que je faisais dans cette aventure. Et la liste serait trop longue pour tous vous citer.

Merci à vous Emma, Jennifer et Laurine, mes demoiselles à l'écoute de mes histoires infinissables sur mes travaux et d'une grande aide quand j'ai dû gérer à la fois le doctorat et le mariage.

À Teidie, Karine et Jonathan qui m'ont conseillée, m'ont boostée et m'ont remonté les bretelles lors des moments de découragement.

À Vanessa, Jean-Louis, Joas et Romane pour leur présence et leur capacité à redonner le sourire à toute personne qui les côtoie. Je suis fière de vous compter parmi mes amis.

Aux étudiants de l'AEUR qui m'ont vu dans mes bons et mes mauvais jours, merci pour votre compréhension et votre amitié.

Table des matières

Chapitre 1

Introduction

Depuis le premier jour de son existence, l'Homme n'a cessé de se développer en puisant ses ressources dans le monde qui l'entourait. Que cela soit pour se nourrir, se vêtir ou se soigner, ses connaissances se sont décuplées pour donner naissance à plusieurs techniques de production de molécules à diverses finalités : synthèse de molécules à fort potentiel énergétique, de plastique, de tissu et même de médicament. Malheureusement, la majeure partie de ces ressources sont constamment en cours d'épuisement et additionné à la pollution issue de certaines méthodes de production, cela a nécessité la mise au point de nouvelles méthodes de transformation non-polluantes : les biotechnologies. Ces nouvelles technologies se veulent plus respectueuses de l'environnement et sont basées sur l'utilisation de la biomasse renouvelable. Parmi elles, les biotechnologies dites « blanches » consistent à employer des systèmes biologiques pour la fabrication, la transformation, mais également la dégradation de molécules. Ces systèmes biologiques peuvent être d'origine naturelle (bactéries, levures, algues) ou synthétique (système acellulaire), où les traditionnelles réactions chimiques sont remplacées par des réactions biochimiques catalysées par des enzymes. Ce domaine de recherche est en constante évolution et porte déjà ses premiers fruits au niveau industriel. Le champ d'application s'étend de la santé, avec la production de médicaments (Paddon *et al.*, 2013), à l'énergie (Boran *et al.*, 2012) ou encore à la production de biomatériaux (Liu *et al.*, 2013). C'est donc dans cette dynamique de croissance des biotechnologies blanches et d'optimisation des systèmes de synthèse que ces travaux s'inscrivent.

Aussi, l'explosion des données massives numériques, vécue ces dernières années, a stimulé le monde de la recherche pour une adaptation à ce nouvel environnement, notamment en déployant des techniques de pointe *in-silico*. Avec des outils de modélisation de voies métaboliques plus ou moins complexes, une nouvelle porte s'est ouverte, offrant la possibilité : de créer de nouvelles voies de synthèses courtes (Nakamura *et al.*, 2012), d'optimiser celles qui existent déjà (Alkim *et al.*, 2015) ou encore de comprendre les réseaux métaboliques existants chez les microorganismes utilisés en industrie (McCloskey *et al.*, 2013). Il va sans dire que la crise sanitaire de la COVID-19 (Haque, 2020), qui a commencé en mars 2020 et que nous vivons encore aujourd'hui, a mis en

exergue l'intérêt des procédés *in-silico* dans divers domaines, soulignant les nombreux avantages qu'ils offrent, tels que : l'économie de temps, la réduction du coût de développement par le biais des modélisations et de leurs prédictions ou encore l'optimisation plutôt rapide des systèmes déjà mis en place.

Le présent travail de recherche exposé dans cette thèse « Modélisation de voies métaboliques pour la production de molécules » se répartit en 6 chapitres :

- Le **chapitre 1** est une introduction bibliographique du sujet de thèse, qui décrit les notions clés abordées dans ces travaux. Aussi, les progrès réalisés dans le domaine de la modélisation de voies métaboliques, les outils créés ainsi que leurs limites sont présentés. Cette partie expose également les diverses techniques de contrôle existants, nous permettant par la suite de nous focaliser sur les objectifs fixés dans cette étude.

- Le **chapitre 2** propose le développement d'une nouvelle méthode de modélisation des voies métaboliques, dite « boîte-grise » (*grey-box*), fondée sur l'usage d'un modèle cinétique, basé sur les connaissances (« boîte blanche »), et comprenant une partie « boîte-noire » avec l'implémentation d'une équation cinétique contenant un terme d'ajustement. Le modèle boîte-grise est bâti à titre d'exemple pour la voie basse de la glycolyse du parasite *Entamoeba histolytica* (Moreno-Sánchez *et al.*, 2008). Ces premiers résultats permettent d'élargir le champ de possibilités de modélisation des voies, en démontrant l'efficacité de ces modèles hybrides pour prédire le flux final de la voie, notamment lorsque peu de paramètres sont connus pour la voie métabolique étudiée.

- Le **chapitre 3** met en avant la création et la comparaison de modèles de trois types, portant respectivement le nom de boîte-blanche (*white-box*), boîte-noire (*black-box*) et boîte-grise (*grey-box*) décrits auparavant. Ces modèles se différenciant par les données de départ utilisées. Ces modèles sont développés également pour la voie basse de la glycolyse du parasite *Entamoeba histolytica*. Ainsi, à partir de données sur la cinétique des enzymes constituants la voie et de données expérimentales, nous avons montré l'efficacité de ces modèles pour prédire le flux final de la voie métabolique et pour identifier les enzymes-clés jouant un rôle important dans le contrôle du flux. Cette étude comparative des différentes méthodes de modélisation permet de poser les bases de notre étude sur la modélisation de réseaux métaboliques et vient enrichir l'arsenal d'outils de modélisation utiles dans divers domaines.

- Le **chapitre 4** propose de construire cette fois des modèles différents, basés sur l'utilisation de données expérimentales brutes et de techniques de modélisation issues

du Machine Learning (ML). Les algorithmes mis au point décrivent des modèles linéaires ou non-linéaires, appliqués pour trois voies métaboliques différentes : (i) la voie basse de la glycolyse du parasite *E. histolytica*, (ii) la voie de détoxification du peroxyde chez le parasite *Trypanosoma cruzi* (González-Chávez et al., 2015) et (iii) le processus de fermentation de la pénicilline à l'échelle industrielle de *Penicillium chrysogenum* (Goldrick *et al.*, 2015). Ce pan de notre étude fournit les critères de base pour la sélection de la méthode optimale de modélisation à appliquer, selon les données existantes et détenues par un futur utilisateur.

- Le **chapitre 5** est consacré à l'établissement d'un système de rétrocontrôle du flux de la voie métabolique appliqué à l'un de nos meilleurs modèles (modèle boîte-grise). L'enjeu ici étant de pouvoir mettre au point un contrôle capable de réguler automatiquement les entrées de notre système, afin de maintenir le flux à un niveau optimal. Ce contrôle a été instauré sur l'exemple de la partie basse de la glycolyse chez *E. histolytica*. L'implémentation de plusieurs régulateurs PID différents met en lumière le bénéfice de cette méthode dans le rétrocontrôle des voies métaboliques dans une perspective de production de molécules.
- Le **chapitre 6** est constitué d'une conclusion générale des travaux menés dans cette étude, et relate quelques perspectives quant à la mise en place d'un modèle optimisé d'une voie métabolique comprenant un système de contrôle permettant de mimer l'équilibre inhérent au microorganisme dans son état naturel et/ou de maintenir le flux d'une voie de production à un niveau optimal.

1.1. Introduction aux voies métaboliques

1.1.1. Origine et composition des voies métaboliques

Tout comme un orchestre se compose de groupes de musiciens jouant à l'unisson pour interpréter une symphonie, chaque organisme est composé de groupements finement coordonnés qui permettent le maintien d'une harmonie en son sein. Ces groupements prennent le nom de **voie métabolique** et participent à la synthèse d'énergie ou celle de molécules nécessaires au bon fonctionnement de l'organisme dans lequel ils se trouvent. La première démonstration de l'existence de ces voies remonte à l'année 1897, avec le chimiste Eduard Buchner, qui grâce à ses expériences sur la levure, démontra qu'il était possible de réaliser le processus de fermentation alcoolique à partir d'un broyat de levures. Cet extrait qu'il prénomma « zymase » lui valut le Prix Nobel en 1907 et constitua la première preuve que la fermentation opérée dans la levure n'était pas le fruit d'une certaine force obscure, mais bien celle de réactions chimiques réalisées par un agent actif à l'intérieur de la cellule vivante (Buchner and Rapp, 1897; Jaenicke, 2007). Ce n'est que plus tard, qu'il fut démontré que cette « zymase » était en fait un ensemble d'agents actifs qui catalysaient différentes réactions d'une voie métabolique importante appelée : la glycolyse (Cornish-Bowden, 1997).

Ainsi, une voie métabolique peut être définie comme étant une succession de réactions biochimiques catalysées, pour la majeure partie d'entre elles, par des protéines appelées enzymes. Comme le montre la figure 1.1, une voie métabolique est initiée par un substrat de départ qui est transformé en une molécule intermédiaire appelée métabolite, pour aboutir après un certain nombre de réactions à un produit final.

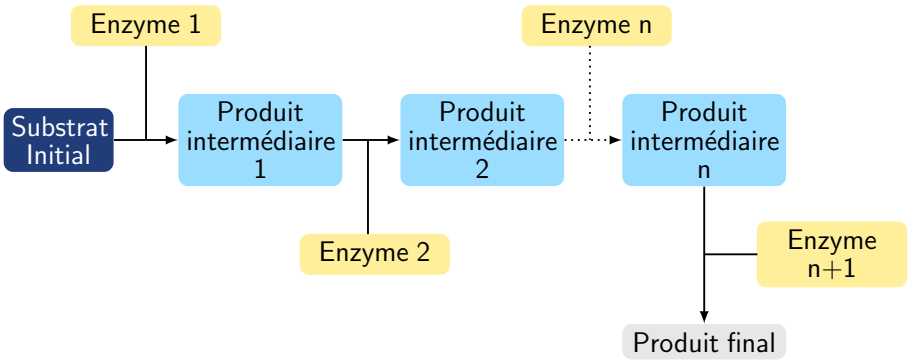

FIGURE 1.1 : *Schéma représentatif d'une voie métabolique.*
Le substrat initial, qui peut être issu de la dégradation de nourriture ou d'une autre voie métabolique, est représenté en couleur bleu foncé. Les intermédiaires métaboliques (métabolites) sont en bleu clair, le nombre

de ces produits peut varier en fonction des voies métaboliques étudiées. Les enzymes réalisent la transformation du substrat initial en produit final (gris). Au cours des réactions, ils peuvent nécessiter la présence de cofacteurs et libérer d'autres molécules (H_2O, CO_2 ou encore de l'énergie sous forme d'ATP par exemple).

Revenons brièvement sur les constituants d'une voie métabolique, en commençant par les substrats. Ces molécules qui servent de base pour la production d'énergie ou de molécules importantes pour l'organisme, peuvent être de nature diverse. Il peut s'agir de sucres plus ou moins complexes (glucose, amidon…), de lipides (acides gras, triglycérides…) et de protéines sous forme de chaînes de résidus d'acides aminés. Ces substrats proviennent soit de l'alimentation de l'organisme étudié (sucre, lipide, protéine), soit d'un sous-produit issu d'une autre voie métabolique (par exemple le pyruvate qui est utilisé comme précurseur de l'acétyl-CoA, le substrat initial du cycle de Krebs). Quelle que soit leur provenance, ces substrats seront spécifiques à l'enzyme qui réalisera sa transformation en métabolite ou produit final. Les métabolites sont également de différents types et résultent de la réaction chimique opérée à partir du substrat. Leur production peut aboutir à la formation : d'énergie, généralement sous forme d'Adénosine Triphosphate (ATP), de dioxyde de carbone (CO2), d'eau (H2O) ou de produits issus de clivage d'une molécule en deux molécules différentes, tels que le glyoxylate et le propionyl-CoA qui sont les résultats de l'action de la malyl-CoA lyase dans l'une des voies d'assimilation de l'acétyl-CoA chez les bactéries (Erb *et al.*, 2010).

Un autre acteur principal de ces voies métaboliques est l'enzyme. Cette molécule de nature protéique sert de catalyseur pour les réactions se déroulant au sein d'un organisme, en venant accélérer la vitesse de la réaction. Là encore, il existe différents types de protéines qui peuvent être classées selon leurs actions au sein des réactions biochimiques (figure 1.2).

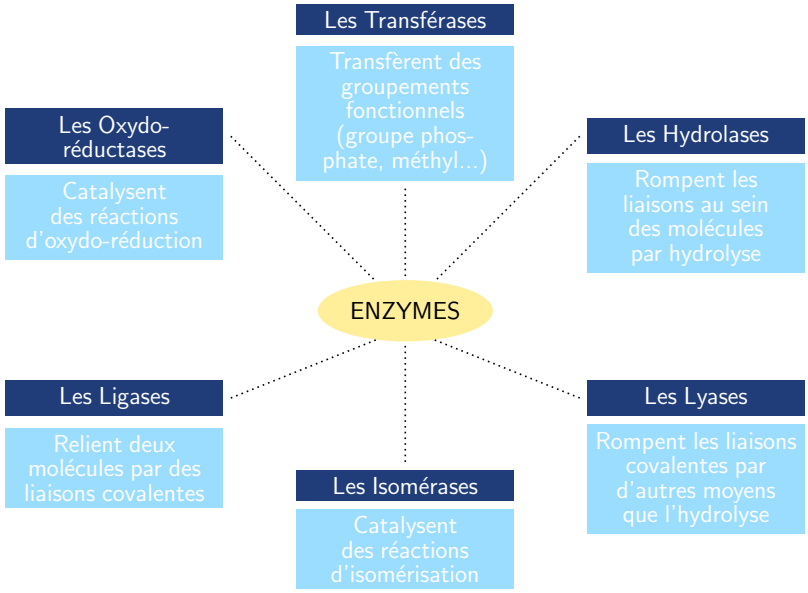

FIGURE 1.2 : *Représentation des différentes classes enzymatiques avec leurs différentes fonctions.*

Parmi ces classes d'enzymes se trouvent, pour n'en citer que quelques-unes : les isomérases, qui convertissent un substrat en l'un de ses isomères ; les lyases, qui rompent des liaisons chimiques présentes dans le substrat créant ainsi deux produits différents ou encore les transférases, qui transfèrent un groupement (phosphate, groupe carboné, azoté…) d'une molécule « donneur de groupement » au substrat.

Aussi, certaines enzymes requièrent la présence de petites molécules pour leur bon fonctionnement : les cofacteurs (figure 1.3). Ces cofacteurs sont généralement classés en deux groupes : les ions métalliques et les coenzymes, qui sont des molécules organiques. Les coenzymes sont soit liés de manière covalente à l'enzyme et portent le nom de groupements prosthétiques, soit liés de manière faible à l'enzyme et portent le nom de cosubstrats.

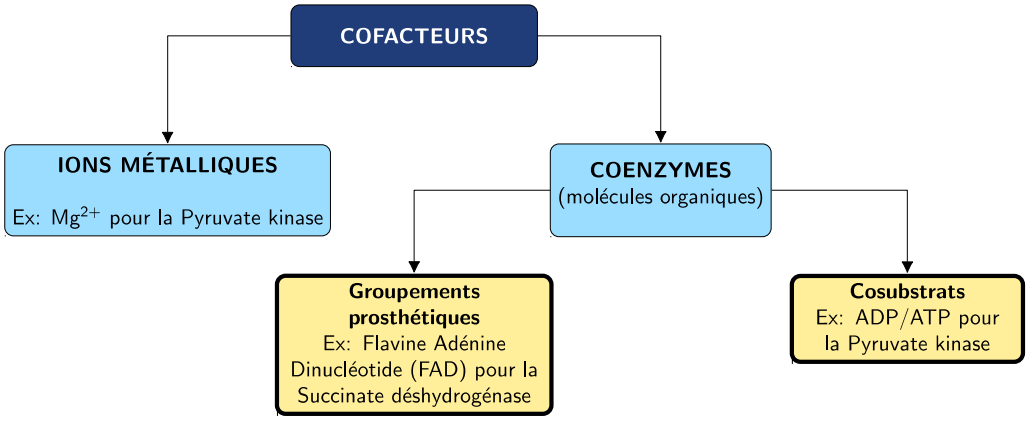

FIGURE 1.3 : *Les différents groupes de cofacteurs existants chez un organisme vivant (Kumar and Barth, 2010; Baranowska et al., 1984; Albracht, 1980).*

Les enzymes peuvent être caractérisées par différents paramètres, qui seront utilisés lors de la modélisation des voies métaboliques plus tard dans cette étude, tels que :

- Sa vitesse (v) et notamment sa vitesse initiale maximale, notée V_{max} ;
- Son affinité avec le substrat, définie par la constante de Michaelis K_m, qui correspond également à la concentration en substrat pour laquelle la vitesse initiale mesurée est égale à la moitié de la vitesse initiale maximale de l'enzyme ;
- Sa constante catalytique ou k_{cat}, qui représente le nombre de moles de substrat transformé par seconde et par mole d'enzyme, quand l'enzyme est saturée en substrat.

D'autres paramètres caractérisant l'enzyme seront rajoutés à la description de la réaction biochimique, selon les molécules participant à cette réaction, comme la constante d'inhibition (K_i), représentant l'affinité de l'enzyme pour un inhibiteur donné.

1.1.2. Exemples de voies de synthèse de molécules

Toute une pléthore de voies métaboliques existe et participe à la synthèse de molécules essentielles à la vie de chaque organisme. Il est important de réaliser que les molécules produites peuvent servir, par exemple, à :

- La production de médicaments : la pénicilline par le champignon *Penicillin chrysogenum* (Nielsen and Jorgensen, 1995) ; la morphine par le pavot somnifère ou *Papaver somniferum* (Onoyovwe *et al.*, 2013) ;
- La production de précurseurs de molécules d'intérêt : le 3-hydroxypropionate, brique de construction pour la production de bioplastique (Jiang *et al.*, 2009); l'isobutanol, pour la production de biocarburant (Wess *et al.*, 2019), par des microorganismes recombinants.

Les molécules produites proviennent de voies métaboliques naturelles (levure, plante…) ou synthétiques générées au sein d'unité de production, telle que des bactéries recombinantes. Ainsi en combinant des tronçons de voies métaboliques différentes et complémentaires, issues d'organismes différents, la synthèse des molécules d'intérêt peut être améliorée (Mukhopadhyay *et al.*, 2008). Une des méthodes d'ingénierie métabolique consiste à identifier les points critiques dans la voie métabolique d'intérêt, à savoir les réactions les plus lentes ou celles produisant des inhibiteurs et/ou des produits dont l'accumulation serait toxique pour l'organisme utilisé pour la production. Puis dans un second temps, à remplacer ces réactions

limitantes par d'autres réactions enzymatiques pouvant provenir d'organismes différents. Cela vient à nouveau mettre en évidence l'importance des techniques de modélisation lors de ces processus de recherche. En effet, le test de ces nouvelles combinaisons en *in-silico* pourrait faire gagner un temps précieux à celui ou celle qui s'investit dans un tel projet de recherche.

Notons que certaines voies de synthèse sont néfastes pour l'Homme, à l'instar de celles des bactéries ou des parasites chez un patient atteint d'une maladie bactérienne ou parasitaire. L'étude de ces voies permettrait l'identification de nouvelles cibles thérapeutiques, participant au développement de nouveaux traitements, utiles notamment lors de cas de résistance aux médicaments déjà en vente sur le marché.

1.2. Techniques de modélisation de voie métabolique

Le développement des nouvelles technologies a permis également la mise en place de techniques de modélisation de voie métabolique. Ces techniques peuvent être classifiées selon leur composition, leur complexité et l'année de leur création (figure 1.4). Cette classification nous permet d'identifier trois groupes de techniques de modélisation. Les modèles basés sur la connaissance, qui ont été parmi les premières méthodes à être développées. Ce sont les méthodes les plus utilisées également pour la modélisation de voie métabolique. Ensuite les modèles basés sur l'utilisation de données expérimentales, qui ont fait leur apparition en biologie plus récemment. Ces modèles ont l'avantage d'être moins compliqués à mettre en place et d'être performants. Et enfin, les modèles hybrides, qui comme leur nom l'indique peuvent contenir à la fois une partie basée sur la connaissance de la voie étudiée et une autre partie basée uniquement sur les données expérimentales acquises. Cette particularité les classe parmi les modèles les plus complexes, mais aussi parmi ceux que l'on cherche à établir encore aujourd'hui.

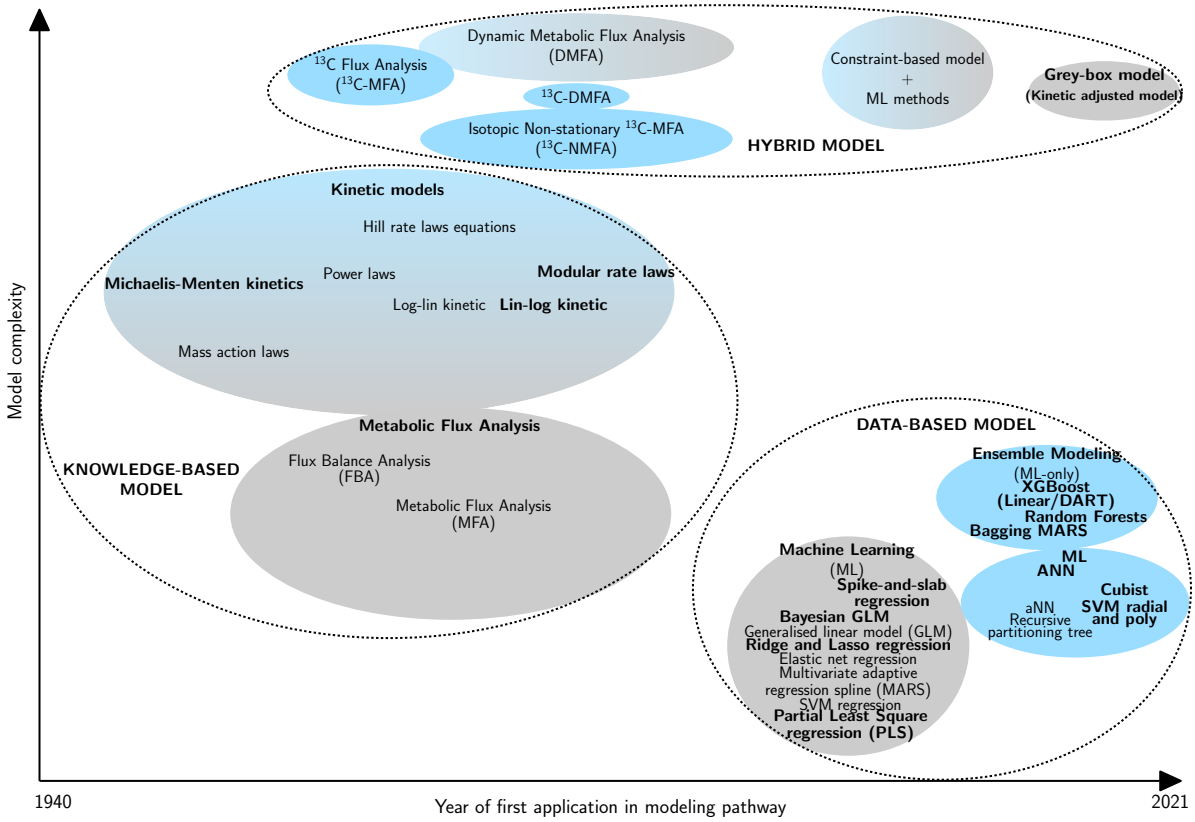

FIGURE 1.4 : *Classification des méthodes de modélisation de voies métaboliques selon leur complexité et l'année de leur première application dans ce domaine.*

La taille de l'ellipse est proportionnelle à la fréquence de la méthode de modélisation dans la littérature. Les méthodes linéaires sont représentées en gris et les méthodes non-linéaires en bleu. Les méthodes en gras sont celles que nous développerons au cours de ces travaux. Le groupe le plus représenté est celui des modèles basés sur les connaissances. Le second groupe est celui des modèles basés sur les données. Et enfin, le dernier groupe, en constante évolution est celui des modèles hybrides.

Nous nous pencherons sur ces différentes techniques de modélisation de voies métaboliques dans chacune des sous-parties qui suivent.

1.2.1. Les modèles basés sur la connaissance

Comme évoqué précédemment, les modèles basés sur la connaissance font partie des premiers modèles de voie métabolique développés en biologie. Ces modèles nécessitent une fine connaissance de la voie étudiée, que ce soit de ses composants (substrat initial, métabolites, produit, enzymes, stœchiométrie), des paramètres enzymatiques (*e.g.* affinité avec le substrat, vitesse de réaction, inhibition) ou encore des équations cinétiques illustrant le mécanisme réactionnel de la catalyse enzymatique (*e.g.* Michaelis-Menten, Ping-Pong, réaction réversible ou non).

Dans cette catégorie, il est possible d'identifier deux groupes différents : les modèles cinétiques et les modèles d'analyse de flux métabolique.

Les modèles cinétiques (« Kinetic models »)

Les modèles cinétiques ou mécanistiques ont eu un impact important dans l'industrie, de sorte qu'ils permettent l'étude dynamique des voies métaboliques, avec les phénomènes d'oscillations pour les voies cycliques ou les transitions d'état physiologique (Garde *et al.*, 2020; Fouchard *et al.*, 2009). Ils contiennent également des informations quant à la régulation de la voie, ce qui est primordial lorsque l'on souhaite comprendre pourquoi l'optimum calculé n'est pas atteint en pratique et où se trouvent les limites du système afin de les surpasser (Wiechert and Noack, 2011). Ce type de modèle intègre l'expression des vitesses, les paramètres enzymatiques, les concentrations initiales en substrats et cosubstrats et les processus de régulation (inhibition ou activation). L'ensemble de ces données font généralement l'objet de recherches bibliographiques (littérature ou bases de données) ou sont les résultats de manipulations expérimentales. Plusieurs types de modèles cinétiques sont définis dans la littérature actuelle et se répartissent en fonction de l'expression mathématique décrivant le comportement des interactions d'une voie sous forme de vitesse de changement, plus couramment appelée **équation cinétique**. Ces équations cinétiques

sont plus ou moins complexes et leur contenu dépend tant du degré de détails des réactions enzymatiques que du champ d'application du modèle créé. Une distinction peut être faite entre les équations cinétiques dérivées de formulations déterministes ou stochastiques, ces dernières ne seront pas développées ici. Les enzymes présentent le plus souvent un comportement déterministe, et il est possible de les décrire par des équations cinétiques mécanistiques (équation de Michaelis-Menten, équation de Hill, loi d'action de masse) ou approximatives (cinétiques lin-log/log-lin, lois de vitesse modulaire, loi de puissance).

La cinétique de Michaelis-Menten

L'un des premiers modèles cinétiques ayant été développés est celui qui intègre les équations de Michaelis-Menten (Michaelis and Menten, 1913; Chance, 1943; Garfinkel *et al.*, 1970). Ce type d'équation est communément utilisé lorsque le substrat est en excès et s'écrit sous la forme suivante :

$$v = \frac{V_{max}[S]}{K_m + [S]}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_{max}}$ est la vitesse initiale maximale de l'enzyme ; $\boldsymbol{K_m}$ est la constante de Michaelis, illustrant l'affinité de l'enzyme pour son substrat, et $[\boldsymbol{S}]$ désigne la concentration en substrat.

À cette équation peuvent venir s'ajouter des inhibiteurs ou des activateurs, bien entendu lorsque ceux-ci sont connus ainsi que leurs paramètres. La glycolyse dans les érythrocytes construit par l'équipe de TA. Rapoport *et al.* (Rapoport *et al.*, 1974) est le premier modèle mathématique développé et utilisant ce type d'équation pour décrire la cinétique de certaines enzymes, telles que l'hexokinase, bisphophoglycérate mutase. Cette étude a permis de mettre en évidence le rôle prépondérant de l'hexokinase et de la phosphofructokinase dans le contrôle du flux de cette voie.

La loi d'action de masse

Déterminée tout d'abord pour décrire les réactions chimiques (Guldberg and Waage, 1867), la loi d'action de masse peut s'avérer utile lors de la modélisation de réactions biochimiques (Shapiro and Shapley, 1965). Cette loi permet de faire abstraction de certains paramètres cinétiques inhérents à l'enzyme pour se concentrer uniquement sur la vitesse de la réaction et les concentrations des métabolites présents. Ainsi elle peut s'écrire sous cette forme :

$$v = k\prod [S_i]^n$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{k}$ représente la constante de vitesse de la réaction ; $[\boldsymbol{S_i}]$ est la(les) concentration(s) en substrat(s) et $\boldsymbol{n}$ est leur valeur stœchiométrique.

L'utilisation de ce type d'équation suppose que le système atteint un état d'équilibre, à une température donnée, où la vitesse de consommation du substrat est égale à celle de la formation du produit.

Les lois de vitesse de Hill (« Hill rate laws »)

Les équations de vitesse de Hill sont également retrouvées dans le groupe des expressions déterministes de la vitesse de la réaction enzymatique. Cette équation est utilisée lorsqu'une régulation est impliquée dans le système biologique représenté par l'enzyme. Elle a été écrite initialement pour décrire le phénomène de la coopérativité lors de la fixation des molécules d'oxygène sur l'hémoglobine (Barcroft and Hill, 1910). Ainsi, pour une enzyme allostérique, la vitesse de la réaction peut s'écrire de la manière suivante :

$$v = \frac{V_{max}[S]^{n_H}}{K_{0,5}^{n_H} + [S]^{n_H}}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_{max}}$ est la vitesse initiale maximale de l'enzyme ; $\boldsymbol{K}_{0,5}$ est la constante de demi-saturation, correspondant à la concentration en substrat pour laquelle la vitesse est égale à la moitié de V_{max} ; $[\boldsymbol{S}]$ désigne la concentration en substrat et $\boldsymbol{n_H}$ désigne le coefficient de Hill ou le coefficient de coopérativité, qui mesure la coopérativité de la liaison entre le substrat et l'enzyme.

Plusieurs études ont montré le bénéfice d'utiliser une telle équation pour décrire certaines enzymes, par exemple celle portant sur la modélisation du métabolisme des purines chez l'Homme (Curto *et al.*, 1998). D'autres travaux ont montré son efficacité pour modéliser des réseaux de régulation de gènes, tels que ceux impliqués dans la réparation de l'ADN (DiStefano, 2013).

En comparaison avec les modèles cinétiques précédents, d'autres modèles utilisent des expressions dites approximatives. Expressions qui se veulent plus simples et qui utilisent moins de paramètres, voyons ensemble quelques-unes de ces expressions.

La loi de puissance (« Power-laws »)

Le modèle développé dans ce paragraphe utilise une équation de vitesse similaire à celle de la loi d'action de masse, qui est appelée la loi de puissance. Et elle peut contenir non seulement les concentrations en substrat, mais aussi celles des produits (Savageau, 1970, 1988).

Une fois encore, ce type d'équation permet de faciliter la modélisation de certaines réactions biochimiques, lorsque le phénomène à modéliser devient trop complexe ou lorsque les paramètres cinétiques des enzymes ne sont pas déterminés.

La cinétique « Lin-Log » ou « Log-Lin »

Les lois cinétiques « Lin-Log » ou « Log-Lin » sont des représentations linéaires de logarithmes pouvant être utilisées lorsque d'importantes concentrations sont présentes au sein de la réaction enzymatique. L'avantage d'utiliser une telle équation réside dans le fait qu'elle prend en compte les changements dans les concentrations d'enzyme. L'expression de la vitesse de réaction prend cette forme-ci pour la cinétique « Lin-Log » (Heijnen, 2005; Alves *et al.*, 2008), pour une enzyme donnée catalysant une réaction avec un nombre $\boldsymbol{i}$ d'espèces $\boldsymbol{M}$ (substrats et produits) :

$$v = \frac{e}{e_0} J_0 \left(1 + \sum_{i=1}^{n} \varepsilon_i^0 log \frac{[M]_i}{[M]_i^0} \right)$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{e/e_0}$ est l'activité enzymatique par rapport à celle à l'état d'équilibre de référence ; $\boldsymbol{J_0}$ est le flux à l'état d'équilibre ; $\boldsymbol{\varepsilon_i^0}$ désigne le coefficient d'élasticité, qui quantifie le degré auquel les espèces de la réaction modifient la vitesse de la réaction ; $[\boldsymbol{M}]_{\boldsymbol{i}}$ est la concentration de l'espèce $\boldsymbol{i}$ de la réaction et $[\boldsymbol{M}]_{\boldsymbol{i}}^{\boldsymbol{0}}$ désigne cette même concentration à l'état d'équilibre de référence.

Un exemple de modèle se basant sur cette cinétique a été développé pour la voie de la glycolyse chez *Escherichia coli* montrant des résultats comparables à un modèle cinétique basé sur l'utilisation d'équations Michaelis-Menten (Tušek and Kurtanjek, 2010). Un autre modèle, construit pour la même voie cette fois chez la bactérie *Lactococcus lactis*, n'a pas donné de résultats satisfaisants par rapport aux modèles existants déjà (del Rosario *et al.*, 2008).

Pour la cinétique « Log-Lin » sa formulation est similaire à celle de « Lin-Log », à la différence que, cette fois, la fonction logarithme est également appliquée sur les effecteurs de la réaction à savoir les enzymes (Hatzimanikatis, 1997).

Les lois de vitesse modulaire (« Modular rate laws »)

Le comportement des enzymes au sein des modèles cinétiques peut être illustré par les lois de vitesse modulaire (Liebermeister *et al.*, 2010). Ces lois font partie des cinétiques approximatives et sont généralement sous la forme standard :

$$v = T \frac{E_0 f_r}{D + D^{reg}}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{T}$ est un terme de paramétrage stœchiométrique contenant les constantes catalytiques ; $\boldsymbol{E}_0$ est la concentration initiale de l'enzyme ; $\boldsymbol{f}_r$ désigne une régulation complète ou partielle ; $\boldsymbol{D}$ est un terme dénominateur pour chaque loi de vitesse et $\boldsymbol{D}^{reg}$ désigne un terme de régulation particulier.

Ces lois de vitesse modulaire présentent l'avantage d'impliquer des phénomènes de régulation, ce qui peut s'avérer utile lorsque nous faisons face à des régulations particulières dans la voie métabolique, comme la régulation allostérique.

Les modèles d'analyse de flux métabolique (« Metabolic Flux Analysis »)

Le second groupe identifié parmi les modèles basés sur la connaissance est le modèle de type analyse de flux métabolique. Dans cette famille, deux modèles sont remarquables : l'analyse de l'équilibre des flux (« Flux Balance Analysis » ou FBA) et l'analyse des flux métaboliques (« Metabolic Flux Analysis » ou MFA). Cette famille de modèles se consacre à l'analyse et la modélisation des flux d'une voie métabolique et présente l'avantage de n'avoir besoin que des stœchiométries de chaque réaction pour le faire.

L'analyse de l'équilibre des flux

La construction de ce premier type de modèle permet le calcul du flux de métabolites au travers du réseau métabolique, en rendant possible la prédiction, par exemple, du taux de croissance d'un organisme ou de la vitesse de production d'un métabolite important à haute valeur ajoutée. Dans ce modèle les réactions métaboliques sont représentées dans un tableau, sous forme d'une matrice numérique, où seront retrouvés les coefficients stœchiométriques de chaque réaction. Ces coefficients imposent des contraintes sur le flux des métabolites, qui permettent la mise en place d'équations équilibrant les entrées et les sorties de la voie et qui posent les limites du système (Orth *et al.*, 2010). Suite à cela, un problème d'optimisation est posé, avec une fonction objective spécifique, restreinte par les équations d'équilibre de masse et les différentes contraintes ajoutées au modèle. À cause de la faible quantité de connaissances qu'ils utilisent, cette méthode permet la modélisation d'importantes voies métaboliques, contenant d'innombrables réactions biochimiques. Aussi, elle permet d'avoir une bonne estimation du comportement de l'organisme dans des conditions données, de prédire des réactions manquantes (les reconstructions de voies métaboliques étant souvent incomplètes) et d'améliorer les constructions dans des organismes recombinants en simulant leur comportant pour ensuite l'optimiser.

Un modèle de ce type a été développé pour la synthèse des graisses dans le tissu adipeux, contenant près de 57 réactions (Fell and Small, 1986). Les résultats de cette modélisation ont mis en

évidence l'intérêt de faire apparaître les équilibres entre la synthèse et la dégradation de chaque intermédiaire ainsi que des cofacteurs présents, afin d'améliorer la prédiction du flux final.

L'analyse des flux métaboliques

Pour ce deuxième type de méthode, aucune concentration ni aucun paramètre cinétique ne sont nécessaires à l'élaboration du modèle de la voie métabolique étudiée. Cette approche vise également à déterminer les valeurs des flux d'un réseau métabolique, à partir de l'hypothèse de base selon laquelle les métabolites internes ne s'accumulent pas dans la cellule, et peut se servir de données de mesures de flux. Ce modèle implique de résoudre un système d'équations linéaires, où les équations sont des bilans de masse des métabolites et les inconnus sont les flux, soumis à des contraintes d'égalité ou inégalité qui tiennent compte entre autres de la non négativité de certains flux (Stephanopoulos, 1999; Antoniewicz, 2015). Plusieurs variantes de cette méthode existent et se distinguent notamment par la présence et l'identité de l'isotope utilisé pour mesurer le flux ou par le fait que l'état d'équilibre est présumé ou non.

Une étude basée sur l'utilisation d'un tel modèle a mis en évidence la présence d'un changement métabolique dans une culture cellulaire utilisée pour la production de molécules d'intérêt, avec une redirection du flux de la voie étudiée vers une autre voie métabolique. Ces résultats servent de boussole aux chercheurs, qui peuvent alors cibler à nouveau leur recherche sur cette partie de la voie, leur permettant par la suite : i) d'identifier des failles dans la voie d'intérêt et ii) d'adapter les conditions de culture pour améliorer la synthèse de la protéine (Sengupta *et al.*, 2011). D'autres recherches ont été menées sur la voie de fermentation de xylose en éthanol par la levure *Candida shehatae* et ont permis de déterminer le rendement théorique maximal de la voie étudiée ce qui ouvre la voie vers l'optimisation du procédé de production d'éthanol (Bideaux *et al.*, 2016).

Un type de modèle basé sur la connaissance que nous ne développerons pas dans ces travaux, mais qu'il est intéressant de citer, est le modèle à l'échelle génomique ou « genome-scale metabolic models » (GEM). En effet, les modèles de type réseau métabolique peuvent servir de base pour l'intégration d'autres données omiques. Ainsi, ils aboutissent à l'obtention d'une carte métabolique complète d'un organisme qui contient à la fois l'ensemble de ses réactions métaboliques et ses données métabolomiques et métagénomiques (Chong and Xia, 2017). Un modèle de ce type a été développé pour la souris afin de prédire des phénotypes particuliers : simulation de souris « knock-out », autrement dit déficientes pour un gène donné, et évaluation de la délétion sur le métabolisme murin. L'une des déficiences modélisées est celle de la lipoprotéine lipase qui est responsable de l'hydrolyse des triglycérides contenus dans les chylomicrons et les lipoprotéines de

très basse densité ou « Very Low Density Proteins » (VLDL). Les souris ayant cette mutation présentent un taux élevé de triglycérides et de VLDL et souffrent de douleurs abdominales récurrentes. Les résultats de la simulation sont cohérents avec les données expérimentales, puisqu'ils ont montré une augmentation du flux vers la synthèse de triglycérides, ainsi qu'une baisse du flux de dégradation de ces mêmes triglycérides (Sigurdsson *et al.*, 2010).

1.2.2. Les modèles basés sur l'utilisation de données expérimentales

La deuxième famille de modèles *in-silico* développés pour la représentation des voies métaboliques est : le modèle basé sur les données expérimentales (« Data-based model »). Nous y retrouvons toutes les méthodes d'**apprentissage automatique** ou « Machine Learning » (ML). Ces techniques se basent sur l'utilisation de données expérimentales issues de l'étude de la voie métabolique d'intérêt et peuvent être classées selon leur capacité à appréhender leur linéarité ou leur non-linéarité. Une distinction de plus a été représentée sur la figure 1.4, en prenant en compte la combinaison de techniques de ML : l'**apprentissage d'ensemble** (« Ensemble Learning »). Cela permet l'utilisation de multiples algorithmes d'apprentissage dans le but d'obtenir une meilleure performance de prédiction, comparée à celle obtenue avec n'importe quel algorithme constituant la méthode d'apprentissage d'ensemble. Il existe en apprentissage automatique deux types de problèmes : les problèmes de régression et les problèmes de classification. Bien que les deux problématiques puissent être étudiées lors de l'étude d'une voie métabolique, nous nous plaçons dans l'étude présente dans l'optique de modéliser une voie de production où la quantification de certaines variables du système est cruciale. Nous aborderons donc la thématique des techniques d'apprentissage automatique en nous basant uniquement sur le volet des problèmes de régression. L'objectif général de ce type d'étude est de prédire une variable quantitative Y (par exemple : la concentration du produit final ou d'autres métabolites, le flux en sortie de la voie étudiée) à l'aide plusieurs variables explicatives X_1, $X_2 \ldots X_n$ (par exemple : la concentration en substrats/cosubstrats/inhibiteurs/activateurs, les paramètres mesurés influençant la production, la concentration en enzymes). Autrement dit, la prédiction se fait par le biais de la recherche d'une fonction linéaire ou non-linéaire de telle sorte que $Y \approx f(X_1,\ X_2 \ldots X_n)$.

Les techniques de ML sont très nombreuses et performantes pour résoudre des problèmes d'ordre biologique. Nous pouvons citer les travaux de I. Shaked *et al.* qui ont mis au point un modèle de machine à vecteurs de support (« Support Vector Machines » ou SVM) pour prédire les effets secondaires éventuels des médicaments agissant sur le métabolisme. Les résultats sont

prometteurs, puisque l'étude a pu identifier de potentiels biomarqueurs, marquant la présence d'états pathologiques liés aux effets secondaires (Shaked *et al.*, 2016). Voyons quelques-unes de ces méthodes et leur principe, plus précisément trois d'entre elles qui feront l'objet de notre étude au **chapitre 3** : les réseaux de neurones artificiels (« Artificial Neural Network » ou ANN), les forêts aléatoires (« Random Forest » ou RF) et une technique d'apprentissage d'ensemble, XGBoost (« Extreme Gradient Boosting »).

Les réseaux de neurones artificiels

Le principe de fonctionnement de ces réseaux de neurones artificiels se base sur celui du réseau neuronal retrouvé dans le cerveau. Ainsi, tout comme notre cerveau va traiter les informations qui nous entourent pour y répondre de manière adaptée ; le réseau de neurones artificiels va traiter l'ensemble des informations qui lui sont données pour ensuite répondre à une problématique posée. Dans notre cas, il s'agira de prédire une variable de notre voie de production à partir d'autres données contenues dans cette voie métabolique. L'un des considérables avantages de cette technique réside dans le fait qu'il va apprendre à résoudre les problèmes que nous lui soumettons à partir des collections des données en entrée, et non pas à partir de règles déjà préétablies (Jain *et al.*, 1996). Aussi, il est capable de prendre en charge des données massives et contenant un grand nombre de variables. Plusieurs de ces modèles ont été construits dans différents domaines d'application :

- En physique : pour effectuer de l'analyse de données dans le domaine de la physique des particules et permettre la classification d'évènements, l'identification de particule (Kolanoski, 1995) ;
- En science de l'environnement : pour prédire le taux de pollution de l'air dans l'environnement (Pawul and Śliwka, 2016) ou encore pour réaliser des prévisions du rendement du traitement des eaux usées (El-Din *et al.*, 2004) ;
- En microbiologie : afin d'identifier des microorganismes en fonction de spectres de masse, ou pour évaluer l'effet de composés sur l'inhibition de bactérie (Basheer and Hajmeer, 2000) ;
- Ou encore en biologie : pour déterminer la composition d'un médicament (Arabzadeh *et al.*, 2019) ou prédire la perméabilité de la peau ou de la barrière hémato-encéphalique lors de l'administration de médicaments (Sutariya *et al.*, 2013).

Le réseau de neurones artificiels est formé à sa base de neurone (appelé également unité), dont le premier modèle fut le perceptron (Basheer and Hajmeer, 2000). Le schéma général de ce neurone

dit formel se trouve en figure 1.5. Ce dernier est comparable au neurone biologique. En effet, il est capable de :

- Recevoir des informations, sous la forme de données expérimentales implémentées dans le modèle ;
- Accorder plus ou moins d'importance aux informations qu'il reçoit en pondérant la valeur d'entrée ;
- Traiter l'information au sein du neurone en réalisant la somme des entrées qu'il a reçue et ;
- Appliquer une fonction d'activation sur la somme pondérée précédente avant de transmettre son résultat au neurone suivant, aboutissant au calcul final de la variable à prédire.

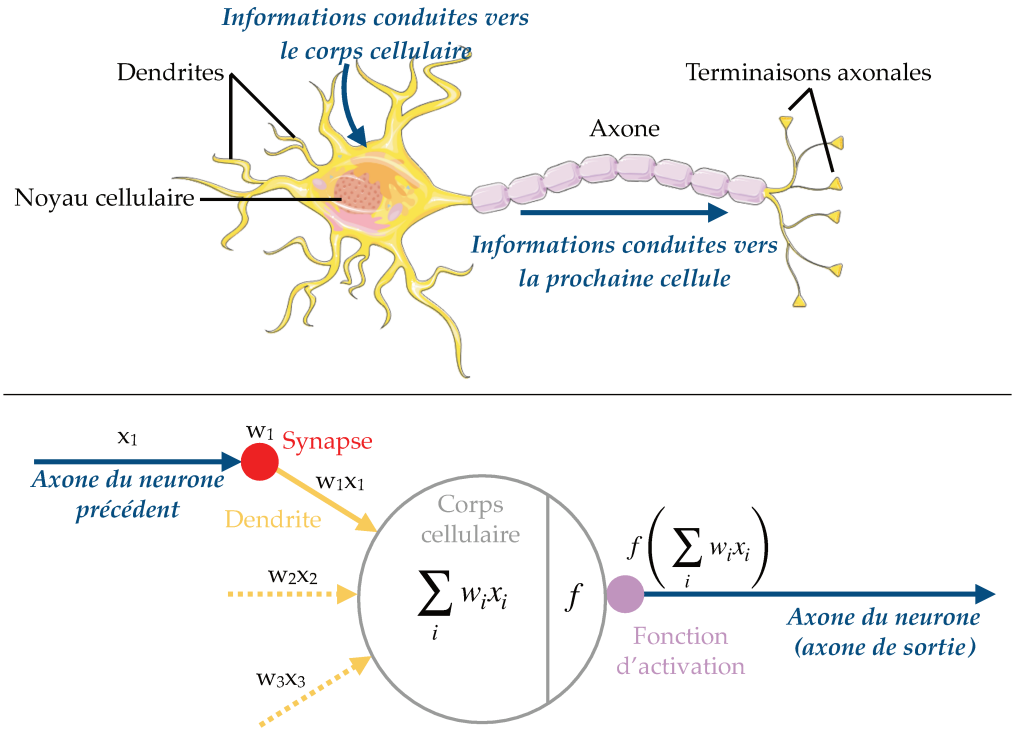

FIGURE 1.5 : *Représentation d'un neurone biologique (panel du haut) comparée à celle d'un neurone formel (panel du bas).*
Tout comme le neurone biologique, le neurone formel reçoit des signaux (x_1) des autres neurones par ses dendrites. Les dendrites transforment ce signal en pondérant l'information reçue (w_1), puis le neurone traite ces informations au niveau de son hypothétique corps cellulaire en leur appliquant des modifications avant de les relayer au neurone suivant.

Un réseau de neurones est composé de plusieurs couches de neurones, comme nous le montre la figure 1.6 :

- La couche d'entrée (« Input layer ») : sert à relayer les données d'entrée aux neurones dans la couche cachée, elle attribue un poids à chaque connexion qu'elle va effectuer avec le neurone suivant.
- La couche cachée (« Hidden layer ») : effectue la somme pondérée des données qu'elle reçoit avant d'y appliquer une fonction d'activation, qui permet de convertir les données d'entrée en sortie et décide si le neurone est activé ou non. Si ce n'est pas le cas, les valeurs en sortie de ce neurone ne seront pas prises en compte dans le résultat final. Plusieurs fonctions d'activation existent, par exemple la fonction logistique (log) ou la tangente hyperbolique (tanh).
- La couche de sortie (« Output layer ») : qui peut aussi être composée de plusieurs neurones en sortie, si le problème posé demande à ce que le modèle prédise plusieurs valeurs en sortie.

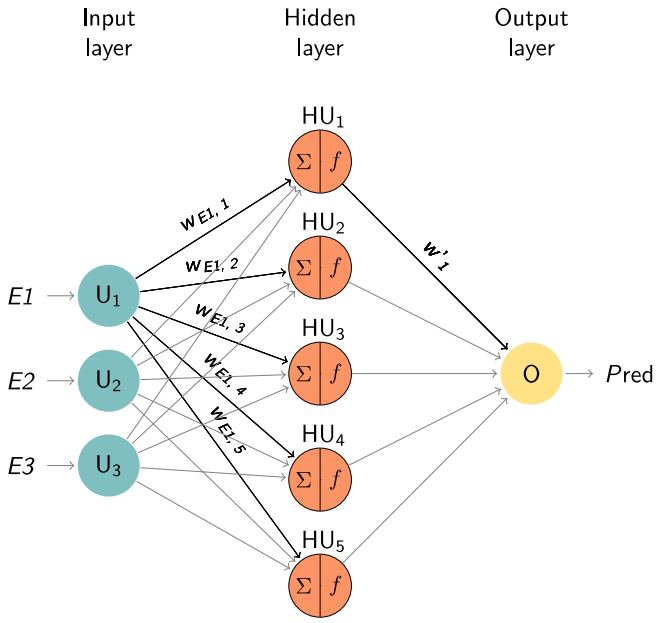

FIGURE 1.6 : *Représentation de l'architecture la plus simple d'un réseau de neurones artificiels. Ce réseau est composé de 3 couches contenant chacune un nombre défini de neurones. La couche d'entrée contient 3 unités auxquelles sont affectées des variables présentes dans notre voie métabolique (activité enzymatique, concentrations en substrats/métabolites/cosubstrats), les valeurs sont pondérées ($w_{E1,1}$). La couche cachée est composée de 5 neurones cachés et opère en son sein la somme pondérée des valeurs d'entrée (Σ) et y applique la fonction d'activation (f). La sortie de la couche cachée est aussi pondérée (w'_1) avant d'être transmise à la couche de sortie qui calcule la valeur finale de la prédiction (concentration en produit final ou flux final de la voie).*

Ces modèles peuvent avoir différentes architectures et selon les connexions effectuées entre les neurones, ils peuvent être classifiés en deux catégories : les réseaux à propagation avant ou *feed-forward*, dans lesquels les signaux ne se propagent que dans un seul sens, et les réseaux récurrents, dans lesquels les signaux peuvent revenir en arrière et alimenter à nouveau la couche précédente ou la même couche.

Les modèles de forêts aléatoires

Les forêts d'arbres décisionnels ou plus simplement les forêts aléatoires se placent dans les méthodes d'apprentissage d'ensemble. En effet, elles se composent d'un grand nombre d'arbres décisionnels dont la structure est présentée en figure 1.7. Ces arbres sont un ensemble de règles de décision qui possèdent une structure arborescente composée de :

- Nœud racine par lequel va transiter l'ensemble de données en entrée ;
- Nœud interne qui va représenter certaines caractéristiques de l'ensemble de données, prenant le nom d'attribut ;
- Branche désigne les règles de décisions qui sont prises au cours de la modélisation et relient les nœuds entre eux ;
- Nœud feuille est, quant à lui, le résultat de l'ensemble des décisions et ne contient pas d'autres branches.

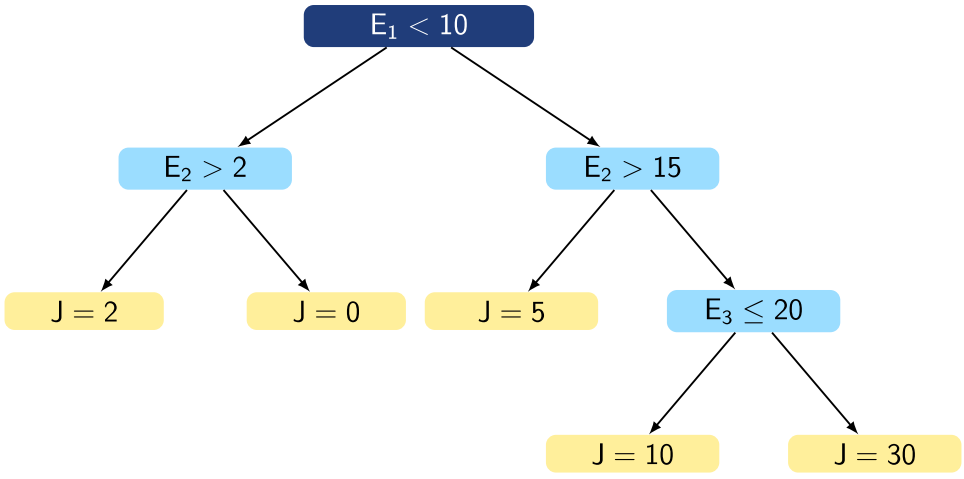

FIGURE 1.7 : *Schéma d'un arbre décisionnel.*
La racine est en bleu foncé, les nœuds internes sont en bleu clair et les feuilles sont en jaune. Les flèches représentent les branches de l'arbre décisionnel. E_i désigne la concentration d'enzyme de la voie métabolique étudiée, J est le flux final de la voie et également la variable à prédire.

Un arbre décisionnel se parcourt donc de la « racine » vers les « feuilles » et présente l'avantage d'être rapide, facilement interprétable et d'être performant pour de grands jeux de données. Le partitionnement des nœuds se fait selon la variable qui réalise le meilleur partage de l'ensemble des données. Ensuite, une succession de coupure de nœuds est réalisée en respectant des règles d'arrêt et des conditions de coupure. Les conditions de coupure, elles, reposent sur des critères mathématiques, comme nous le montre la figure 1.7. L'exemple de cette figure illustre un modèle d'arbre décisionnel conçu pour un problème de régression, où la variable à prédire est le flux en sortie de la voie métabolique et les prédicteurs sont les concentrations enzymatiques des catalyseurs présents dans cette même voie.

Aussi, lorsque l'arbre est construit, le nombre de feuilles peut parfois être trop important, nous procédons alors à l'élagage de l'arbre jusqu'à l'obtention d'un équilibre entre la complexité de l'arbre et la précision de la prédiction. Le choix de l'arbre final peut se faire à l'aide d'une méthode de validation croisée testant différentes versions élaguées de l'arbre construit. Une fois la construction de l'arbre validée, la valeur de la variable à prédire (Y) peut être déterminée et correspond à la moyenne des valeurs de Y associées aux individus présents dans la feuille.

La méthode de forêts aléatoires a été développée en 2001, dans l'optique d'accroître la performance des arbres décisionnels seuls (Breiman, 2001). Elle est composée d'une combinaison d'arbres décisionnels dont nous avons vu la croissance dans le paragraphe précédent. Cette configuration nous permet alors de prendre en compte, non plus les résultats d'un seul arbre décisionnel, mais le résultat d'un ensemble d'arbres décisionnels différents (figure 1.8). Plusieurs méthodes de combinaison de ces arbres existent, dont deux en particulier : le *Bagging* et le *Bootstrapping* (Liaw and Wiener, 2002).

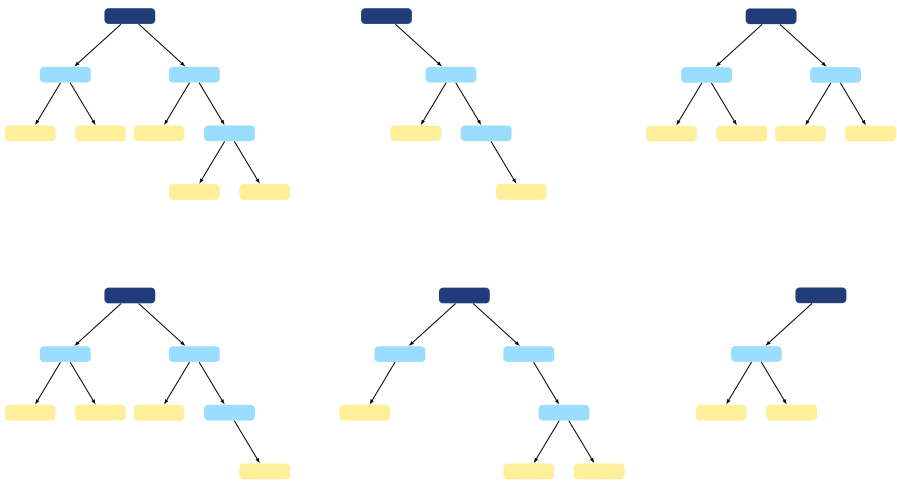

FIGURE 1.8 : *Représentation d'un modèle de forêts aléatoires.*
Le modèle illustré est composé de six arbres décisionnels avec une architecture différente.

Les forêts aléatoires ont montré leur efficacité pour la prédiction de variable biologique, telle que la réponse aux médicaments, en utilisant des données contenant la sensibilité aux médicaments et l'expression basale des gènes de la lignée cellulaire étudiée (Riddick *et al.*, 2011). D'autres études ont montré que ce type de modèle pouvait prédire avec précision le rendement de fruits, selon l'irrigation de la culture ou les émissions de gaz à effet de serre majeur produit par les activités agricoles (Fukuda *et al.*, 2013; Philibert *et al.*, 2013).

L'algorithme XGBoost

Le dernier modèle abordé au sein de cette partie est celui de XGBoost (« *Extreme Gradient Boosting* »). Il fait aussi partie des méthodes d'apprentissage d'ensemble référencées dans la figure 1.4. Cet algorithme est basé sur le modèle des arbres de régression auxquels une implémentation de *boosting* a été ajoutée (Chen and Guestrin, 2016). Ainsi, le modèle va créer un premier arbre de régression qu'il va évaluer. L'évaluation faite, le modèle va appliquer une pondération à chaque donnée selon la prédiction qui a été faite par le premier modèle, avant de construire le second arbre de régression et ainsi de suite. Le résultat de prédiction correspondra alors à l'agrégation de l'ensemble des arbres construits au cours du processus, qui de ce fait sera un peu plus long, car il se fait de manière séquentielle. Les avantages de cette méthode sont sa vitesse accrue par rapport à d'autres algorithmes se basant sur l'utilisation d'arbre de décision, sa capacité à restituer l'importance des variables présentes dans l'ensemble de données et sa performance à traiter des données qui ne sont pas linéaires.

L'efficacité de cette méthode a fait d'elle un modèle incontournable, utilisé dans plusieurs domaines de recherche. En chimie de l'environnement, par exemple, XGBoost a été le meilleur modèle pour prédire la concentration de matière particulaire (< 2,5 μm) présente dans une ville, à partir de données satellites et météorologiques (Zamani Joharestani *et al.*, 2019). Couplé aux nombreuses données chimiques et biochimiques, XGBoost se révèle être parmi les meilleurs modèles dans le domaine de la recherche pharmaceutique, par exemple pour prédire l'activité biologique de molécules avec l'aide de données structurales de ces composés (Babajide Mustapha and Saeed, 2016). Une autre étude plus récente, dans le domaine de la santé, a également utilisé l'algorithme XGBoost pour établir un modèle de pronostic du COVID-19 permettant la prédiction du risque de mortalité et la distinction entre les cas critiques et les cas graves (Yan *et al.*, 2020).

1.2.3. Les modèles hybrides

Enfin, au sein des méthodes utilisées pour la modélisation de voies métaboliques, nous retrouvons les modèles hybrides (figure 1.4). Ces modèles hybrides sont nombreux et sont issus de la combinaison de modèles de nature différente ou de la combinaison entre un modèle *in-silico* couplé à une expérience *in-vitro*.

Modèles dérivés des méthodes d'analyse de flux métabolique

Dans cette dernière catégorie, nous retrouvons les modèles dérivés des méthodes d'analyse du flux métabolique. Parmi ces techniques, il existe celles qui utilisent des traceurs isotopiques stables (^{13}C) pour récolter de nouvelles données et les « injecter » par la suite au sein du modèle d'analyse de flux métabolique. En effet, la manipulation expérimentale préalable consiste à ajouter l'isotope stable dans une culture cellulaire pour un temps donné (>3 h), ce qui aboutit à l'incorporation de l'isotope dans les intermédiaires métaboliques et les molécules produites. Si le traceur isotopique est bien choisi par le manipulateur, les distributions de cet isotope dans la cellule dépendront fortement des valeurs des flux métaboliques intracellulaires. Par conséquent, les mesures du marquage au ^{13}C, par exemple, pourront être utilisées comme des contraintes supplémentaires pour estimer les flux de la voie métabolique étudiée, dans un modèle de type analyse de flux métabolique. La prédiction du flux résulte alors, non plus de la stœchiométrie seule de la voie, mais également de données expérimentales nouvellement produites. Le traçage de l'isotope stable dans la voie, se fait par le biais de techniques d'analyse spectrale, comme la spectrométrie de masse ou l'analyse HPLC (« *High Performance Liquid Chromatography* ») ou chromatographie en phase liquide à haute performance (Antoniewicz *et al.*, 2007; Antoniewicz, 2015). Cette méthode porte le nom de ^{13}C–MFA (« *Metabolic Flux Analysis* ») lorsque l'état d'équilibre métabolique est supposé et ^{13}C–DMFA (« *Dynamic Metabolic Flux Analysis* »), si ce n'est pas le cas. Une application de cette technique de modélisation a été faite avec un grand jeu de données issu de la voie de production de lysine chez *Corynebacterium glutamicum* (Wiechert *et al.*, 1997). L'estimation des flux d'un tel réseau métabolique complexe a été faite avec succès par cette technique, utilisant comme donnée de départ une immense quantité de données expérimentales mesurées de façon très précise. Il est intéressant de noter que, dans cette étude, les hypothèses sur leur stœchiométrie du métabolisme énergétique n'étaient pas nécessaires pour la détermination des flux ; ces paramètres pouvant même être estimés par les résultats de prédiction.

Le temps nécessaire pour atteindre l'état d'équilibre dans un système dépendra de plusieurs facteurs, à savoir l'activité métabolique relative de la cellule, le substrat utilisé en tant que traceur

et la composition du milieu de culture. Il s'avère parfois que des systèmes importants développés au niveau industriel n'atteignent jamais cet état d'équilibre, d'où l'importance de l'étude des voies dans un état dit non-stationnaire (Drysch *et al.*, 2003; Kelleher, 2001) par la méthode de ^{13}C-DMFA. Un exemple de ce type de modèle a été développé pour le processus de fermentation chez la bactérie *E. coli* en vue de la synthèse de 1,3-propanediol (PDO) au niveau industriel (Antoniewicz *et al.*, 2007). Les flux estimés par le modèle ont permis l'obtention de profils temporels détaillés des flux intracellulaires au cours de la fermentation, et ont fourni des informations précieuses pour expliquer le phénomène de surproduction industrielle de PDO dans ce microorganisme. Nous pouvons également citer le pendant de cette méthode n'utilisant pas de traceur isotopique : le DMFA, qui est plutôt utilisé pour identifier des changements métaboliques dans une voie (Antoniewicz, 2015; Leighty and Antoniewicz, 2011).

Dans certains cas, l'état d'équilibre est atteint dans le métabolisme, mais il ne l'est pas pour l'isotope ajouté dans la culture. Par conséquent, une autre technique a été établie pour pouvoir utiliser les données issues du traçage au carbone 13 (^{13}C) dans un modèle d'analyse de flux métabolique. Cette méthode a été baptisée ^{13}C–NMFA pour « isotopic Non-stationary ^{13}C-Metabolic Flux Analysis » (Antoniewicz, 2015; Nöh *et al.*, 2007).

Notons que nous pouvions classer ces méthodes parmi celles utilisant des données, néanmoins comme ces analyses expérimentales se font dans le même laps de temps que l'analyse métabolique, il nous est apparu convenable de les placer parmi les modèles hybrides.

Modèles à base de contraintes combinées à des méthodes de Machine Learning (ML)

D'autres techniques consistent à combiner des modèles à base de contraintes, comme les modèles d'analyse de flux métabolique, avec des méthodes d'apprentissage automatique. Il existe une multitude de combinaisons envisageables de ces deux types de méthodes et il ne nous est guère possible de tous les recenser dans cette présente étude (Zampieri *et al.*, 2019; Rana *et al.*, 2020).

Cependant, nous pouvons citer quelques-unes de ces méthodes ainsi que l'application qui a été faite par la suite. La première est celle développée sur la plateforme nommée MFlux, mise en place pour prédire le métabolisme central bactérien par apprentissage automatique à partir de données issues de modèles ^{13}C-MFA, pour ainsi prédire le profil de flux d'un microorganisme (Wu *et al.*, 2016). Trois méthodes d'apprentissage automatique ont été testées : l'algorithme de Machine à vecteurs de support ou « *Support Vector Machine* » (SVM), les K plus proches voisins ou « *K-nearest neighbors* » (kNN) et les arbres de décision. Des trois algorithmes développés, celui du SVM a donné les meilleurs résultats de prédiction du flux à partir de données sur : l'espèce bactérienne, le

type de substrat, le taux de croissance, les conditions en oxygène et les méthodes de culture. D'autres modèles, même s'ils sont peu nombreux, ont été développés pour l'exploitation de données de fluxomique pour : estimer le flux à partir des conditions génétiques et environnementales (Oyetunde *et al.*, 2019), par une combinaison de contraintes stœchiométriques et de méthodes d'apprentissage automatique (kNN, arbres décisionnels, SVM) ou encore pour prédire des données protéomiques (prédiction de constante catalytique) par une combinaison de méthodes FBA avec des forêts aléatoires ou des modèles ANN (Heckmann, 2018).

D'autres combinaisons ont été faites, notamment, pour prédire la production de molécules, comme le xylitol chez *E. coli* (Yousoff *et al.*, 2017). Dans cette étude, le modèle a été mis en place pour prédire la production du xylitol, molécule incorporée dans plusieurs produits pharmaceutiques, chez différentes souches de la bactérie exprimant ou non certains gènes. Cela afin de faciliter, par la suite, l'identification des meilleures conditions de knock-out, autrement dit d'inactivation génique, pour obtenir une production encore plus intense de xylitol.

Modèles de type boîte-grise (ou « *Grey-box model* »)

Les derniers modèles présentés dans cette partie à propos des modèles hybrides concernent des modèles développés par nos soins à partir de modèles cinétiques de voies métaboliques. Ces modèles ont été bâtis afin d'optimiser le processus de production modélisé et de parvenir à estimer le flux final des voies étudiées, reconstruites expérimentalement en *in-vitro* (Lo-Thong *et al.*, 2020).

Malgré l'absence de ce type de modèle dans le domaine de la biologie, nous pouvons nous attarder sur certains exemples appliqués dans d'autres domaines d'investigation. C'est le cas des modèles de type boîte-grise développés dans le secteur de l'énergie renouvelable (Li and Wen, 2014). Ces modèles hybrides utilisent des descriptions physiques simplifiées pour simuler le comportement complexe de systèmes énergétiques des bâtiments. L'utilisation de ces modèles permet de réduire les exigences fixées en matière d'ensemble de données et le temps de calcul nécessaire. Il a été démontré qu'un réseau de résistance et de capacité (modèle boîte-grise) a pu prédire de manière précise la charge de refroidissement des bâtiments, avec seulement des données de 1 à 2 semaines (Braun and Chaturvedi, 2002). Toujours dans le domaine énergétique, un autre modèle de type boîte-grise a été construit pour prédire le temps (température de l'air, humidité relative et radiation solaire) afin d'estimer la charge thermique du bâtiment et déterminer le réglage optimal du système de gestion de bâtiment (Zhou *et al.*, 2008).

Revenons toutefois dans notre domaine de recherche : la biologie. Si ces modèles boîte-grise, comme nous les concevons, ne sont pas référencés dans la littérature, il existe des modèles qui, similaires à ceux présentés dans la partie précédente, sont issus de la combinaison de modèles, par

exemple de modèles cinétiques de type mécanistique et des modèles de *Machine Learning*. Un modèle de ce type a été construit et combine plusieurs modules : un module composé de 2 méthodes cinétiques pour modéliser une voie métabolique et un module *Machine Learning* appliquant un algorithme de type SVM. Ainsi, tandis que le premier module s'occupe de modéliser la voie, le second module prédit les paramètres du modèle cinétique en utilisant des informations d'une base de données (Pan *et al.*, 2017).

1.2.4. Les limites de ces méthodes

Bien que ces méthodes dépeintes plus tôt présentent beaucoup d'avantages, telle une pièce de monnaie, elles présentent une autre face avec des limites qu'il nous reste à surmonter pour optimiser nos modélisations. Commençons par les modèles basés sur la connaissance. Cela n'étonne guère personne qu'une des limites des modèles cinétiques soit la connaissance en elle-même. Il n'est pas rare que les paramètres cinétiques ne soient pas connus ou sont indéfinissables dans une gamme de données (Kim *et al.*, 2018). Ou que le mécanisme réactionnel de l'enzyme soit inconnu ou tellement complexe qu'il devient difficile à en définir l'équation. Si ces cas de figure se présentent : des méthodes existent et s'appliquent à estimer les paramètres, par exemple à partir de données expérimentales de la réaction ; en ce qui concerne l'équation cinétique, la préférence se fera toujours envers la plus simple formulation. Aussi, certains puristes poseront la question des origines des paramètres cinétiques utilisés dans le modèle : peut-on vraiment utiliser des paramètres mesurés en *in-vitro* pour des modèles censés illustrer une voie métabolique *in-vivo* (Wiechert and Noack, 2011) ? Afin de combler le fossé entre l'*in-vitro* et l'*in-vivo*, des efforts sont fournis actuellement pour développer des milieux de culture synthétiques qui miment l'environnement intracellulaire (van Eunen *et al.*, 2010).

Une autre interrogation que nous pouvons soulever, est le transfert des données entre les différents organismes. Il est, en effet, de coutume d'utiliser les paramètres cinétiques d'un organisme différent, lorsque ceux de l'organisme étudié sont inconnus. Aussi, des manipulations expérimentales supplémentaires sont à effectuer afin de pallier le manque d'informations sur les enzymes.

De même, certains facteurs peuvent influencer la vitesse de réaction d'une enzyme, comme le pH, la température ou la présence de complexe enzymatique (Almquist *et al.*, 2014). Et malencontreusement, les effets de ces facteurs ne sont pas représentés dans les équations cinétiques.

Pour ce qui est des modèles d'analyse de flux métabolique, une des plus importantes limites réside dans leur utilisation de la stœchiométrie uniquement et qu'ils ne peuvent contenir d'autres informations quantitatives sur la voie (mesures expérimentales). En effet, il est très important au niveau industriel de savoir, non seulement, quel serait le rendement maximal réalisable, mais également de savoir pourquoi cet optimum n'est pas atteint en pratique ; d'où l'intérêt d'avoir un modèle contenant des régulations. Le fait que les modèles d'analyse de l'équilibre des flux n'utilisent pas de paramètres cinétiques les réduit dans leurs capacités de prédiction et ils ne peuvent pas estimer les concentrations en métabolites par exemple. Également, ils peuvent faire l'estimation des flux uniquement à l'état d'équilibre et non pas de manière dynamique (Orth *et al.*, 2010). Une autre limite de l'analyse des flux métaboliques est que le nombre de contraintes appliquées au modèle est souvent insuffisant pour analyser toutes les voies métaboliques importantes. La prédiction du flux dans ces modèles dépend des contraintes choisies, aussi beaucoup de temps est à investir pour une construction de qualité.

De plus, tout comme les modèles cinétiques, les modèles de type FBA ne se concentrent que sur une partie du métabolisme de l'organisme ; l'absence de certaines réactions, dû à des lacunes métaboliques ou une annotation incomplète du génome d'un métabolisme dans une base de données, peut introduire un biais dans la détermination du flux (Raman and Chandra, 2009).

L'une des limites notables pour la mise en place de modèles à l'échelle génomique est son temps de construction (Chong and Xia, 2017). Un tel modèle est effectivement long à mettre en place, allant de six mois pour des bactéries bien connues et possédant un génome de taille moyenne, à deux ans pour la reconstruction du métabolisme humain (Thiele and Palsson, 2010). Toutefois, il offre la possibilité de fusionner des réseaux métaboliques issus de différents organismes, afin de faciliter l'étude des interactions entre ces organismes (exemple : interaction hôte–microbiome). Aussi, certaines mesures faites au niveau de l'expression des gènes peuvent induire en erreur le modèle. C'est le cas des gènes codant les enzymes faisant partie d'un complexe enzymatique. Dans cette configuration, une forte expression du gène codant pour l'enzyme ne signifie pas nécessairement une activité plus forte de cette enzyme, puisque celle-ci est limitée par l'expression des autres enzymes qui composent le complexe (Zampieri *et al.*, 2019).

Passons maintenant à l'analyse des faiblesses présentes au sein des modèles basés sur l'utilisation de données. Ces modèles permettent généralement une bonne prédiction des voies métaboliques à partir de données expérimentales. Néanmoins, ils sont basés sur des données obtenues par des expériences qui sont sujettes aux biais et aux erreurs ; un ensemble de données de qualité est alors requis si l'on souhaite développer ce type de modèle (Zampieri *et al.*, 2019). Si ce critère-là est rempli, il reste encore le choix du modèle. Ces modèles d'apprentissage

automatique sont performants également dans certaines conditions, qui posent également les limites à leur utilisation : un grand nombre de données dans le jeu de départ utilisé et un nombre de variables à prédire relativement faible. Ce qui n'est pas le cas, par exemple, des modèles cinétiques qui sont capables de prédire plusieurs données à la fois : flux métabolique, concentrations en protéines, en métabolites ou en produits (van Riel *et al.*, 2021). De plus, contrairement aux autres modèles, les modèles de type Machine Learning n'ont pas de base « biologique » sur laquelle s'appuyer pour construire les relations entre les données qui lui sont confiées ; ce qui nous amène à nous interroger sur la fiabilité et l'interprétabilité de ces modèles (Culley *et al.*, 2020).

Enfin, si nous nous penchons sur les modèles hybrides, la plupart d'entre eux ont été mis au point pour combler un déficit présent au départ dans l'algorithme d'origine. Ils présentent peu de limites. Parmi celles-ci nous pouvons citer : le temps de développement, la connaissance disponible sur la voie métabolique et la compréhension du système dans sa globalité. Ceci est vrai dans le cas de modèles combinés ou de modèles de type boîte-grise. Plusieurs interrogations peuvent être énoncées à ce sujet. Comment déterminer les meilleures combinaisons qui vont améliorer la prédiction de notre modèle ? Avons-nous des données d'une grande diversité pour notre sujet d'étude, à savoir des données expérimentales et des données cinétiques sur la voie de production ?

1.3. De la modélisation vers le contrôle de la voie métabolique

Une fois le modèle de notre voie métabolique bâti, il est possible de s'intéresser à l'addition de mécanismes de contrôle sur cette voie. Une distinction peut être effectuée entre les différents processus de contrôles. Nous avons en premier lieu les régulations existantes naturellement au sein des voies métaboliques étudiées. Rappelons-le, beaucoup de processus de production au niveau industriel utilisent des microorganismes (recombinants ou non) pour produire la molécule d'intérêt. Il est donc important de prendre en considération ces régulations lors d'une modélisation de la voie de synthèse. Puis, nous avons des systèmes qui ont été délibérément mis en place sur des cultures cellulaires ou des bioréacteurs pour réguler le système de production. Dans les sous-parties qui suivent, nous nous attèlerons à développer les deux notions de contrôle de voie métabolique citées ci-dessus ; ce qui nous mènera à la troisième partie de cette section : la modélisation de systèmes de contrôle d'une voie.

1.3.1. Les régulations internes présentes dans une voie métabolique

L'équilibre présent naturellement dans une cellule vivante, ou à une échelle plus grande dans un organisme vivant, est régi par divers mécanismes qui vont permettre la régulation des réactions qui s'y déroulent. De la même manière, ces mécanismes permettent à l'organisme de s'adapter face aux changements métaboliques ou environnementaux qui lui seraient imposés. Cet équilibre qui maintient l'organisme en vie est donc un processus dynamique et est désigné sous le terme d'homéostasie. Ainsi, des *stimuli* vont être envoyés et réceptionnés par la cellule qui adaptera son métabolisme en conséquence en activant ou inhibant l'expression de gènes, par exemple. Ce type de régulation qui est inhérent à la vie de l'organisme n'est pas à omettre lors d'une étude sur les voies métaboliques de production, puisqu'il est susceptible d'être à l'origine d'une variation voire d'une baisse de la production de la molécule.

Parmi les régulations applicables à la majeure partie des enzymes, se trouvent celles opérées par le pH et la température (Frieden *et al.*, 1976; Ju *et al.*, 2004). Il existe en effet, pour ces enzymes une valeur optimale de pH (pH_{opt}) et de température (T_{opt}) à laquelle l'activité de l'enzyme sera maximale. Pour le pH, si l'enzyme se trouve dans un milieu où celui-ci est très inférieur/supérieur au pH_{opt}, sa charge ne sera plus la même. Sachant que la charge de la protéine est importante, tant

pour la stabilité de l'enzyme que pour sa liaison avec son substrat, ce changement de pH induira une modification de l'activité enzymatique et donc du fonctionnement global de la voie métabolique dans laquelle se trouve l'enzyme, dénaturée dans les cas les plus extrêmes. La température, elle, permet généralement l'accélération de la réaction, mais si elle est très inférieure au T_{opt}, l'enzyme est moins efficace voire inactive. Et si elle est très supérieure au T_{opt}, cela peut entraîner la dénaturation de l'enzyme qui est un processus irréversible.

Une autre régulation peut se mettre en place en amont de l'activité de l'enzyme en elle-même, à savoir au niveau de son expression. À cet effet, des phénomènes d'induction ou de répression de l'expression du gène codant pour l'enzyme peuvent réguler son expression et par extension son activité. Les responsables de ces changements sont de multiples natures. Il peut s'agir des substrats de ces enzymes. Nous en avons un exemple avec l'exoglucanase (ß-D-cellobiohydrolase) chez *Aspergillus niger* (Hanif *et al.*, 2004). L'expression de cette enzyme est contrôlée par la source de carbone utilisée au départ. Si du glucose est utilisé, il a été observé une répression du gène codant pour cette enzyme. Par contre, si la culture de champignon est mise en présence de cellobiose, nous voyons une augmentation de l'expression de ce gène. De cela, nous pouvons bien imaginer qu'à l'inverse, une déplétion en substrats ou même en cofacteurs engendrerait des changements au sein de la voie métabolique étudiée ; faisant de ces molécules des leviers importants pour réguler la production d'une molécule.

Un autre procédé de régulation, créé par les organismes pour maintenir l'équilibre dans leur métabolisme, est le contrôle de l'activité même des enzymes. Cela peut se faire par différents moyens au vu de la diversité d'enzymes qui existent (*e.g*, régulation allostérique, activation ou inhibition par des effecteurs présents dans la voie métabolique).

D'abord, et nous en avons parlé brièvement dans les sections précédentes : la régulation allostérique. En particulier, certaines enzymes possédant plusieurs sous-unités ont la propriété d'être allostérique et peuvent changer de conformation lorsqu'elles sont liées à un effecteur. Cet effecteur se lie sur un site régulateur, différent du site actif sur lequel se fixe le substrat. Si l'effecteur est un activateur de l'enzyme, celle-ci va changer de conformation, prenant une forme à forte affinité pour le substrat. Nous retrouvons ce cas de figure lors de la dernière étape de la glycolyse, catalysée par la pyruvate kinase (Jurica *et al.*, 1998). Cette enzyme peut être activée de manière allostérique par le fructose-1,6-bisphophate. Au contraire, si l'effecteur est un inhibiteur de l'enzyme, la fixation de celui sur l'enzyme va la contraindre à garder une conformation qui présente une faible affinité pour le substrat. Un exemple traditionnel est celui de l'hexokinase, où le glucose-6-phosphate est un inhibiteur allostérique de l'enzyme (Ureta, 1976). Lors de la modélisation d'une voie métabolique contenant ce type d'enzyme, cette régulation est intégrée au

niveau de l'équation cinétique décrivant l'enzyme allostérique. Un mécanisme similaire à celui de l'allostérie a été également illustré chez des enzymes monomériques appelées enzymes mnémoniques (Porter and Miller, 2012). Ces enzymes peuvent prendre une conformation différente de celle à l'état initial suite à la libération du produit de la réaction, engendrant ainsi une cinétique différente, similaire à celle des enzymes allostériques. Ce mécanisme a été développé notamment pour l'hexokinase L1 des germes de blé (Ricard *et al.*, 1974; Meunier *et al.*, 1974).

Ensuite, l'activité d'une enzyme est régulée également par d'autres molécules appelées activateurs ou inhibiteurs (Lopina, 2017). Cette fixation à l'enzyme est souvent réversible. Les activateurs enzymatiques qui sont : des ions, de petites molécules organiques telles que des peptides, des protéines et des lipides, simplifiés sous le terme cofacteurs dans la partie précédente. À l'instar de la protéine kinase C, activée par l'ion Ca^{2+}, qui se lie sur un site spécifique qui lui est dédié. Cette fixation a pour effet de changer la conformation de l'enzyme et accroître son activité (Huang, 1989). Des complexes « activateurs » peuvent être formés pour activer l'enzyme cible et augmenter son activité. Un des plus connus est le groupe formé par la calmoduline (CaM) liée aux ions calcium dans les cellules eucaryotes. La fixation de Ca^{2+} a pour effet de modifier la conformation de la protéine (CaM), lui donnant la possibilité de se lier à des enzymes pour les rendre actives. Parmi ces enzymes nous retrouvons par exemple l'adénylate cyclase, qui est responsable de la synthèse d'adénosine monophosphate cyclique (AMPc), un second messager important dans la transduction du signal dans beaucoup de voies de signalisation cellulaire (Neil *et al.*, 1985).

Les inhibiteurs des enzymes, à l'inverse des activateurs, se lient aux enzymes pour arrêter la réaction catalytique (Lopina, 2017). Ils peuvent être divisés en deux groupes différents : les inhibiteurs irréversibles et les inhibiteurs réversibles. Ceux du premier groupe sont responsables de modifications covalentes de l'enzyme : phosphorylation, méthylation, formation d'un pont disulfure. Avec par exemple la phosphorylation de la glycogène synthase, responsable de la biosynthèse de glycogène, qui entraîne son inactivation (Fang *et al.*, 2000). Le mécanisme d'action des inhibiteurs réversibles est variable : ils peuvent se lier sur le site actif (**inhibition dite compétitive**), sur un site différent du site actif (**inhibition dite non-compétitive**) ou sur le complexe enzyme-substrat (**inhibition dite incompétitive**). Ces deux types d'effecteurs (activateurs et inhibiteurs réversibles) peuvent être rajoutés également dans une modélisation d'une voie métabolique, notamment dans des modèles cinétiques où les équations cinétiques des enzymes seront modifiées par l'ajout de paramètres relatifs aux effecteurs.

Avant de clore ce paragraphe sur les régulations internes présentes dans une voie métabolique, voyons ensemble un dernier exemple de régulation : le rétrocontrôle. Il existe aussi des molécules

appelées répresseurs qui diminuent le taux de synthèse des enzymes créant ainsi un système de rétrocontrôle (« *feed-back regulation/control* ») au sein de la voie. Par exemple, dans la voie de synthèse du cholestérol, lorsque les niveaux de cholestérol cellulaire (produit de la voie) sont élevés, le taux de dégradation de l'HMG CoA réductase (« 3-Hydroxy-3-méthylglutaryl coenzyme A réductase »), enzyme qui participe à la synthèse du cholestérol, est multiplié par 10 (Espenshade, 2013).

1.3.2. Les systèmes de contrôle d'une voie de production

En dépit des systèmes existants pour contrôler les voies métaboliques, lorsque nous nous intéressons à la synthèse d'une molécule, nous recensons les systèmes qui ont été mis en place sur des cultures cellulaires ou des bioréacteurs par l'Homme pour réguler le système. Ces systèmes de contrôle ont la possibilité d'être construits à différents endroits de notre système de production, énoncés en figure 1.9, à savoir : les paramètres physico-chimiques du milieu de culture ; les concentrations en substrat et cofacteur ; la concentration et/ou l'activité des enzymes, qui sont modifiables par génie génétique ; la concentration en produit par rétrocontrôle. Analysons alors les principales méthodes que nous retrouvons lors de la production d'une molécule.

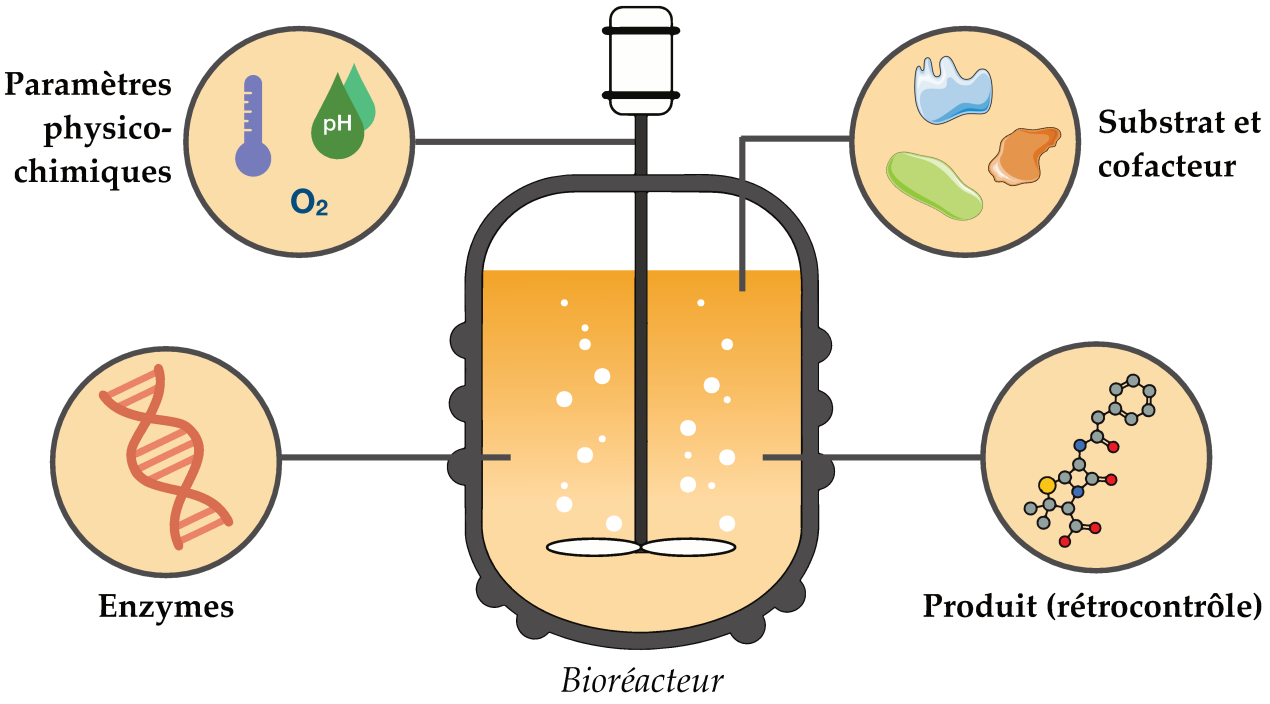

FIGURE 1.9 : *Résumé des voies d'action possible pour établir un contrôle au sein d'une voie métabolique. Les contrôles peuvent être conduits sur différents angles d'attaque dans un bioréacteur : 1) les paramètres physico-chimiques (pH, température, taux de dioxygène), 2) les concentrations en substrat et cofacteurs, 3) l'expression/activité des enzymes et 4) la concentration en produit par des systèmes de rétrocontrôle.*

Contrôle des paramètres physico-chimiques

Afin de réguler la synthèse de la molécule d'intérêt, il est possible d'agir sur le milieu de culture, en contrôlant les paramètres physico-chimiques que sont le pH, la température ou encore le taux de dioxygène. Comme nous l'avons vu précédemment, le pH et la température ont une influence sur l'activité des enzymes, et pour certaines productions ces paramètres vont varier et entraîner une modification du rendement final. De ce fait, le contrôle de ces paramètres est un procédé couramment utilisé pour contrôler une production. Il en est de même pour le taux de dioxygène, utilisé par le microorganisme responsable de la production de molécule d'intérêt si nous sommes dans un système cellulaire. La production de dihydrogène par fermentation du xylose dans l'obscurité est affectée par le changement du pH du milieu, freinant le procédé de fermentation (Calli *et al.*, 2008). Aussi, un contrôle du pH a été ajouté sur des bioréacteurs. Ce contrôle consiste en l'ajout d'un transmetteur de pH équipé d'une électrode sonde de pH et des pompes dosant la quantité de NaOH et de HCl à rajouter dans le milieu pour ajuster le pH.

Ces procédés de contrôle sont aussi employés par des équipes de recherche pour développer des réacteurs qui miment les conditions physiologiques ou pathologiques de cultures cellulaires (Mazzei *et al.*, 2008; Giusti *et al.*, 2017). Une étude a été faite sur un bioréacteur multi-compartimenté, sur lequel les chercheurs ont rajouté un système de contrôle des paramètres suivants : pH, flux de gaz, la température ; afin de mimer un environnement *in-vivo* pour simuler par la suite des environnements physiologiques ou pathologiques (Mazzei *et al.*, 2008). Le programme qui code pour le système de contrôle a une architecture inspirée du système nerveux humain avec : 1) un cerveau : le cœur du système de contrôle, 2) une carte du corps : tableau utilisé pour envoyer et recevoir des messages et 3) les « roblets » : représentent les appendices du système, qui lisent les données de l'environnement et convertissent les signaux du cerveau en action.

Contrôle des entrées du système (substrats, cofacteurs, enzymes)

Une autre manière de réguler la voie de production serait de moduler la concentration en substrats, cofacteurs et en enzymes. Un contrôle de ce type a été construit pour optimiser la production d'éthanol chez *S. cerevisiae* (Mesquita *et al.*, 2019). Dans ces travaux, un modèle métabolique à l'échelle du génome a été utilisé grâce à sa capacité d'estimer les flux inter- et intracellulaires de métabolites dans différentes conditions environnementales. Ainsi grâce aux mesures prédites, un algorithme de contrôle a pu estimer les corrections à effectuer sur le flux de substrat et d'oxygène et appliquer ces modifications sur l'alimentation en oxygène et le débit

d'alimentation en substrat de départ. Ce contrôle a permis un contrôle précis d'oxygène, nécessaire dans plusieurs bioprocédés qui requièrent une aération restreinte.

Tandis que les systèmes de contrôle des concentrations en substrat et en cofacteur peuvent être « facilement » mis en place sur des systèmes de production au moyen de sondes et de pompes, ceux des enzymes nécessitent une méthode un peu plus recherchée. Avec le développement de méthodes sophistiquées d'ingénierie métabolique et de techniques de « *gene-editing* », il est possible d'améliorer les voies métaboliques en opérant différentes actions sur les gènes codant nos enzymes (Cheng *et al.*, 2017), comme :

- Surexprimer ceux codant pour les enzymes qui interviennent dans la voie métabolique d'intérêt ;
- Supprimer l'expression des ceux qui codent des enzymes dites « concurrentes », qui utilisent le même substrat que nos enzymes d'intérêt ;
- Éditer le gène d'une ou plusieurs enzymes d'intérêt pour modifier sa structure et augmenter ainsi son affinité pour le substrat de départ, par exemple.

Dans cette même rubrique, nous retrouvons également la reconstruction de voie métabolique dans des systèmes cellulaires ou acellulaires en combinant des enzymes d'origines différentes pour améliorer le rendement final (Petroll *et al.*, 2019).

Les exemples cités ci-dessus font partie des méthodes « classiques » utilisées pour réguler une production dans un milieu de culture. Néanmoins, récemment, d'autres systèmes ont vu le jour, permettant une régulation plus dynamique de la voie de production. Ils consistent à placer une enzyme, importante dans la régulation du flux, sous un **contrôle dynamique** (Tan and Prather, 2017). En d'autres termes, l'expression de cette enzyme charnière sera induite soit lorsqu'un métabolite sera ajouté dans le milieu de culture, soit lorsqu'il y aura une accumulation suffisante d'un métabolite produit naturellement par le microorganisme dans le milieu de culture (biocapteur), ou soit lorsque le taux de croissance du microorganisme sera suffisant pour permettre le passage à la voie de production. Le contrôle peut aussi être mis en place sur une voie importante pour la survie de l'organisme et parallèle à la voie de production étudiée. Une illustration de ces méthodes a été développée pour la production d'isobutanol chez *S. cerevisiae* (Tan *et al.*, 2016). Chez ce microorganisme, trois enzymes existent pour catalyser la première réaction de la glycolyse. Deux des gènes codants pour l'hexokinase 2 et la glucokinase 1 ont été supprimés pour diminuer la concentration de leur produit et rediriger le flux de la glycolyse vers la voie de production. Et le gène restant, codant pour la troisième enzyme (hexokinase 1) catalysant la première réaction de la glycolyse chez cette levure a été mise sous contrôle d'un promoteur actif en

absence de doxycycline dans le milieu (figure 1.10). Ainsi, lorsque le milieu ne contient pas cette molécule, la levure peut croître librement dans le milieu de culture et lorsque la doxycycline est rajoutée, l'hexokinase 1 n'est pas exprimée et le flux de glucose est redirigé vers la voie de production. L'ajout de ce système a permis une augmentation de la production d'isobutanol, en plus d'une diminution de la production de sous-produit (éthanol).

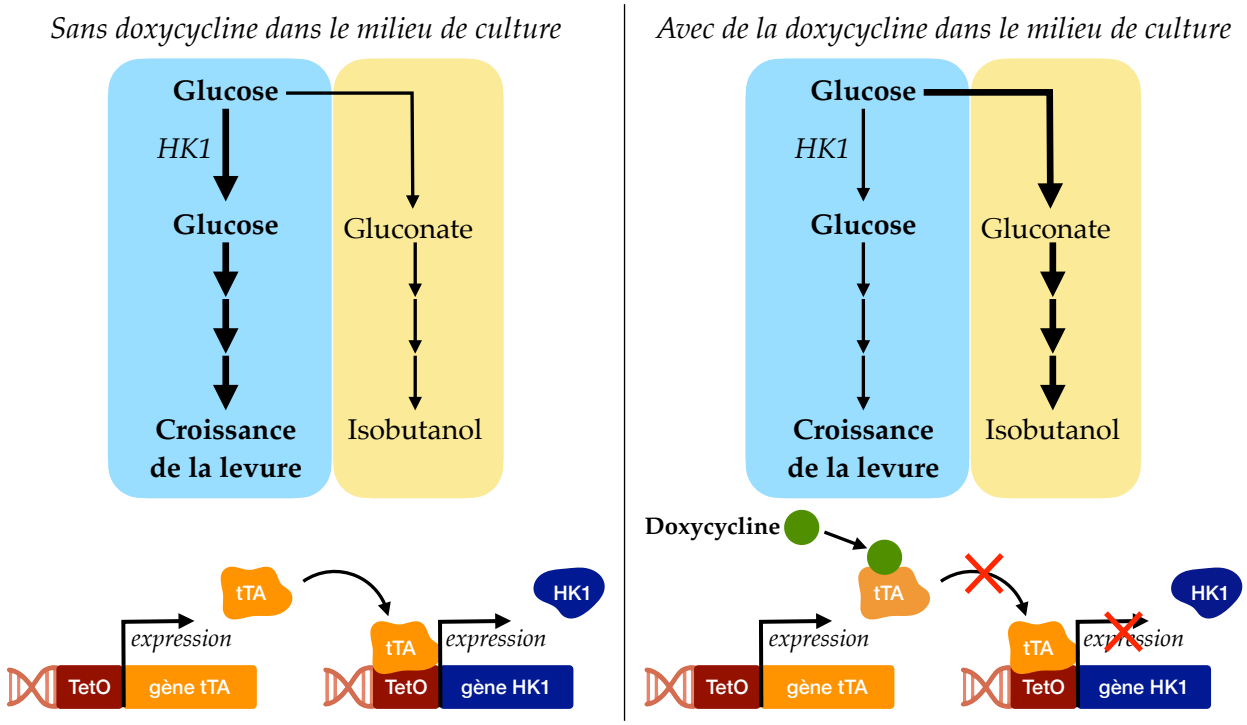

FIGURE 1.10 : *Schéma illustrant la construction génique bâtie chez le microorganisme étudié pour promouvoir la production d'isobutanol.*
La séquence métabolique en gras est celle qui est majoritaire dans le microorganisme pour le cas étudié. tTA : protéine transactivatrice de la tétracycline ; TetO : séquence opératrice spécifique où peut se lier la protéine tTA, qui active alors l'expression du gène en aval.

Système de rétrocontrôle (produits)

La régulation des produits synthétisés par la voie de production est aussi importante si nous souhaitons établir un contrôle efficace de cette voie. Il arrive parfois que l'accumulation de la molécule produite soit néfaste pour l'organisme ; à l'image du 1,4-butanediol qui peut être toxique pour la bactérie *E. coli* (Wehrs *et al.*, 2019; Burgard *et al.*, 2016). Des systèmes de drainage des flux issus des bioréacteurs ont été développés à cet effet (Young *et al.*, 2012). Une autre alternative est d'utiliser un réacteur différent, et préférer un système semi-fermé, par exemple, lorsque la molécule synthétisée est toxique pour le microorganisme qui la produit (Qureshi, 2009).

Les systèmes de rétrocontrôle ou « *feed-back control* » se servent de la quantité de produits ou de métabolites présents dans le milieu pour apporter un régulation sur le système. Cette régulation

peut avoir plusieurs finalités, comme par exemple maintenir le flux de production à un certain niveau. Une proposition de régulation par rétroaction fut élaborée pour un processus de fermentation de la levure (Smets *et al.*, 2000). La finalité de l'ajout de ce système est de permettre une production maximale de biomasse. Deux méthodes de contrôle sont proposées et se basent sur l'utilisation des données mesurées en ligne : la concentration en substrat (alimentant le réacteur) et celle de la biomasse ; ou d'une seule donnée : la concentration en substrat. Ces contrôles, en boucle fermée, tiennent compte des mesures de concentration mesurées et sont ajoutés dans une équation permettant l'ajustement de l'alimentation du réacteur en substrat.

Deux autres systèmes de rétrocontrôle ont été proposés pour optimiser la production d'éthanol et de biomasse par la levure *S. cerevisiae* (Persad *et al.*, 2013). L'un d'entre eux se base sur l'utilisation de mesures effectuées des concentrations en éthanol et des mesures de masse cellulaire. Il calcule ensuite l'erreur entre la concentration mesurée et la valeur cible pour déterminer les ajustements à effectuer sur les entrées du bioréacteur. Ce système est un contrôle de type Proportionnel-Intégral-Dérivé (ou PID) que nous développerons plus en détail dans la section suivante. Les résultats obtenus avec l'ajout de ce type de régulation ont montré la performance de ces rétrocontrôles pour maximiser la biomasse et l'éthanol produits.

Par ailleurs, certains systèmes développés dans le paragraphe précédent, à l'instar des processus de contrôle dynamique, peuvent être aussi classés dans cette régulation par rétrocontrôle sur le système. Bien que le contrôle soit effectué sur des enzymes, il est amorcé et dépend dans certains cas de la concentration en produit, et peut donc être classé dans cette catégorie.

Dans cette partie, nous avons apprécié l'intérêt de mettre en place des systèmes de contrôle dans une voie métabolique dans le but de produire une molécule. Attelons-nous, dans la prochaine section, à présenter les méthodes *in-silico* actuelles employées comme systèmes de contrôle.

1.3.3. La modélisation de systèmes de contrôle

Les technologies actuelles nous permettent d'établir des modélisations informatiques (*in-silico*) de systèmes de contrôle, avant de les mettre en place sur de « vrais » processus de production. Notre étude s'intéresse à la mise en place de procédé de contrôle sur un modèle de voie métabolique. Ainsi, afin de faire le point sur les procédés déjà modélisés et d'en distinguer les plus intéressants, voyons ensemble trois types de systèmes de contrôle : ceux basés sur des systèmes

d'Intelligence Artificielle, ceux modélisant des rétrocontrôles et enfin un système intéressant spécifique, le contrôle de type Proportionnel-Intégral-Dérivé (ou PID).

Contrôle basé sur des systèmes d'Intelligence Artificielle

Grâce à leur performance à résoudre différents types de problèmes, les systèmes d'Intelligence Artificielle (IA) sont intéressants pour établir une régulation. Si nous considérons en plus, que les fermentations microbiennes sont difficiles à suivre en ligne (« *on-line measurements* »), les systèmes intelligents pourraient être d'une grande aide pour les contrôler. En effet, la force de ces systèmes réside dans leur capacité à tirer des enseignements des performances du processus de production, à en tirer des déductions, grâce à des raisonnements proches du raisonnement humain, pour ensuite anticiper les problèmes (Patnaik, 1997). L'opérateur pourra alors par la suite y pallier en agissant sur différents paramètres du système.

Nous retrouvons un modèle de ce type dans le contrôle de cultures multi-espèces, utilisées pour de la bio-production (Treloar *et al.*, 2020). Ces cultures multi-espèces seraient plus productives comparées aux cultures mono-espèce, néanmoins elles sont rarement utilisées en pratique à cause du contrôle difficile des espèces constitutives de chaque culture. L'utilisation d'un modèle de contrôle de type IA a été efficace pour réguler les co-cultures dans un bioréacteur et optimiser la sortie du processus de production. Ce contrôle a été modélisé en utilisant une technique d'apprentissage par renforcement, combinée à des réseaux de neurones. La technique d'apprentissage par renforcement permet à un agent (contrôleur physique du réacteur, robot…) d'apprendre les actions à effectuer à partir des observations faites de son environnement. Dans le cas du bioréacteur étudié, contenant les différentes cultures, l'agent va décider, à chaque instant t, s'il doit fournir le nutriment (celui spécifique à l'une des espèces) à l'environnement ou non. L'objectif est soit de maintenir à un niveau donné les microorganismes ou de maximiser le produit en sortie. Ainsi, selon la distance du niveau de population des microorganismes avec la valeur cible ou du niveau de produit synthétisé avec la valeur cible, l'agent délivrera ou non une certaine quantité en substrat au microorganisme concerné.

Comme il a été énoncé plus tôt, l'un des moyens de réguler une voie métabolique serait d'agir sur les paramètres physico-chimiques. Une étude sur la fermentation alcoolique dans la levure a montré l'efficacité d'un modèle, utilisant la méthode des réseaux de neurones artificiels, à contrôler la température au sein du bioréacteur (Nagy, 2007). L'un des avantages à utiliser cet algorithme est le fait qu'il ne nécessite pas de connaissances détaillées sur le processus de production, par rapport à d'autres techniques de modélisation.

Système de rétrocontrôle *in-silico*

Il existe aussi des méthodes menant à la modélisation de systèmes de rétrocontrôle sur des voies de production. Parmi elles, se trouvent les modèles de contrôle de type boucle-fermée qui opèrent des ajustements sur une variable à contrôler en comparant la sortie du système à une valeur de consigne donnée par l'instructeur. Ils sont généralement composés de :

- Un capteur : qui mesure la valeur réelle de la sortie du système ;
- Un contrôleur : qui compare cette valeur à la valeur consigne et qui génère un signal de commande ;
- Un actionneur ou organe de commande : qui correspond à l'appareil qui va effectuer la nouvelle commande pour ajuster la variable à contrôler.

Un système de contrôle de type boucle-fermée a été créé pour réguler la production de 1,3-propanediol lors d'un procédé de fermentation en continu (Bei *et al.*, 2019). Dans cet exemple, ils ont établi deux variables de contrôle : le taux de dilution et la concentration en glycérol (substrat), qu'ils ont par la suite intégré, sous forme d'une combinaison linéaire, à une loi de contrôle linéaire. Le contrôle ainsi intégré a pour but de maximiser la concentration en 1,3-propanediol après avoir mesuré sa concentration.

Ce type de système de contrôle peut être appliqué dans d'autres domaines également, comme celui de la recherche pharmaceutique. Un système de contrôle linéaire à boucle-fermée a été construit pour adapter la quantité d'anesthésique à administrer aux patients en fonction des exigences de l'intervention chirurgicale et de l'état clinique général du patient (Silva *et al.*, 2015).

Un des modèles phares utilisé pour établir des systèmes rétrocontrôles est : la régulation Proportionnelle-Intégrale-Dérivée. Voyons, dans ce prochain paragraphe la constitution de ce système de contrôle, son fonctionnement et quelques exemples d'application.

La régulation Proportionnelle Intégrale Dérivée (PID)

Le système de contrôle Proportionnel-Intégral-Dérivé ou plus simplement régulateur PID est couramment utilisé en industrie pour établir un contrôle au sein du système (Chevalier *et al.*, 2018; Wang, 2020). La modélisation de ce type de système s'avère très intéressante, puisqu'il s'agit d'un système classique, facile à mettre en place en pratique et robuste. C'est un régulateur qui fait partie d'un système boucle-fermée et qui sert à maintenir une sortie, de sorte qu'il n'y ait pas ou très peu d'erreur entre la variable de sortie et la consigne entrée par l'opérateur. Le rétrocontrôle exercé est le résultat de la combinaison de trois actions de contrôle représentées en figure 1.11 :

- Terme « proportionnel » (P) qui correspond au calcul de l'erreur, entre la variable de sortie et la valeur consigne, multipliée par un gain (K_P).
- Terme « intégral » (I), où l'intégrale de l'erreur est ajoutée. L'erreur entre la consigne et la mesure est intégrée par rapport au temps et est aussi multipliée par un gain (K_I). Le terme intégral est rajouté pour corriger l'erreur statique, qui demeure lorsque l'erreur est infime et que le terme proportionnel n'est plus suffisant pour modifier la valeur de sortie pour atteindre la valeur consigne. Il résulte en un modèle plus stable.
- Terme « dérivé » (D), comme son nom l'indique utilise l'erreur sous sa forme dérivée (par rapport au temps) et multipliée par un troisième gain (K_D). Le contrôle PI (avec les deux termes Proportionnel et Intégral uniquement) peut parfois aboutir à un dépassement de la consigne ; le terme dérivé permet de limiter ce dépassement. Ainsi, ce terme a pour objectif d'anticiper le comportement futur de l'erreur et de freiner le système de contrôle en appliquant une action dans le sens opposé. Ce terme est à utiliser avec précaution, puisqu'il peut conduire à une usure prématurée des actionneurs, dûe à une utilisation excessive pour opérer de « petites » modifications sur le système.

Ce régulateur peut contenir :

- le premier terme uniquement et prend le nom de régulateur de type P pour Proportionnelle ;
- les deux premiers termes et prend le nom de régulateur de type PI pour Proportionnelle-Intégrale ou ;
- les trois termes à la fois et prend le nom de régulateur de type PID pour Proportionnelle-Intégrale-Dérivée. Une fois la commande calculée par le régulateur PID, les modifications sont envoyées vers le système de production, dans notre cas, il s'agira d'envoyer ces commandes vers le modèle de la voie métabolique étudiée.

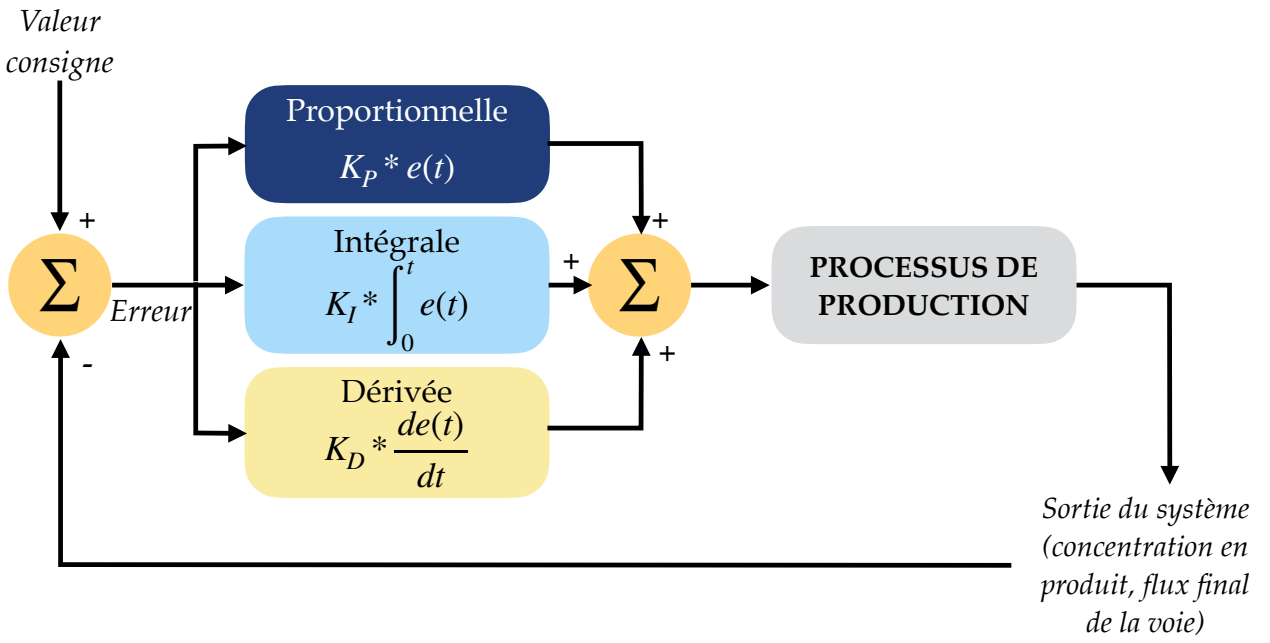

FIGURE 1.11 : *Schéma représentatif d'un système de rétrocontrôle PID.*
La sortie du système peut être dans notre cas la concentration en produit final ou le flux final d'une voie de production. Cette sortie est mesurée et comparée à la valeur consigne enregistrée. L'erreur $e(t)$ *est ensuite intégrée au régulateur PID, qui calcule la commande finale à appliquer sur le processus de production.* K_P, K_I *et* K_D *représentent les gains de chaque terme de correction, respectivement celui du terme Proportionnelle, Intégrale et Dérivée. Il s'agit des paramètres du système de contrôle à régler.*

Plusieurs modélisations de ce type de contrôle existent pour contrôler :

- La température lors de la production d'éthanol dans un bioréacteur : le régulateur agit sur la température et réalise des corrections de cette variable en fonction du flux en sortie du réacteur (Kumar *et al.*, 2019).
- Le pH de la culture-mixte faite avec *Lactococcus lactis* et *Kluyveromyces marxianus* et ainsi assurer une production optimale de nisine, un peptide antimicrobien. Les variables mesurées en sortie sont le taux d'oxygène dissous et le pH, tandis que la variable de contrôle est le pH de manière indirecte, via le taux d'oxygène dissous (Shimizu *et al.*, 1999).
- Le taux de croissance d'*Escherichia coli*, pour la production de protéine recombinante, en modulant le taux d'alimentation en substrat (Galvanauskas *et al.*, 2019).
- L'administration de médicaments anesthésiques pour un blocage neuromusculaire, afin d'éviter des phénomènes de sous-dosage ou de surdosage lors d'une administration automatisée de médicaments (Medvedev *et al.*, 2019).

1.4. Objectifs de la thèse

L'ensemble des recherches réalisées dans ce travail de thèse s'inscrit dans le cadre du développement de méthodes de i) modélisation de voie métabolique, utilisée à des fins de production de molécule et ii) modélisation de système de contrôle de cette voie métabolique. L'enjeu principal étant de parvenir à un modèle intégrant un contrôle de la voie métabolique, qui puisse être appliqué, sans encombre, à d'autres voies métaboliques. Pour ce faire, nous nous sommes appliqués à choisir nos modèles d'application, en gardant bien en mémoire, qu'ils servent uniquement d'exemples pour enrichir nos propos sur la modélisation faite d'une voie de production et de son système de contrôle.

Ces voies de production trouvent leur origine dans le monde qui nous entoure, et de manière plus précise dans les microorganismes (bactéries, levures, champignons, parasites). Une revue des modélisations de voie métabolique déjà existantes a été faite et nous a conduit vers le choix du parasite, en termes de microorganisme de départ. Il est très peu question, en effet, de production de molécules à grande échelle par le biais de parasites ; si ce n'est celle de l'équipe de M. Aydogdu *et al.* pour la production d'antigènes de *Leishmania infantum* en vue de la production de vaccins ou de kit de diagnostic (Aydogdu *et al.*, 2019). Il nous a paru intéressant de nous pencher sur des voies métaboliques issues de ces microorganismes.

Parmi les voies métaboliques naturelles qui existent, nous avons porté notre attention sur deux d'entre elles : la glycolyse chez le parasite *Entamoeba histolytica* et la voie de détoxification du peroxyde chez *Trypanosoma cruzi*. La première voie mène à la production de pyruvate, qui a un rôle pivot dans la production d'autres molécules d'intérêt synthétisées à partir de sucres. La seconde voie est celle de la réduction du peroxyde libérant de l'eau au passage. L'intérêt d'avoir choisi cette deuxième voie ne réside pas tant dans la molécule qu'elle produit, contrairement au premier exemple, mais dans sa capacité à éliminer les molécules néfastes pour le microorganisme, que nous nous trouvions dans une culture cellulaire ou acellulaire (Hu *et al.*, 2020; Martínez, 2016; Bowie *et al.*, 2020).

Concernant le mécanisme de contrôle développé dans le modèle, nous nous sommes orientés vers la modélisation d'un régulateur PID, en raison de la facilité de son implémentation dans un réacteur et de la rareté des travaux faisant recours à de tels mécanismes de contrôle de voie métabolique, notamment pour le contrôle du flux en sortie d'une voie et ce, bien que le régulateur PID soit beaucoup utilisé dans d'autres champs disciplinaires (physique, génie des procédés). Le choix de cette variable de sortie, comparée aux autres variables en sortie d'un système de

production, s'est fait selon le type de mesures disponibles dans la majeure partie des études utilisant un bioréacteur.

Les travaux présentés dans les chapitres suivants ciblent les axes suivants (figure 1.12) :

1. Modélisation des voies métaboliques par une méthode au moyen de modèles cinétiques ou de modèles hybrides basés sur des modèles cinétiques (méthode dite « boîte-grise »).
2. Modélisation des voies métaboliques par des techniques d'apprentissage automatique, basées sur l'utilisation de données expérimentales.
3. Implémentation d'un système de contrôle PID sur le modèle de voie métabolique.

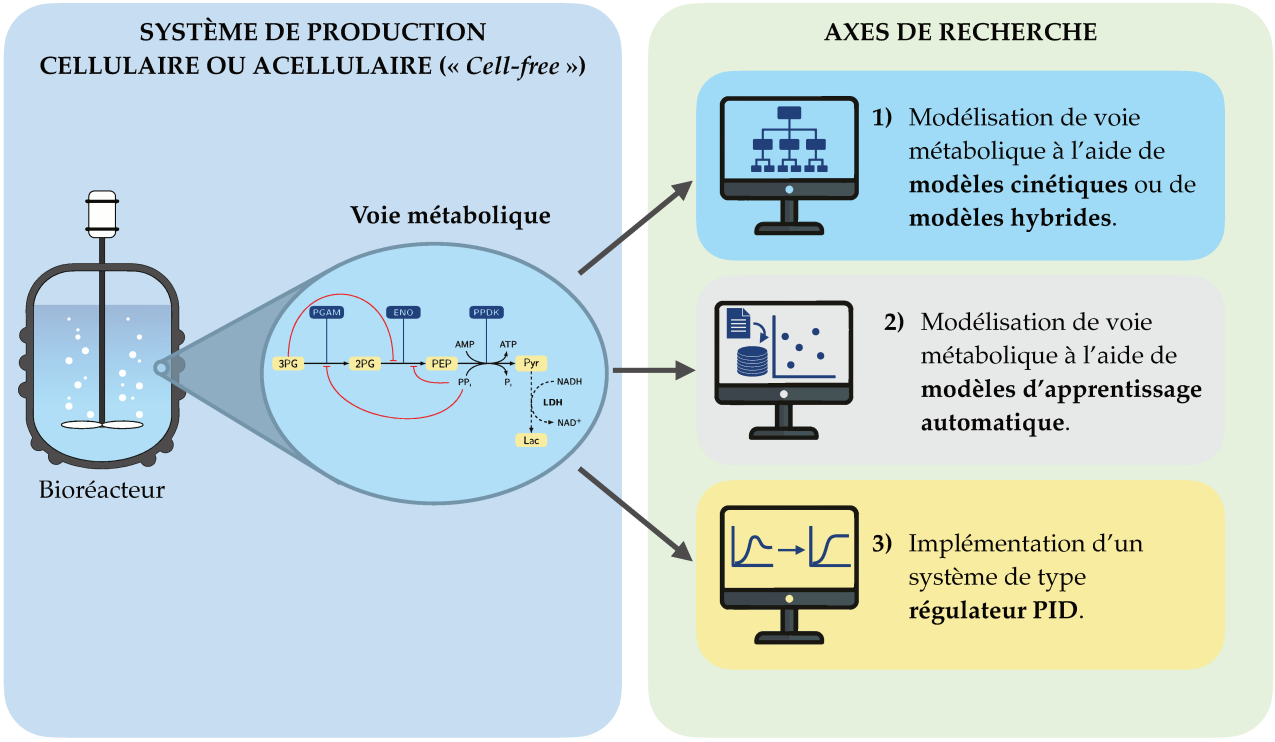

FIGURE 1.12 : *Schéma bilan des objectifs et du positionnement de ce travail de recherche dans le domaine de la modélisation de voie métabolique pour la production de molécule.*
Le cadre général de ces travaux est le développement de système de production de molécule, grâce à la modélisation de voies métaboliques. Trois principaux axes de recherches sont définis pour ces travaux et sont décrits dans le panel de droite.

Les chapitres qui suivent exposent les résultats essentiels réalisés au cours de ces travaux de thèse, suivis des conclusions de ces travaux et des perspectives de recherche.

Chapitre 2

Modélisation de voies métaboliques par une méthode dite « boîte-grise »

Comme il a été annoncé précédemment, l'un des axes de ces travaux réside dans la modélisation de voie métabolique à l'aide de modèles cinétiques ou hybrides. Les modèles cinétiques sont les représentations les plus détaillées des voies métaboliques, décrivant chaque réaction biochimique de la voie avec des équations cinétiques. De ce fait, ils constituent un outil robuste pour la description et l'analyse d'une voie métabolique et un bon point de départ pour le développement d'un modèle hybride optimisé.

Les modèles cinétiques « classiques » que nous développons dans ce chapitre présentent des équations cinétiques, de type mécanistique, adaptées au schéma réactionnel des enzymes contenues dans la voie métabolique. Elles présentent des formulations allant des plus simples (par exemple : cinétique michaélienne) aux plus complexes (par exemple : réaction réversible impliquant plusieurs substrats).

Aussi, l'utilisation de ces modèles implique la connaissance :

- De chaque réaction enzymatique de la voie,
- Des paramètres cinétiques des enzymes,
- Des mécanismes réactionnels de ces enzymes, et enfin,
- De la concentration des constituants de cette voie (substrats, cosubstrats, enzymes, inhibiteurs, activateurs).

Or, il arrive que cette connaissance ne soit pas complète pour la voie métabolique étudiée ou que les mécanismes réactionnels des enzymes ne soient pas connus ou trop complexes pour être modélisés de cette manière.

Ce présent chapitre se propose de développer un modèle de la voie basse de la glycolyse du parasite *Entamoeba histolytica* (Moreno-Sánchez *et al.*, 2008) de deux manières : par un modèle cinétique classique et par un modèle hybride. Décrit en tant que nouvelle méthode de modélisation de voies métaboliques, ce modèle est appelé modèle « boîte-grise » (*grey-box*). Cette nouvelle méthode est fondée sur l'usage d'un modèle cinétique mécanistique et comprend une partie « boîte-noire » avec l'implémentation d'une équation cinétique contenant un terme d'ajustement.

Les modèles cinétiques sont développés au moyen d'un logiciel nommé COPASI pour « ***CO****mplex* ***PA****thway* ***Si****mulator* » (Hoops *et al.*, 2006). Ce logiciel est spécialisé dans la conception, l'analyse et l'optimisation de réseaux métaboliques par la genèse de modèles cinétiques. Une validation des modèles créés est faite à partir de données expérimentales mesurées en *in-vitro* (Moreno-Sánchez *et al.*, 2008). La construction du modèle hybride se base sur le meilleur modèle cinétique obtenu et validé auparavant. Ce nouveau modèle hybride présente une particularité : la présence d'un terme d'ajustement au sein d'une des équations cinétiques, dans le but de simplifier cette équation et optimiser le flux final en sortie de la glycolyse. L'ajustement de ce nouveau terme est également effectué à partir des données expérimentales obtenues par l'équipe du Département de Biochimie de l'Institut National de Cardiologie Ignacio Chávez[1] basée au Mexique, et avec laquelle nous collaborons.

Les résultats exposés dans ce chapitre analysent l'efficacité d'un modèle cinétique à représenter une voie métabolique, contenant des enzymes aux mécanismes complexes et détaillent la conception et la mise au point d'un modèle inexistant jusqu'à présent : le modèle boîte-grise. Ce qui permet d'élargir le panel de techniques de modélisation des voies, en démontrant l'efficacité de ces modèles hybrides pour prédire le flux final de la voie, notamment lorsque peu de paramètres sont connus pour la voie métabolique étudiée.

Ces travaux constituent une partie de l'article présenté au **chapitre 3** publié dans *Scientific Reports* (Lo-Thong et al., 2020).

[1] Departamento de Bioquímica, Instituto Nacional de Cardiología Ignacio Chávez. Mexico City

Identification of flux checkpoints in a metabolic pathway through white-box, grey-box and black-box modeling approaches[2]

Ophélie Lo-Thong, Philippe Charton, Xavier F. Cadet, Brigitte Grondin-Perez, Emma Saavedra, Cédric Damour and Frédéric Cadet

ABSTRACT

Metabolic pathway modeling plays an increasing role in drug design by allowing better understanding of the underlying regulation and controlling networks in the metabolism of living organisms. However, despite rapid progress in this area, pathway modeling can become a real nightmare for researchers, notably when few experimental data are available or when the pathway is highly complex. Here, three different approaches were developed to model the second part of glycolysis of *E. histolytica* as an application example, and have succeeded in predicting the final pathway flux: one including detailed kinetic information (white-box), another with an added adjustment term (grey-box) and the last one using an artificial neural network method (black-box). Afterwards, each model was used for metabolic control analysis and flux control coefficient determination. The first two enzymes of this pathway are identified as the key enzymes playing a role in flux control. This study revealed the significance of the three methods for building suitable models adjusted to the available data in the field of metabolic pathway modeling, and could be useful to biologists and modelers.

[2] Article plus détaillé dans le **chapitre 3**.

2.1. Introduction

L'*Entamoeba histolytica* est un protozoaire existant sous deux formes : le kyste infectieux ou le trophozoïte amiboïde. L'être humain est son hôte naturel principal, infecté généralement par l'ingestion de kystes présents dans les aliments ou l'eau, contaminés par de la matière fécale. Les kystes matures sont ronds, entre 10-15 μm de diamètre, et sont dotés d'une paroi (figure 2.1).

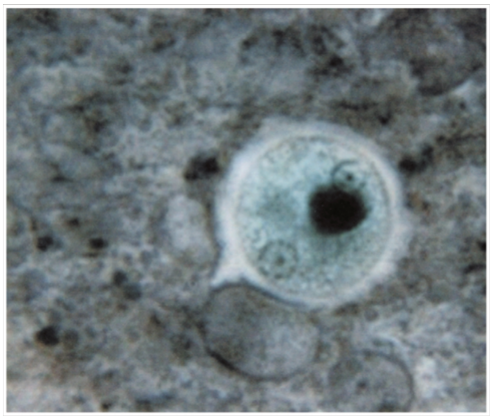

FIGURE 2.1 : *Photographie de l'amibe E. histolytica dans les selles, sous sa forme de kyste. La coloration du noyau et du cytoplasme a été faite au noir de chlorazol. Les noyaux sont clairement visibles. L'image a été empruntée des travaux de Samuel L. Stanley Jr (Stanley and Li, 2001).*

Une fois dans l'organisme hôte, le parasite survit à l'acide de l'estomac et traverse le système digestif pour ensuite se transformer en trophozoïte. Cette forme de l'amibe est très mobile comparée au kyste et peut prendre plusieurs formes possibles (forme pléomorphe). Cette mobilité, qui est à l'origine de son caractère infectieux, est rendue possible par la production d'énergie. Énergie produite lors de la conversion anaérobique du glucose en pyruvate, puis en éthanol. Cette voie métabolique est donc essentielle à la survie du parasite et contient une dizaine de réactions allant du glucose à l'éthanol.

L'infection par ce parasite (amibiase) peut parfois être asymptomatique. Mais elle peut revêtir des formes les plus sévères, avec l'apparition de colite amibienne, avec une diarrhée glairo-sanglante, et de formes extra-intestinales, avec des abcès au niveau du foie et des lésions cutanées. Ces formes dangereuses se mettent en place lorsque la forme trophozoïte adhère aux cellules épithéliales du côlon. Elle lyse ensuite les cellules par apoptose et envahit les autres tissus. Bien que l'infection par cette souche ne soit pas majoritaire, parmi celles causées par les *Entamoeba*, elle est responsable de plus de 100 000 morts par an (Kantor *et al.*, 2018). Beaucoup de recherches ont été faites à son sujet, en vue du développement d'un vaccin (Roncolato *et al.*, 2015; Quach *et al.*,

2014; Stanley, 2006; Min *et al.*, 2016) ou de molécules thérapeutiques pouvant soigner l'infection (Bansal *et al.*, 2004; Orozco *et al.*, 2009; Gonzales *et al.*, 2019; Mi-ichi *et al.*, 2019).

Parmi ces études se trouve celle de l'équipe de R. Moreno-Sánchez *et al.* qui s'est focalisée sur la représentation *in-vitro* des deux segments de la glycolyse pour valider la modélisation qu'ils avaient faite de la voie (Moreno-Sánchez *et al.*, 2008). Le modèle développé dans cette étude a éclairé les enzymes importantes dans la régulation du flux de la glycolyse chez le parasite, à savoir l'hexokinase pour la partie haute de la glycolyse (allant du glucose au dihydroxyacétone phosphate) et la pyruvate phosphate dikinase (PPDK) pour la partie basse de la glycolyse (allant du 3-phosphoglycérate au pyruvate).

Le modèle développé dans ces travaux est un modèle cinétique utilisant un logiciel conçu avant COPASI : GEPASI (Mendes, 1993). Ce logiciel simule la cinétique de systèmes de réactions biochimiques et permet également l'ajustement de ces modèles et l'optimisation de fonctions présentes dans le modèle et l'analyse du contrôle métabolique de la voie.

Dans cette étude, nous nous intéresserons à la modélisation de la partie basse de la glycolyse, illustrée en figure 2.2, à l'aide de modèles cinétiques, bâtis sur COPASI. Ce choix s'est basé sur le fait que ce segment comporte une enzyme qui est une cible thérapeutique ou un marqueur diagnostique potentiel pour l'infection par ce parasite : PPDK (Stephen *et al.*, 2008; Saidin *et al.*, 2014; Ali and Nozaki, 2007). Une fois ce modèle bâti, un processus de validation est réalisé avec l'aide des données expérimentales (flux final) récupérées dans les travaux précédents (Moreno-Sánchez *et al.*, 2008). Puis, nous nous attacherons au processus de développement d'un modèle cinétique aux prédictions de flux améliorées.

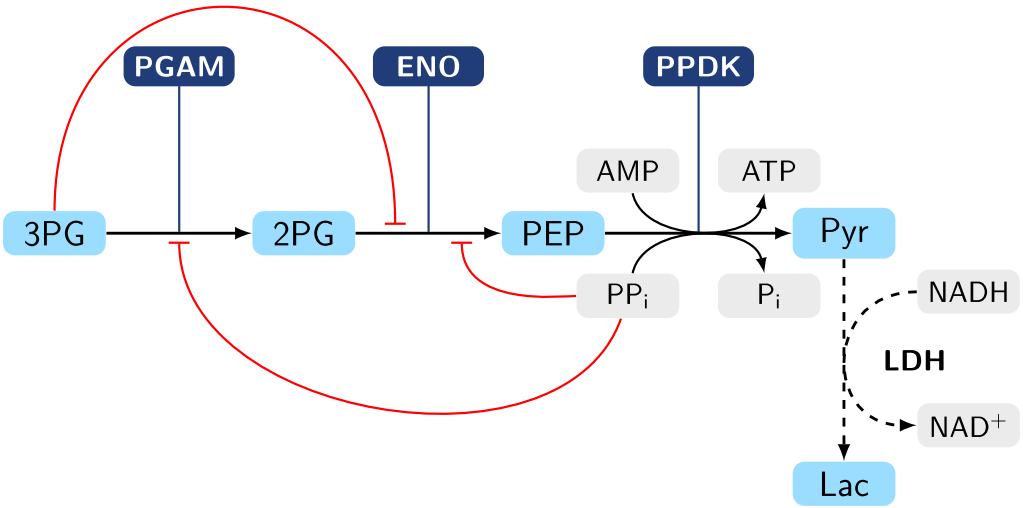

FIGURE 2.2 : *Schéma de la voie basse de la glycolyse d'E. histolytica.*
Formation de pyruvate (Pyr) à partir de 3- phosphoglycérate (3PG). La formation de L-lactate (Lac), en lignes pointillées, ne fait pas partie de la voie naturelle ; cependant, la lactate déshydrogénase (LDH) a été

ajoutée afin de suivre expérimentalement le flux final en in-vitro et d'établir un état quasi-stationnaire ; elle est donc gardée dans la modélisation. Les inhibitions sont représentées en rouge. PGAM, 3-phosphoglycérate mutase ; 2PG, 2-phosphoglycérate ; ENO, énolase ; PEP, phosphoénolpyruvate ; PPDK, pyruvate phosphate dikinase (Moreno-Sánchez et al., 2008).

2.2. Méthodes

2.2.1. Données expérimentales utilisées

Afin d'évaluer la performance de nos modèles à prédire le flux en sortie de la voie métabolique modélisée, nous utilisons les données expérimentales provenant d'une reconstruction *in-vitro* du segment de la voie basse de la glycolyse (Moreno-Sánchez *et al.*, 2008). Ces données sont composées des activités des trois enzymes contenues dans ce segment (figure 2.2) et du flux final mesuré de manière *in-vitro* pour chaque combinaison d'activité enzymatique testée. L'ensemble de données expérimentales comporte 184 combinaisons d'activités enzymatiques ainsi que la valeur correspondante du flux en sortie du segment de la voie étudiée.

2.2.2. Modélisation de la voie basse de la glycolyse à l'aide d'un modèle cinétique

Comme nous l'avons énoncé un peu plus tôt, la modélisation de la voie basse de la glycolyse de notre parasite d'intérêt est faite sur le logiciel libre d'accès COPASI (Version 4.24). Même s'il ne possède pas d'interface graphique, permettant à l'utilisateur de créer sa voie sous forme de schéma, comme le proposent d'autres logiciels de modélisation, COPASI permet la construction de modèles cinétiques en suivant les étapes suivantes :

- La création des entrées pour tous les constituants de la voie (substrats, cosubstrats, métabolites, produit, enzymes) ;
- Le renseignement des informations sur la stœchiométrie et les concentrations de chaque espèce ;
- La définition des réactions enzymatiques et de leur cinétique de réaction ;
- La création des fichiers de sortie (prédictions des concentrations/flux) ;
- Le réglage et lancement de la simulation.

Suite à la définition du modèle, le programme COPASI construit les équations différentielles qui régissent le comportement du système et les résout lors de la simulation. Les résultats produits peuvent être importés dans des feuilles de calcul ou être exportés sous forme de graphiques.

Le premier modèle créé utilise des concentrations en substrat et en cosubstrat, menant à l'obtention d'un état quasi-stationnaire en *in-vitro*, qui sont données dans le tableau 2.1. En ce qui

concerne les équations cinétiques utilisées pour décrire les réactions enzymatiques, chacune est spécifique à l'enzyme qu'elle décrit.

Metabolite	Pseudo-steady state concentrations (in μM)
3PG	4000
AMP	200
PP_i	1700
ATP	3000
P_i	10000

TABLEAU 2.1 : Concentrations des métabolites utilisés dans le modèle cinétique.
Données extraites des travaux précédents et repris dans nos travaux (Moreno-Sánchez *et al.*, 2008; Lo-Thong *et al.*, 2020). 3PG, 3-phosphoglycérate; AMP, adénosine monophosphate ; PP_i, pyrophosphate inorganique ; ATP, adénosine triphosphate ; P_i, phosphate inorganique.

Ainsi, pour la première enzyme modélisée, PGAM, l'équation utilisée est celle d'une réaction réversible de type Michaelis-Menten, avec l'ajout de l'inhibition compétitive par le pyrophosphate inorganique (PP_i) :

$$v = \frac{V_f \frac{[3PG]}{K_{m3PG}} - V_r \frac{[2PG]}{K_{m2PG}}}{1 + \frac{[3PG]}{K_{m3PG}} + \frac{[2PG]}{K_{m2PG}} + \frac{[PP_i]}{K_{iPP_i}}}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_f}$ et $\boldsymbol{V_r}$ sont les vitesses initiales maximales de l'enzyme respectivement dans le sens « forward » (vers la synthèse de 2-PG) ou « reverse » (vers la synthèse de 3-PG) ; $\boldsymbol{K_m}$ est la constante de Michaelis pour les différents substrats indiqués, $\boldsymbol{K_i}$ est la constante d'inhibition pour l'inhibiteur indiqué (PP_i et 3-PG), $[\mathbf{2PG}]$ et $[\boldsymbol{PP_i}]$ sont les concentrations respectives en 3-PG, 2-PG et PP_i.

Pour la seconde enzyme, l'énolase (ENO), l'équation utilisée est celle d'une réaction michaélienne réversible et présentant deux inhibiteurs compétitifs :

$$v = \frac{V_f \frac{[2PG]}{K_{m2PG}} - V_r \frac{[PEP]}{K_{mPEP}}}{1 + \frac{[2PG]}{K_{m2PG}} + \frac{[PEP]}{K_{mPEP}} + \frac{[PP_i]}{K_{iPP_i}} + \frac{[3PG]}{K_{i3PG}}}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_f}$ et $\boldsymbol{V_r}$ sont les vitesses initiales maximales de l'enzyme respectivement dans le sens « forward » (vers la synthèse de PEP) ou « reverse » (vers la synthèse de 2-PG) ; $\boldsymbol{K_m}$ est la constante de Michaelis pour les différents substrats indiqués, $\boldsymbol{K_i}$ est la constante d'inhibition pour les inhibiteurs indiqués (PP_i et 3-PG) et [**2PG**], [$\boldsymbol{PEP}$], [$\boldsymbol{PP_i}$] et [**3PG**] sont les concentrations respectives en 2-PG, PEP, PP_i et 3-PG.

Enfin, l'équation utilisée dans notre modèle pour PPDK est la plus simple que nous ayons trouvée pour décrire cette réaction. Cette équation illustre le mécanisme d'une enzyme ayant trois réactants au départ : l'équation Tri-Réactants (« *Ter-Reactants* »). Elle s'écrit sous la forme suivante :

$$v = \frac{V_f\left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}}\right)}{K_{mA}\mathrm{B} + K_{mB}\mathrm{A} + K_{mC}\mathrm{B} + K_{mB}\mathrm{C} + \frac{V_f K_{mQ}\mathrm{P}}{V_r K_{eq}} + \frac{V_f K_{mP}\mathrm{Q}}{V_r K_{eq}} + \frac{V_f K_{mQ}\mathrm{R}}{V_r K_{eq}} + \frac{V_f K_{mR}\mathrm{Q}}{V_r K_{eq}} + \mathrm{ABC} + \frac{V_f \mathrm{PQR}}{V_r K_{eq}}}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_f}$ et $\boldsymbol{V_r}$ sont les vitesses initiales maximales de l'enzyme respectivement dans le sens « forward » (vers la synthèse de Pyr) ou « reverse » (vers la synthèse de PEP) ; **A**, $\boldsymbol{B}$ et $\boldsymbol{C}$ sont les concentrations respectives des trois substrats PEP, AMP et PP_i ; $\boldsymbol{P}$, $\boldsymbol{Q}$ et $\boldsymbol{R}$ sont les concentrations respectives des trois produits Pyr, ATP et P_i ; $\boldsymbol{K_m}$ est la constante de Michaelis pour les différentes espèces indiquées et $\boldsymbol{K_{eq}}$ est la constante d'équilibre de la réaction.

Pour la LDH, une simple équation provenant de la loi d'action de masse est utilisée ($v = k \times [Pyr]$), où $\boldsymbol{k}$ est une constante de vitesse, $k = 2\,000\ min^{-1}$ et [$\boldsymbol{Pyr}$] est la concentration en Pyr.

Un modèle cinétique utilisant cette équation pour PPDK a été développé par l'équipe de Moreno-Sánchez *et al.* (Moreno-Sánchez *et al.*, 2008). Le modèle développé ici diffère du modèle initial avec : l'utilisation de concentrations en substrat et cosubstrat correspondant à celles au pH de 6 et en présence d'ATP et de P_i dans le tableau 2.1, ainsi que de paramètres, notamment les $\boldsymbol{V_r}$ qui ont été recalculés par nos soins (tableau 2.2), en se basant sur les proportions d'enzymes utilisées à un pH de 6 et en présence d'ATP et de P_i dans le modèle (Moreno-Sánchez *et al.*, 2008; Saavedra *et al.*, 2007).

Les paramètres cinétiques de chacune des réactions de la voie basse de la glycolyse sont regroupés dans le tableau 2.2. Ils sont déterminés expérimentalement de manière *in-vitro* ou calculés à partir de données expérimentales.

Enzyme	K_m (μM)	K_i (μM)	K_{eq}	V_{max} (mU)
PGAM	473 (3PG) 106 (2PG)	173 (PPi)		Vf = 75 Vr = 67.24
ENO	86.4 (2PG) 102 (PEP)	137 (PPi) 610 (3PG)		Vf = 328.5 Vr = 66.61
PPDK	30 (PEP) 2 (AMP) 91 (PPi) 221 (Pyr) 597 (ATP) 1342 (Pi)		0.73	Vf = 196.5 Vr = 12.28

TABLEAU 2.2 : Paramètres cinétiques utilisés dans notre modèle cinétique de la partie basse de la glycolyse.
Les constantes de Michaelis (K_m) et les constantes d'inhibition (K_i) sont en μM ; K_{eq} est la constante d'équilibre de la réaction. Les vitesses initiales maximales de l'enzyme dans le sens « *forward* » ou « *reverse* » (V_f et V_r) sont en mU. Les données sont issues d'une étude antérieure (Moreno-Sánchez *et al.*, 2008), les valeurs de V_r ont été calculées à partir des proportions d'enzymes (Saavedra *et al.*, 2005).

2.2.3. Méthodes d'optimisation du modèle cinétique

Après une comparaison des flux prédits et des flux expérimentaux, il s'est avéré que l'équation cinétique de la dernière enzyme de la voie basse de la glycolyse n'était pas adaptée pour modéliser son comportement. Par conséquent, afin d'établir un modèle plus précis de la voie basse de la glycolyse, nous opérons une modification de l'équation cinétique de PPDK.

PPDK : Utilisation d'une équation Uni Uni Bi Bi Ping-Pong (UUBB)

La première modification effectuée sur le modèle se définit par l'utilisation d'une équation différente pour PPDK, décrivant plus précisément son mécanisme d'action. Cette équation est celle d'un mécanisme Ping-Pong Uni Uni Bi Bi (ou UUBB), déterminée dans des travaux précédents (Varela-Gómez *et al.*, 2004) :

$$v = \frac{V_f V_r \left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}} \right)}{\mathrm{D}}$$

Avec au dénominateur $D = V_r K_{iB} K_C A + V_r K_C A B + V_r K_B A C + \frac{V_f}{K_{eq}} K_{iR} K_Q P + \frac{V_f}{K_{eq}} K_R P Q +$

$V_r K_{iB} \frac{K_C}{K_{iQ}} A Q + \frac{V_f}{K_{eq}} K_Q P R + \frac{V_f}{K_{eq}} K_P Q R + \frac{V_f}{K_{eq}} K_Q P R + V_r K_A B C + \frac{V_f}{K_{eq}} \frac{(K_{iR} K_Q)}{K_{iC}} C P + V_r \frac{K_c}{K_{iQ}} A B Q +$

$\frac{V_f}{K_{eq}} \frac{K_R}{K_{iA}} A P Q + \frac{V_f}{K_{eq}} \frac{K_P}{K_{CB}} B Q R + V_r \frac{K_A}{K_{iP}} A C P + V_r \frac{K_A}{K_{iR}} B C R + \frac{V_f}{K_{eq}} \frac{K_P}{K_{iB}} B Q R + V_r \frac{K_C}{(K_{iQ} K_{iiC})} A B C Q +$

$\frac{V_f}{K_{eq}} \frac{K_Q}{K_{iC}} C P R + \frac{V_f}{K_{eq}} \frac{K_Q}{(K_{iC} K_{iiC})} C P R + V_r \frac{K_A}{(K_{iR} K_{iiC})} B C Q R + V_r A B C + \frac{V_f}{K_{eq}} P Q R + \frac{V_f}{K_{eq}} \frac{(K_{iR} K_Q)}{K_{iA}} A P$

avec $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_f}$ et $\boldsymbol{V_r}$ sont les vitesses initiales maximales de l'enzyme respectivement dans le sens « forward » (vers la synthèse de Pyr) ou « reverse » (vers la synthèse de PEP) ; **A**, $\boldsymbol{B}$ et $\boldsymbol{C}$ sont les concentrations respectives des substrats PEP, AMP et PP_i ; $\boldsymbol{P}$, $\boldsymbol{Q}$ et $\boldsymbol{R}$ sont les concentrations respectives des produits Pyr, ATP et P_i ; $\boldsymbol{K_m}$ est la constante de Michaelis pour les différentes espèces indiquées, $\boldsymbol{K_{eq}}$ est la constante d'équilibre de la réaction ; $\boldsymbol{K_i}$ et $\boldsymbol{K_{ii}}$ sont respectivement la constante de dissociation du substrat/produit et la constante de l'inhibiteur qui affecte l'interception ($1/\boldsymbol{V_{max}}$).

Les paramètres utilisés dans cette équation sont dans le tableau 2.2 et le tableau 2.3 ci-dessous.

Constante	Valeur (en μM)
$\mathbf{K_{ii\_Pi}}$	7200
$\mathbf{K_{i\_Pyr}}$	2300
$\mathbf{K_{i\_Pi}}$	23000
$\mathbf{K_{i\_ATP}}$	140
$\mathbf{K_{ii\_PPi}}$[a]	1000
$\mathbf{K_{i\_PEP}}$[a]	1000
$\mathbf{K_{i\_AMP}}$[a]	1000
$\mathbf{K_{PPi\_AMP}}$[a]	1000
$\mathbf{K_{i\_PPi}}$[a]	1000

TABLEAU 2.3 : Paramètres spécifiques de l'équation UUBB pour la réaction catalysée par PPDK.

[a] Les données ont été fixées de manière arbitraire.

PPDK : Utilisation d'une équation modifiée de UUBB

Une fois l'équation cinétique de PPDK remplacée par l'équation précédente, la prochaine modification consiste en l'estimation des paramètres fixés de manière arbitraire dans l'équation UUBB. Cette estimation s'est faite par le biais de l'outil « *Parameter Estimation* » proposé par le logiciel COPASI. Cet outil permet la détermination des valeurs des paramètres cinétiques présents dans le modèle. Les nouveaux paramètres sont calculés sur la base d'un jeu de données, des résultats d'expériences en temps réel par exemple, qui est importé sur COPASI. Le logiciel ajuste alors les valeurs afin de minimiser la différence entre le modèle et les données expérimentales (Hoops *et al.*, 2006). Le descriptif de la méthode d'estimation des paramètres est faite au **chapitre 3**.

PPDK : Utilisation d'une équation Bi Bi Ping-Pong (BBPP)

Une autre équation cinétique a été également testée pour décrire la réaction de PPDK, celle d'un mécanisme de type Bi Bi Ping-Pong :

$$v = \frac{V_f\left(\mathrm{AB} - \frac{\mathrm{PQ}}{K_{eq}}\right)}{AB + K_{mB}\mathrm{A} + + K_{mA}\mathrm{B}\left(1 + \frac{Q}{K_{iQ}}\right) + \frac{V_f}{V_r K_{eq}}\left[\mathrm{K}_{mQ}\mathrm{P}\left(1 + \frac{A}{K_{iA}}\right) + Q(K_{mP} + P)\right]}$$

Où $\boldsymbol{v}$ est la vitesse de la réaction ; $\boldsymbol{V_f}$ et $\boldsymbol{V_r}$ sont les vitesses initiales maximales de l'enzyme respectivement dans le sens « forward » (vers la synthèse de Pyr) ou « reverse » (vers la synthèse de PEP) ; $\mathbf{A}$ et $\boldsymbol{B}$ sont les concentrations respectives des substrats PEP et PP_i ; $\boldsymbol{P}$ et $\boldsymbol{Q}$ sont les concentrations respectives des produits Pyr et P_i ; $\boldsymbol{K_m}$ est la constante de Michaelis pour les différentes espèces indiquées et $\boldsymbol{K_{eq}}$ est la constante d'équilibre de la réaction.

Nous utilisons les paramètres figurant dans le tableau 2.2 pour construire ce nouveau modèle.

2.2.4. Modélisation de la voie basse de la glycolyse à l'aide d'un modèle hybride

Enfin, le dernier modèle cinétique que nous développons, dans cette étude, est un modèle hybride créé à partir d'un modèle cinétique construit auparavant, où tous les paramètres cinétiques sont connus (modèle utilisant l'équation Tri-Réactants). Afin d'améliorer la prédiction de ce modèle COPASI, l'équation cinétique de PPDK est changée par celle d'une équation réversible Tri-réactants modifiée comme suit :

$$v = \frac{V_f\left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}}\right)}{K_{mA}\mathrm{B} + K_{mB}\mathrm{A} + K_{mC}\mathrm{B} + K_{mB}\mathrm{C} + \frac{V_f K_{mQ}\mathrm{P}}{V_r K_{eq}} + \frac{V_f K_{mP}\mathrm{Q}}{V_r K_{eq}} + \frac{V_f K_{mQ}\mathrm{R}}{V_r K_{eq}} + \frac{V_f K_{mR}\mathrm{Q}}{V_r K_{eq}} + \mathrm{ABC} + \frac{V_f \mathrm{PQR}}{V_r K_{eq}} + \mathrm{X}}$$

Avec le terme d'ajustement $\boldsymbol{X = \alpha \left| V_f - V_{f0} \right|}$ au dénominateur, $\boldsymbol{\alpha}$ est un nombre estimé à partir des données, $\boldsymbol{V_{f0}}$ est la vitesse initiale maximale de PPDK dans le sens « *forward* » de la reconstitution *in-vitro* et $\boldsymbol{V_f}$ la vitesse initiale maximale de PPDK dans le sens « *forward* » du modèle.

Ce modèle particulier a été construit car, bien que le modèle UUBB ait pu prédire assez bien le flux final lorsque les activités de PGAM et ENO variaient, il surestimait le flux lorsque l'activité de PPDK variait. Cependant, ce même modèle prédisait bien le flux, avec les paramètres enzymatiques utilisés dans la reconstitution *in-vitro*. Par conséquent, un terme d'ajustement a été ajouté, afin de diminuer la vitesse de PPDK avec α. Comme V_f de PPDK est égal à V_{f0} lorsque l'on fait varier l'activité de PGAM ou ENO, α est multiplié par $V_f - V_{f0}$, de sorte que le terme d'ajustement soit nul lorsque $V_f = V_{f0}$, et que les flux prédits ne soient pas modifiés lorsque l'activité de PGAM ou ENO varie. De plus, pour s'assurer que le terme d'ajustement soit positif, nous avons utilisé la valeur absolue $\left| V_f - V_{f0} \right|$.

Pour déterminer la valeur α, nous évaluons d'abord une gamme de valeurs allant de 0 à 4.10^6 avec un large pas de 10^6. La gamme et le pas utilisés sont ensuite réduits, de 10^6 à 1, jusqu'à ce que nous obtenions de meilleurs résultats en termes d'erreur quadratique moyenne (ou RMSE pour « *Root-Mean-Square Error* »), et de coefficient de détermination (R^2) entre les données prédites et les données expérimentales. Il est important de noter que ce paramètre α n'a aucune signification biologique et est déterminé par une méthode d'apprentissage basée sur les données, d'où le nom de « boîte-grise » pour ce modèle.

2.2.5. Validation des modèles cinétiques

Une fois les modèles construits, ils sont validés à l'aide de l'ensemble de données expérimentales citées au paragraphe 2.2.1. La performance de chaque modèle est alors évaluée en fonction du RMSE et R^2 déterminés. L'équation pour le calcul du RMSE est la suivante :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}$$

Où $\boldsymbol{Y_i}$ et $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ sont respectivement les valeurs observées et les valeurs prédites, **n** est le nombre total de valeurs et **i** = 1, 2…n.

L'équation pour déterminer le R^2 est :

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Où $\boldsymbol{Y_i}$ et $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ sont respectivement les valeurs observées et les valeurs prédites, **n** est le nombre total de valeurs et **i** = 1, 2…n.

2.3. Résultats et discussion

2.3.1. Prédiction du flux final de la voie de la glycolyse par les modèles cinétiques

Dans la section qui suit, plusieurs modèles cinétiques de la voie basse de la glycolyse d'*E. histolytica* ont été développés. Chacun de ces modèles est réalisé grâce à différentes données expérimentales issues de la littérature et sera utilisé pour la prédiction du flux en sortie de ce segment de voie métabolique.

Modèle cinétique utilisant l'équation Tri-réactants pour la réaction de PPDK

Lors de notre étude bibliographique sur la voie de la glycolyse chez ce parasite, il nous est apparu que PPDK était l'une des enzymes les plus épineuses à modéliser, en raison de son mécanisme réactionnel complexe (Varela-Gómez *et al.*, 2004). De même, face à ce genre de situation, il est admis en modélisation d'appliquer de préférence l'équation la plus simple pour décrire le mécanisme réactionnel d'une enzyme. Pour PPDK, il a donc été convenu d'utiliser l'équation Tri-réactants, où chaque paramètre est connu et dont aucun n'est défini de manière arbitraire. Nous l'avons également mentionné plus tôt, cette équation fut utilisée par une étude de Moreno-Sánchez *et al.* (Moreno-Sánchez *et al.*, 2008), à la seule différence que nous appliquons cette fois au modèle des paramètres et des concentrations en substrat et en cosubstrat utilisés lors des expériences *in-vitro*. Les résultats obtenus après une simulation d'une heure sont présentés à la figure 2.3.

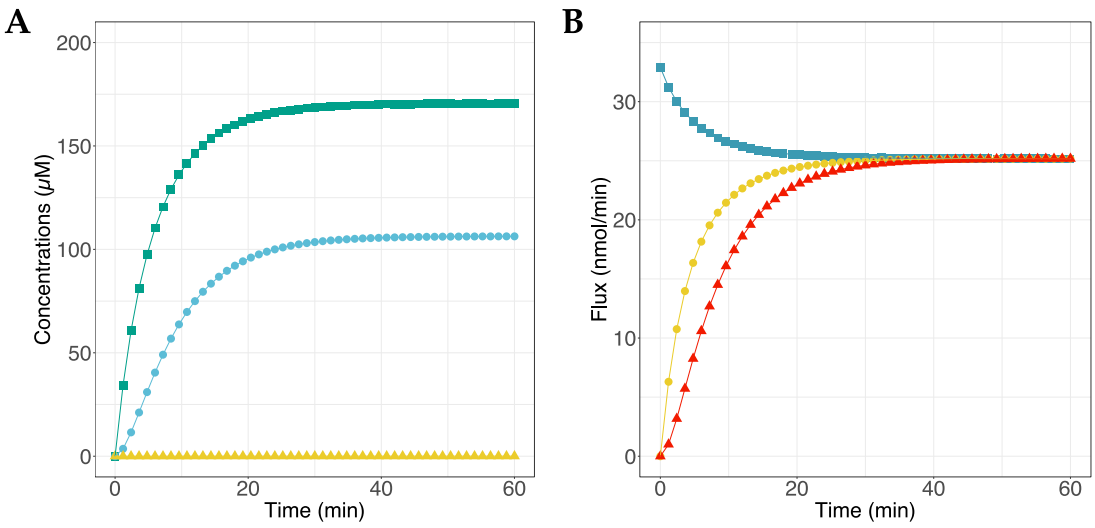

FIGURE 2.3 : *Prédictions des concentrations en métabolites et des flux par le modèle utilisant l'équation Tri-Réactants.*

(A) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles) et en jaune pour Pyr (triangles).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles) et en rouge pour PPDK (triangles).*

Nous observons que les concentrations de métabolites prédites sont supérieures à celles mesurées pendant la reconstitution *in-vitro* : 2-PG≈ 170.45 μM (2-PG$_{exp}$ = 58 ± 29 μM), PEP≈ 106.3 μM (PEP$_{exp}$ = 37 ± 16 μM) et Pyr≈ 1,26.10^{-2} μM[3] (figure 2.3A). Il semble que le modèle surestime les concentrations de 2-PG et PEP, nous ne pouvons statuer sur la prédiction de la concentration en Pyr car celle-ci n'a pas été mesurée lors des manipulations expérimentales. Alors que cette voie métabolique est considérée comme l'une des principales voies synthétisant de l'énergie chez ce parasite, nous sommes étonnés de constater la faible concentration en pyruvate prédite par notre modèle, cette concentration pouvant s'élever jusqu'à 450 μM dans des extraits cellulaires de trophozoïtes (Varela-Gómez *et al.*, 2004).

Cette différence majeure pourrait s'expliquer par le fait qu'une seule partie de la voie et du métabolisme d'*E. histolytica* est modélisée dans ce modèle UUBB. En effet, il a été démontré que la réaction catalysée par PPDK n'était pas une spontanée, avec un ΔG^0 = +0.19 kcal/mol (Varela-Gómez *et al.*, 2004). Comme le conclut cette étude, la réaction dans les conditions physiologiques des trophozoïtes n'est pas proche de son équilibre thermodynamique. Ainsi, la synthèse « totale » d'ATP par cette enzyme ne serait donc pas atteinte. Cela serait possible dans le cas où l'on observe une accumulation des substrats et une diminution des produits dans le parasite. L'ajout de certaines enzymes, qui remplissent ces conditions, dans le système pourrait alors conduire à la synthèse nette d'ATP par PPDK.

Pour ce qui est de la variable d'intérêt : le flux final de la voie, nous analysons la prédiction des flux de la glycolyse en figure 2.3 B. Ce flux est estimé à 25.17 nmol·min^{-1} par notre modèle, nous constatons que cette valeur est proche de la valeur observée qui est de 27 nmol·min^{-1}. Aussi, il est important de considérer que le modèle a réussi à mimer l'état quasi-stationnaire du système en *in-vitro*, étant donné la convergence des courbes de flux de chaque enzyme au bout d'une heure de simulation. Afin d'évaluer la performance du modèle, nous l'utilisons pour estimer un ensemble de données expérimentales (figure 2.4). Notre modèle prédit globalement bien les flux de la partie basse de la glycolyse à partir de l'activité des enzymes contenues dans ce segment, avec R^2 = 0.88 et RMSE = 3.39 nmol·min^{-1}. Cependant, si nous regardons une seule variation d'activité à la fois _ autrement dit une seule couleur à la fois sur la figure 2.4 _ nous remarquons que ce modèle prédit

[3]La concentration en pyruvate (Pyr) n'a pas été mesurée expérimentalement.

mieux le flux final lorsque l'activité de PGAM ou celle de ENO varie, que lorsqu'il s'agit de celle de PPDK.

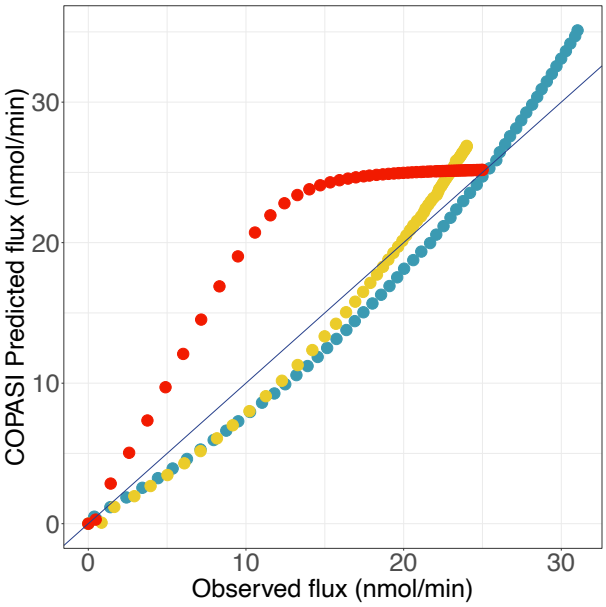

FIGURE 2.4 : *Prédictions des flux par le modèle Tri-Réactants à partir des données expérimentales in-vitro. Les couleurs correspondent aux variations de l'activité de chaque enzyme : PGAM en bleu, ENO en jaune et PPDK en rouge.*

Ces résultats nous suggèrent que l'équation utilisée pour la réaction catalysée par PPDK n'est pas adaptée. Nous entamons donc la recherche approfondie d'une équation appropriée pour l'enzyme PPDK.

Modèle cinétique utilisant l'équation UUBB pour la réaction de PPDK

Le second modèle construit par nos soins est un modèle cinétique, où la réaction de PPDK est modélisée par une équation de type UUBB (modèle UUBB). Lorsque nous comparons les résultats obtenus avec ce modèle et en reconstitution *in-vitro* (Moreno-Sánchez et al., 2008), nous observons que ce modèle surestime toujours les concentrations des métabolites : 2-PG ≈ 188.65 μM, PEP ≈ 256.43 μM ; et celle du produit final est de Pyr ≈ $1{,}12.10^{-2}$ μM (figure 2.5 A).

En ce qui concerne le flux final de la voie, qui correspond à celui de la dernière réaction (PPDK), il est de 22.36 nmol·min^{-1} pour le modèle cinétique UUBB (figure 2.5 B) et de 27 ± 7 nmol·min^{-1} pour la valeur mesurée expérimentalement. Il est notable que les flux convergent plus ou moins vers la même valeur, ce qui nous indique que ce modèle atteint toujours un état quasi-stationnaire. Contrairement aux prédictions de concentration, nous voyons que la différence est moins importante entre le flux prédit et le flux observé. Si nous comparons ces résultats à ceux obtenus

par le premier modèle (modèle Tri-Réactants), nous remarquons que sa capacité à prédire des variables de sortie (concentrations, flux) après un temps donné est amoindrie.

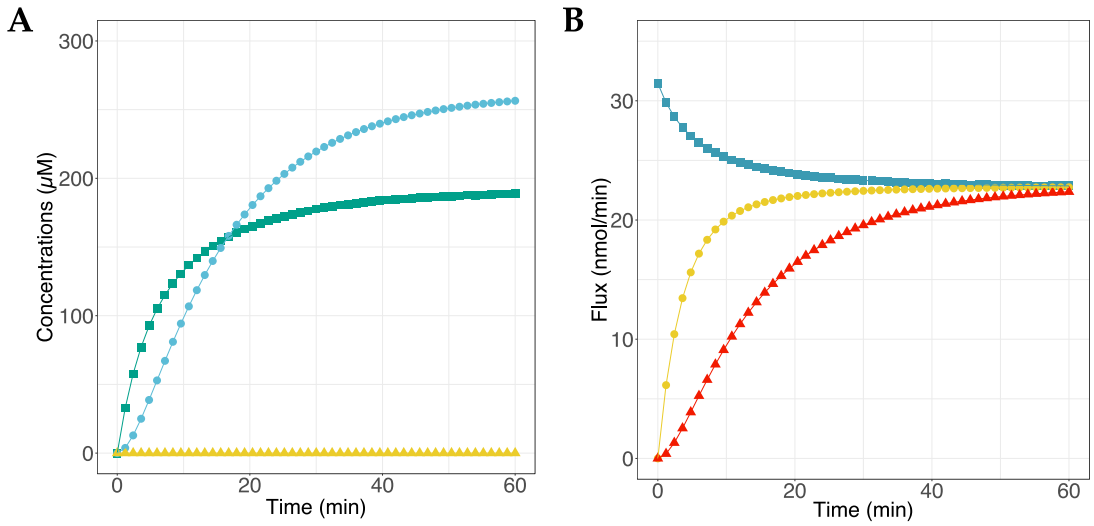

FIGURE 2.5 : *Prédictions des concentrations en métabolites et des flux par le modèle UUBB.*
(A) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles) et en jaune pour Pyr (triangles).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles) et en rouge pour PPDK (triangles).*

Toutefois, pour valider le modèle, nous évaluons sa capacité à prédire l'ensemble de données expérimentales. Il s'en dégage que le modèle UUBB prédit assez bien le flux final de la voie (R^2 = 0.94 et RMSE = 2.43 nmol·min^{-1}), notamment lorsque l'activité des enzymes PGAM ou ENO subit des variations (figure 2.6). Lorsque l'activité de PPDK varie, nous apercevons un léger déclin des flux prédits à partir de 22 nmol·min^{-1}. Les indicateurs statistiques de performance que nous utilisons nous révèlent effectivement que le modèle est plus efficace pour prédire les flux lorsque nous faisons varier l'activité de PGAM (R^2 = 0.98 et RMSE = 2.96 nmol·min^{-1}), puis celles de ENO (R^2 = 0.97 et RMSE = 2.3 nmol·min^{-1}) et enfin celle PPDK (R^2 = 0.94 et RMSE = 1.95 nmol·min^{-1}).

Nous en concluons que pour parvenir à un modèle efficace, il est non seulement nécessaire d'avoir des équations cinétiques précises et capables de mimer le mécanisme de la réaction enzymatique, mais aussi d'avoir des paramètres appropriés. Il est donc indispensable d'effectuer des modifications au modèle présent, afin d'obtenir de meilleures prédictions, tant en termes de concentration que de flux.

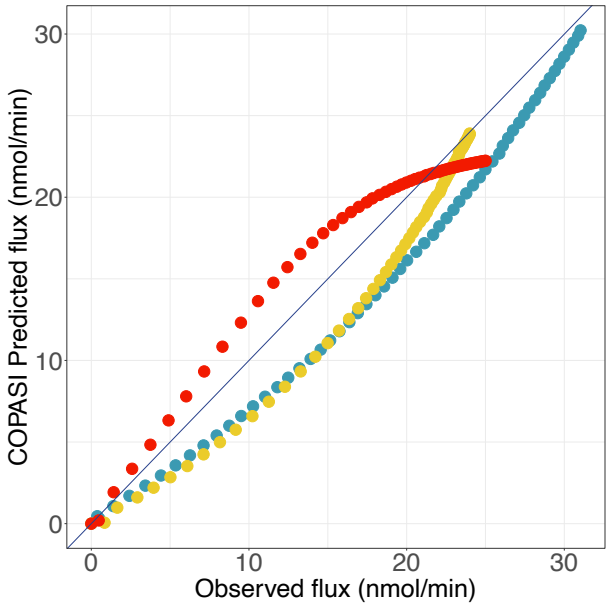

FIGURE 2.6 : *Prédictions des flux par le modèle UUBB à partir des données expérimentales in-vitro. Les couleurs correspondent aux variations de l'activité de chaque enzyme : PGAM en bleu, ENO en jaune et PPDK en rouge.*

Modèle cinétique utilisant l'équation UUBB modifiée pour la réaction de PPDK

L'équation UUBB, telle que présentée dans la partie Méthodes, comporte un grand nombre de paramètres cinétiques fixés de manière arbitraire (tableau 2.3). Dans un premier temps, nous essayons d'améliorer le modèle en faisant une estimation des cinq paramètres non définis, à l'aide de l'outil d'estimation de paramètres de COPASI. Les meilleurs résultats sont obtenus par la méthode « *Particle Swarm* » ou essaims particulaires ; et les nouvelles valeurs estimées sont présentées dans le tableau 2.4.

Constante	Valeur du modèle UUBB (en μM)	Valeur estimée par COPASI (en μM)
K_{ii_PPi}	1000	9,999,100
K_{i_PEP}	1000	9,933,660
K_{i_AMP}	1000	2,224.9
K_{PPi_AMP}	1000	9,999,870
K_{i_PPi}	1000	9,948,210

TABLEAU 2.4 : Estimation des paramètres de l'équation UUBB pour la réaction catalysée par PPDK.

Une fois la modification effectuée, une simulation d'une heure est faite afin d'évaluer la performance de ce nouveau modèle. Les résultats sont présentés en figure 2.7. Pour ce nouveau modèle, nous constatons que les valeurs prédites pour les concentrations des métabolites 2–PG et PEP sont toujours supérieures à celles mesurées en *in-vitro* (figure 2.7 B). Quant au flux final, il augmente faiblement par rapport au modèle UUBB initial (~22.9 nmol·mol^{-1}), mais il reste très inférieur au flux expérimental pour les conditions utilisées lors de la reconstitution de la partie basse de la glycolyse (figure 2.7 B).

Nous en concluons que l'estimation des paramètres de l'équation UUBB n'a pas été fructueuse, au vu grand du nombre de variables à estimer, sans compter la faible quantité de données expérimentales en notre possession. Nous pouvons également nous interroger sur le type de données à utiliser lors d'une estimation de paramètres. Alors qu'il est conseillé sur COPASI d'ajuster les paramètres avec des données provenant d'expériences en temps réel (pendant une durée donnée) ou menant à un état d'équilibre (Hoops *et al.*, 2006) ; les données que nous avons utilisées ici sont tout autre puisqu'il s'agit de notre ensemble de données présenté dans la partie Méthodes. Nous prenons donc la décision d'employer une équation cinétique avec moins de paramètres non déterminés pour PPDK.

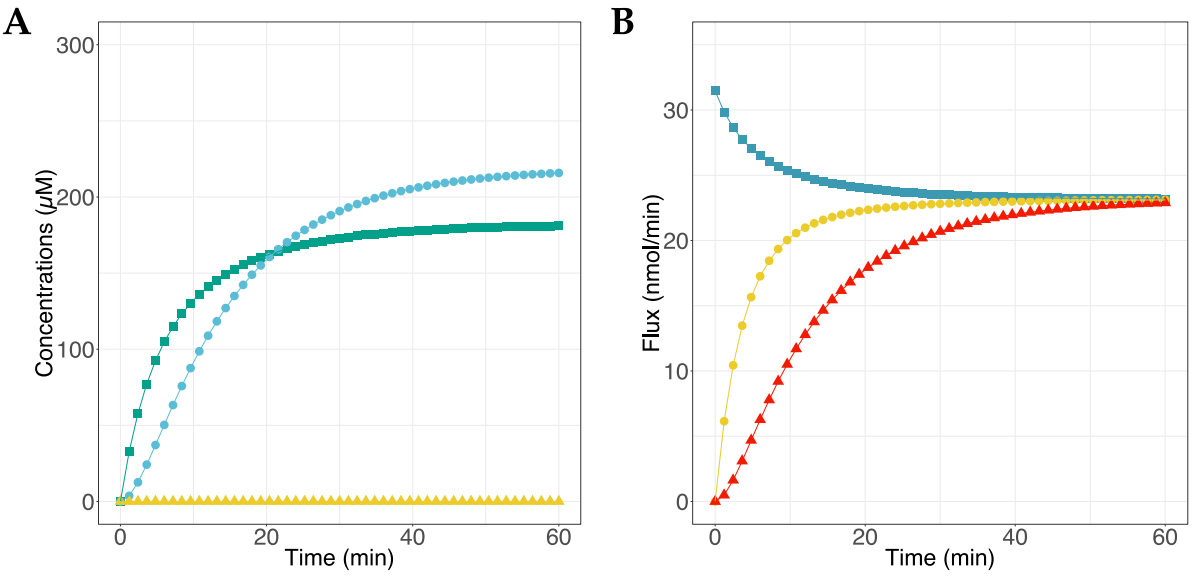

FIGURE 2.7 : *Prédictions des concentrations en métabolites et des flux par le modèle UUBB amélioré.* ***(A)*** *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles) et en jaune pour Pyr (triangles).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles) et en rouge pour PPDK (triangles).*

Par ailleurs, même si les améliorations sont moindres au niveau du flux et des concentrations prédits dans les conditions utilisées en *in-vitro* ; nous observons une optimisation des prédictions

des données expérimentales par ce nouveau modèle UUBB (figure 2.8). En effet, nous notons une diminution de la valeur de RMSE (toutes données confondues) comparée à celle du modèle précédent et une légère amélioration du R^2, avec : $R^2 = 0.95$ et RMSE = 2.06 nmol·min^{-1}.

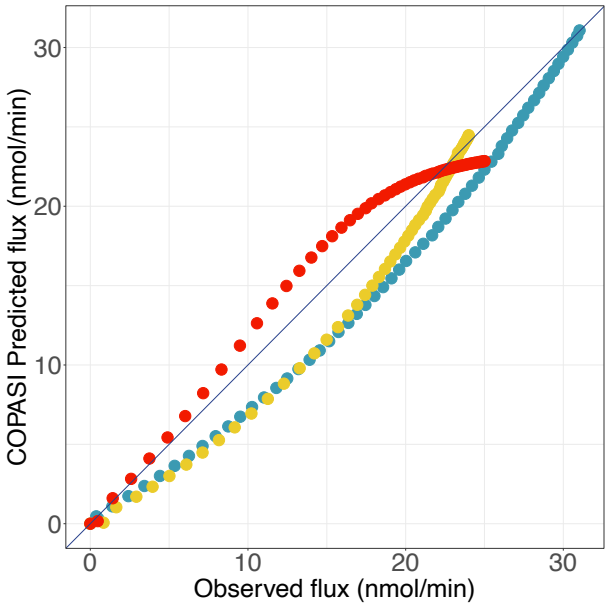

FIGURE 2.8 : *Prédictions des flux par le modèle UUBB amélioré à partir des données expérimentales in-vitro.*
Les couleurs correspondent aux variations de l'activité de chaque enzyme : PGAM en bleu, ENO en jaune et PPDK en rouge.

Modèle cinétique utilisant l'équation BBPP pour la réaction de PPDK

Comme il l'a été prescrit dans le paragraphe ci-dessus, nous avons échangé l'équation UUBB de PPDK par une équation plus simple, celle illustrant le mécanisme Bi Bi Ping-Pong (BBPP). Cette équation ne contient que deux termes non déterminés de manière expérimentale (K_{i_PEP} et K_{i_Pi}) et dont la valeur a été fixée arbitrairement à 1000 μM. Nous remarquons que l'utilisation de cette nouvelle équation semble améliorer les prédictions des concentrations et du flux (figure 2.9).

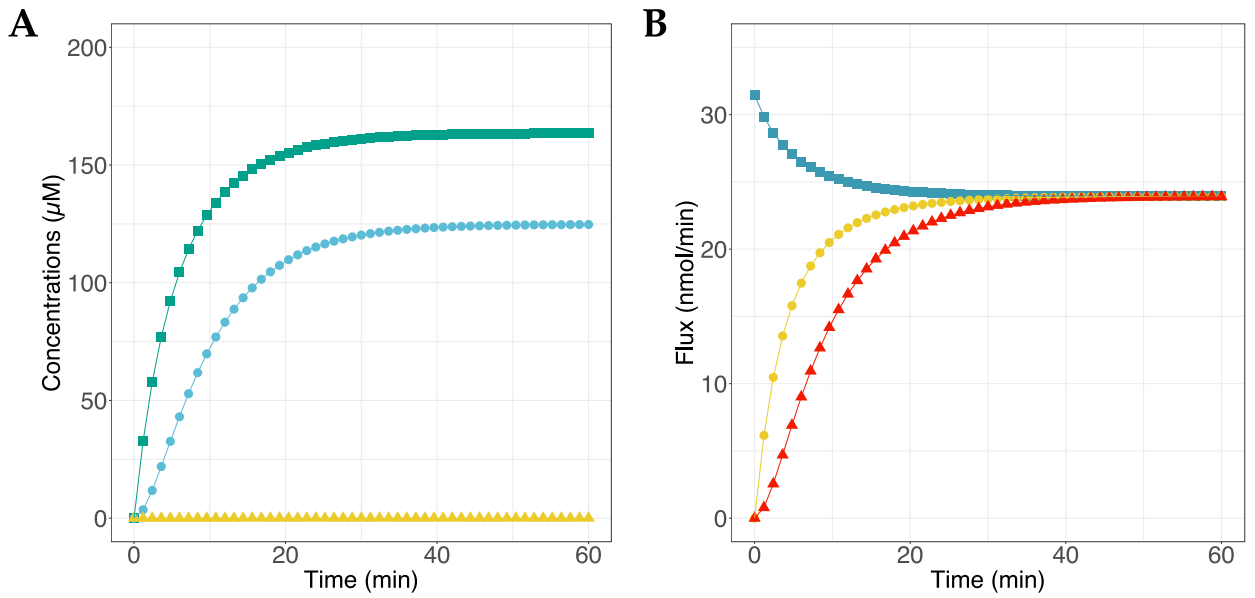

FIGURE 2.9 : *Prédictions des concentrations en métabolites et des flux par le modèle BBPP.*
(A) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles) et en jaune pour Pyr (triangles).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles) et en rouge pour PPDK (triangles).*

La concentration prédite en métabolite est celle qui se rapproche le plus de la valeur expérimentale comparée à celles des modèles précédents avec 2-PG ≈ 163.5 μM (figure 2.9 A). Ce n'est pas le cas de la concentration pour PEP qui est à 124.73 μM. Pour ce qui est de l'estimation du flux, le flux final (23.9 nmol·min^{-1}) est plus élevé que celui prédit par les modèles UUBB, mais il ne surpasse pas celui prédit par notre premier modèle Tri-Réactants (figure 2.9 B). Nous en concluons que la réduction du nombre de paramètres non déterminés ne suffit pas pour incliner notre modèle vers de meilleures prédictions.

L'ensemble des données expérimentales est une dernière fois prédit par ce modèle BBPP et les résultats sont fournis en figure 2.10. Nous rapportons pour l'ensemble des données un R^2 de 0.89 et un RMSE de 2.74 nmol·min^{-1}. Si ce modèle surpasse celui utilisant l'équation Tri-Réactants (voir la partie « *Modèle cinétique utilisant l'équation Tri-réactants pour la réaction de PPDK* »), il ne devient néanmoins pas meilleur que les modèles UUBB. Si, à première vue, cela peut être déstabilisant qu'un modèle comportant de nombreuses valeurs fixées arbitrairement présente de meilleurs résultats qu'un modèle où chaque paramètre est défini ; nous pouvons faire l'hypothèse que ces équations cinétiques complexes utilisées illustrent mieux le mécanisme de PPDK. Cela passe par exemple par l'intégration d'acteurs ou de régulations supplémentaires dans l'équation, à savoir : l'inhibition provoquée par certains produits de la réaction, un ordre particulier dans le mécanisme d'action.

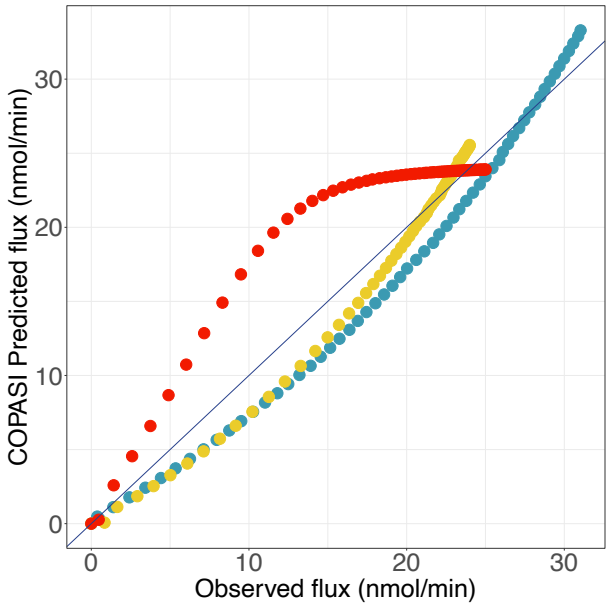

FIGURE 2.10 : *Prédictions des flux par le modèle BBPP à partir des données expérimentales in-vitro. Les couleurs correspondent aux variations de l'activité de chaque enzyme : PGAM en bleu, ENO en jaune et PPDK en rouge.*

Ainsi, nous prenons la décision de bâtir un modèle particulier contenant une équation cinétique pour PPDK où tous les paramètres ont été définis et où la partie du mécanisme réactionnel que nous ne maîtrisons pas encore est représentée par un terme d'ajustement.

2.3.2. Amélioration de la prédiction du flux par le modèle hybride boîte-grise

Lors de la construction du modèle cinétique pour la voie métabolique étudiée, nous avons constaté que le mécanisme réactionnel de PPDK était complexe à modéliser (Varela-Gómez *et al.*, 2004) et qu'il impliquait l'emploi d'équation cinétique tout aussi complexe. C'est pourquoi nous avons élaboré une équation qui réunit :

- Les avantages d'un modèle de connaissance de type « boîte-blanche » : avec des paramètres cinétiques connus, afin de prédire la partie du mécanisme de PPDK que nous pouvons définir ;
- Les avantages d'un modèle de type « boîte-noire » : où la compréhension des mécanismes n'est par essentielle, avec l'ajout d'un terme d'ajustement défini dans la partie Méthodes.

Nous lui donnons donc le nom de : modèle « boîte-grise ».

Tout d'abord, nous déterminons la valeur du paramètre $\boldsymbol{\alpha}$ contenu dans le terme d'ajustement $\boldsymbol{X = \alpha \left| V_f - V_{f0} \right|}$ de l'équation cinétique de PPDK. Afin de déterminer la valeur optimale de $\boldsymbol{\alpha}$, nous testons une gamme de valeurs allant de 0 à 5.10^6 et identifions la meilleure valeur autour de $3{,}09.10^6$; en-dessous de cette valeur, le flux est surestimé, et au-dessus, le flux final est sous-estimé (figure 2.11).

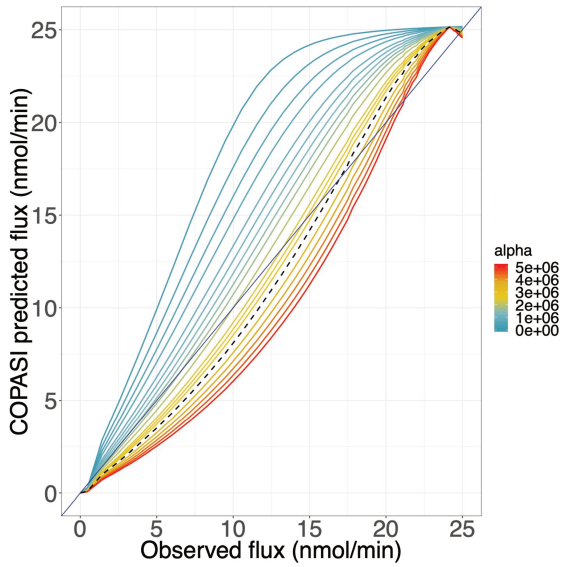

FIGURE 2.11 : *Prédictions du flux par le modèle boîte-grise lorsque l'activité de PPDK varie. La ligne pointillée représente la courbe obtenue avec le meilleur terme d'ajustement pour (3,008,970).*

Aussi, lorsque nous faisons varier le terme α, nous observons que cela n'a aucune influence sur la prédiction du flux lorsque les activités enzymatiques de PGAM et de ENO sont soumises à des variations (figure 2.12). Cela nous conforte dans notre choix de l'expression du terme d'ajustement, terme qui influence bien la prédiction du flux lorsque l'activité de PPDK varie, mais qui ne change pas la performance du modèle à prédire le flux quand les activités de PGAM et ENO varient.

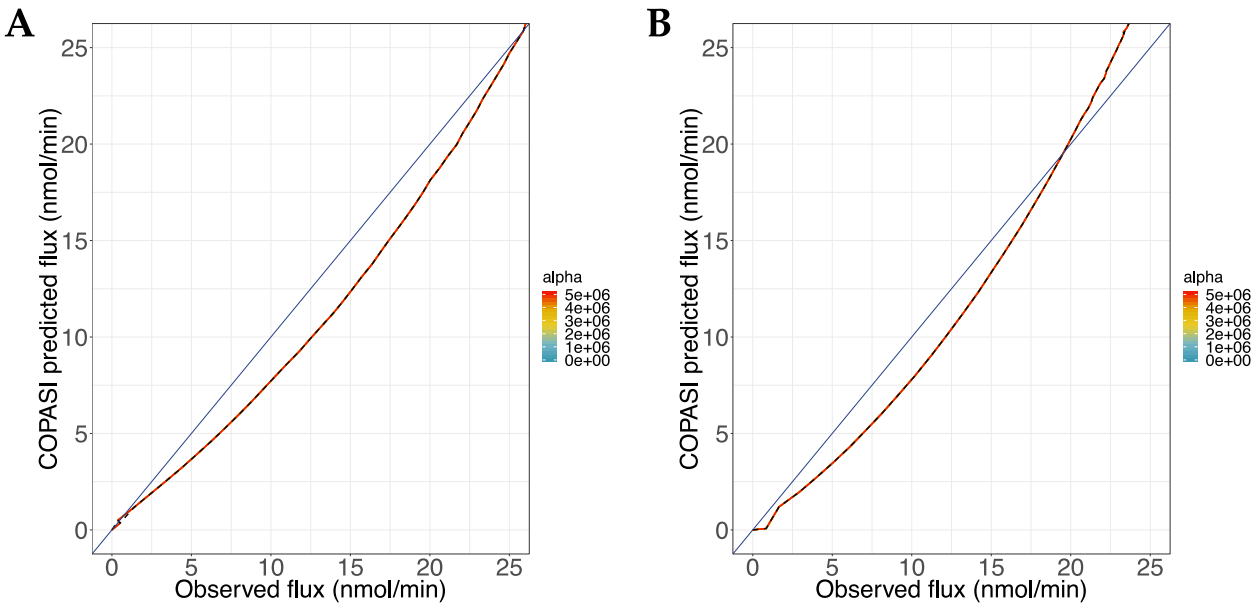

FIGURE 2.12 : *Effet de la variation du terme* α *dans le terme d'ajustement du modèle COPASI.*
(A) *Prédictions du flux par le modèle lorsque l'activité de PGAM varie.* ***(B)*** *Prédictions du flux par le modèle lorsque l'activité de ENO varie. La ligne pointillée représente la courbe obtenue avec le meilleur terme d'ajustement pour (3,008,970).*

La prochaine étape consiste à analyser la capacité du modèle à prédire les concentrations des métabolites ainsi que le flux final au bout de 60 min (figure 2.13). Nous observons que les concentrations de métabolites sont similaires à celles obtenues par le modèle Tri-Réactants (figure 2.13 A). En ce qui concerne les flux, ils sont aux alentours de 25 nmol·min^{-1}, tout comme le modèle Tri-Réactants développé au début de ces travaux (figure 2.13 B).

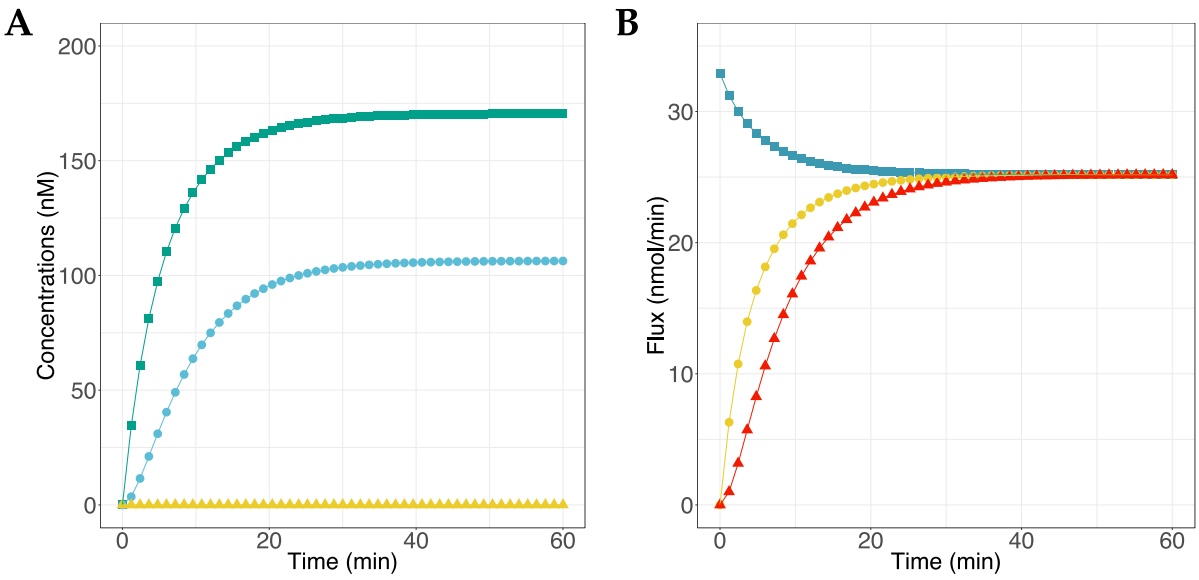

FIGURE 2.13 : *Prédictions des concentrations en métabolites et des flux par le modèle boîte-grise.*

(A) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles) et en jaune pour Pyr (triangles).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles) et en rouge pour PPDK (triangles).*

Si le modèle boîte-grise présente des difficultés à prédire les concentrations et les flux en conditions utilisées pendant la reconstitution *in-vitro*, ce n'est pas le cas lorsque l'on fait varier les différentes activités enzymatiques. De manière remarquable, une amélioration significative des prédictions de flux a été obtenue, notamment lorsque l'activité de PPDK varie, par rapport à l'ensemble des autres modèles construits auparavant (figure 2.14). Collectivement, ces résultats valident l'utilisation du terme d'ajustement dans l'équation cinétique pour améliorer un modèle d'une voie métabolique construit avec COPASI.

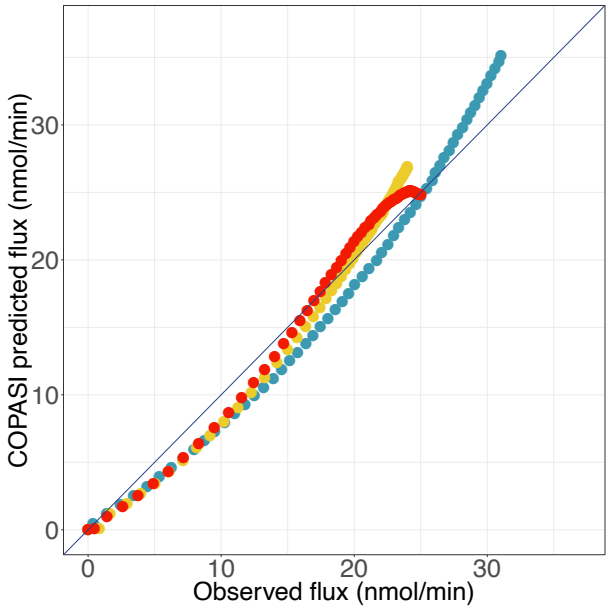

FIGURE 2.14 : *Prédictions des flux par le modèle boîte-grise à partir des données expérimentales in-vitro. Les couleurs correspondent aux variations de l'activité de chaque enzyme : PGAM en bleu, ENO en jaune et PPDK en rouge.*

2.3.3. Comparaison des performances des différents modèles

Lors de la construction des modèles, nous nous sommes attachés à :

- Prédire les sorties du système (concentrations des métabolites et flux) au bout d'une heure de simulation, en utilisant les conditions utilisées lors de la reconstitution *in-vitro* ;
- Prédire le flux final d'un ensemble de données expérimentales constitué d'activités enzymatiques.

Pour le premier point, nous remarquons que les modèles cinétiques ne parviennent pas à estimer les concentrations de 2-PG et PEP retrouvées en *in-vitro* à l'état quasi-stationnaire. Les estimations se rapprochant le plus des valeurs expérimentales sont celles obtenues par le modèle Tri-Réactants. Pour le flux final, la valeur la plus proche est celle prédite par le modèle Tri-Réactants et le modèle boîte-grise (~25 nmol·min^{-1}). Ces résultats peuvent être dus à des régulations de la glycolyse chez ce parasite qui n'apparaissent pas dans les modèles. Néanmoins, comme le flux est assez bien prédit après une heure, nous pouvons également faire l'hypothèse que nos modèles utilisent des paramétrages qui ne permettent pas la bonne prédiction des concentrations (par exemple : fixation des valeurs de substrats/cosubstrats).

Si nous nous intéressons à la prédiction des données expérimentales, nous distinguons globalement une meilleure performance du modèle boîte-grise par rapport aux autres modèles ; et ce quelle que soit l'enzyme dont l'activité varie (tableau 2.5). Ces résultats nous montrent aussi que la performance du modèle boîte-grise est semblable à celle du modèle Tri-Réactants. Pourtant, une observation plus attentive nous révèle une amélioration des prédictions pour le modèle boîte-grise lorsque l'activité de PPDK est variée (R^2 = 0.99 et RMSE = 1.22 nmol·min^{-1}).

Modèle	R^2	RMSE	PGAM		ENO		PPDK	
			R^2	RMSE	R^2	RMSE	R^2	RMSE
Tri-Réactants	0.88	3.39	0.98	2.02	0.98	1.78	0.8	5.27
UUBB	0.94	2.43	0.98	2.96	0.97	2.3	0.94	1.94
UUBB améliorée	0.95	2.06	0.98	2.59	0.98	1.89	0.96	1.6
BBPP	0.89	2.74	0.98	2.16	0.98	1.43	0.83	4.02
Boîte-grise	**0.98**	**1.71**	**0.98**	**2.02**	**0.98**	**1.78**	**0.99**	**1.22**

TABLEAU 2.5 : Comparaison des métriques statistiques pour chaque modèle bâti dans ce chapitre. RMSE est exprimée en nmol·min^{-1}. Les deux premières colonnes correspondent aux calculs sur l'ensemble du jeu de données, les suivantes correspondent aux métriques calculées sur la partie du jeu de données où l'on fait varier l'activité de PGAM, de ENO ou de PPDK.

Pour les autres modèles intégrant des valeurs fixées de manière arbitraire (UUBB, UUBB améliorée et BBPP), ils présentent de meilleurs résultats que le premier modèle bâti. De plus, si l'utilisation de ces équations améliore les prédictions lorsque l'activité de PPDK varie, ce n'est pas le cas pour les autres enzymes ; pour lesquelles nous notons une augmentation du RMSE. Cela fait ressortir l'un des importants avantages du modèle boîte-grise : le renforcement du modèle initial,

où la performance était moindre, à l'aide de données expérimentales, sans aucune altération sur la performance du modèle pour le reste des données (quand l'activité de l'enzyme PGAM ou celle de l'enzyme PPDK est variée).

2.4. Conclusion

Les modèles cinétiques, s'ils sont bien construits, peuvent être de considérables outils pour la compréhension des mécanismes au sein d'un organisme (Stanford *et al.*, 2013; Kerkhoven *et al.*, 2014). L'un des organismes phares dans ce domaine est *Saccharomyces cerevisiae*, une levure déjà utilisée en expérimental pour comprendre la biologie des systèmes eucaryotes (Botstein and Fink, 1988) et en industrie pour produire des molécules (Sulieman *et al.*, 2018). Aussi dans cette étude, nous avons construit des modèles cinétiques représentant la partie basse de la glycolyse chez *Entamoeba histolytica*. Ces modèles ont nécessité la connaissance et l'utilisation : des concentrations initiales de substrats et cosubstrats, des paramètres cinétiques de chaque enzyme (PGAM, ENO et PPDK) et des équations cinétiques précises pour chacune d'entre elles.

Au vu des résultats peu concluants du premier modèle, plusieurs modèles ont été construits avec un objectif précis : modifier l'équation de PPDK afin d'améliorer les prédictions générales du système. Les équations utilisées augmentaient effectivement la performance du modèle, surtout pour la prédiction des flux lorsque l'activité de PPDK était variée.

Mais ils présentent également deux limites :

- L'utilisation d'équations contenant plusieurs paramètres définis de manière arbitraire et difficiles à estimer expérimentalement et,
- L'augmentation du RMSE des modèles lorsque les deux autres activités enzymatiques étaient variées (PGAM ou ENO).

Après ces tentatives d'amélioration du modèle initial, un nouveau modèle a vu le jour : le modèle boîte-grise. Ce dernier contient en majeure partie des équations cinétiques ainsi qu'une partie boîte-noire avec l'ajout d'un terme d'ajustement, dont le réglage se fait à partir de l'ensemble de données expérimentales. Le terme d'ajustement a été ajouté sur le modèle où tous les paramètres cinétiques étaient définis et il est ajusté selon les résultats obtenus par ce même modèle.

Malgré la difficulté que présentent ces modèles à prédire les concentrations des métabolites dans des conditions données, nous avons pu valider leur utilisation pour la prédiction du flux final de la voie métabolique étudiée, grâce aux valeurs de R^2 et RMSE calculées entre les valeurs prédites et observées du jeu de données expérimentales. De nouvelles possibilités d'optimisation des modèles cinétiques peuvent être envisagées pour la prédiction des concentrations : la mesure des paramètres cinétiques manquants de PPDK, l'obtention d'un plus large ensemble de données

expérimentales contenant cette fois des mesures des concentrations en métabolites ou en produit final. Nous pouvons également envisager l'utilisation de techniques différentes pour estimer les paramètres manquants (Kim *et al.*, 2018; Matera *et al.*, 2019).

2.5. Discussion et conclusion du chapitre

Ce second chapitre a été consacré à l'élaboration d'un modèle cinétique de la voie basse de la glycolyse chez *E. histolytica* pour prédire des informations importantes, telles que le flux final de la voie. Ce type de modèle, comme nous l'avons détaillé au **chapitre 1**, est un système très détaillé qui intègre des informations sur chaque acteur participant à la production de la molécule finale.

Un modèle initial a été créé et intégrait des équations cinétiques simples pour la description de chaque réaction enzymatique. Ce modèle a été amélioré grâce à différentes modifications opérées sur l'une des équations cinétiques intégrées au modèle (celle de PPDK). À chaque modification effectuée, une validation du modèle était opérée par le biais de la prédiction d'un jeu de données expérimentales. Suite à la modélisation de trois modèles contenant des paramètres dont la valeur avait été fixée par arbitrairement, nous nous sommes décidés à revenir à un modèle avec des équations dont les paramètres avaient déjà été définis et nous y avons rajouté un terme d'ajustement.

L'ajout de ce terme d'ajustement dans l'équation cinétique de PPDK a permis l'amélioration des prédictions du flux à partir des activités enzymatiques. En effet, le mécanisme d'action étant complexe _ et nous avons insisté sur ce point pendant ce chapitre _ nous avons gardé la plus simple équation définissant ce mécanisme, où les paramètres étaient connus et définis, et nous y avons rajouté un terme d'ajustement qui illustrait la partie des régulations qui n'étaient pas représentées dans l'équation de départ. Nous pouvons penser également que cet ajout permettrait, dans des cas où nous voudrions modéliser une seule voie métabolique en condition *in-vivo*, de représenter les régulations qui sont provoquées par d'autres molécules présentes dans l'organisme ; qui ne le sont pas évidemment si nous nous plaçons en condition *in-vitro*, comme dans la présente étude. Ainsi est née une méthode hybride de modélisation de voies métaboliques : le modèle boîte-grise. Ce modèle peut être appliqué à d'autres voies, en particulier lorsque ces voies contiennent des réactions qui ne sont pas encore caractérisées, ou pour lesquelles les équations cinétiques existantes ne suffisent pas à modéliser correctement la réaction catalysée. Ces premiers résultats enrichissent le panel de méthodes existant pour la modélisation des voies, même si des progrès sont encore à fournir pour la prédiction des concentrations.

De plus, ces travaux ont mis en lumière les difficultés possibles rencontrées lors de la modélisation d'une voie métabolique : une connaissance incomplète de la voie métabolique étudiée, de nombreux paramètres cinétiques non définis ou encore des mécanismes réactionnels complexes à modéliser. Dans notre application, lors de la modélisation de la glycolyse, nous avons

utilisé une équation complexe avec des paramètres non définis, que nous avons essayé de modéliser. Cette équation (UUBB) avait l'avantage de représenter au mieux le mécanisme réactionnel de l'enzyme PPDK (Varela-Gómez *et al.*, 2004). Néanmoins, nous n'avons pas pu obtenir une estimation satisfaisante des paramètres manquants. Cela pourrait être expliqué par plusieurs raisons : par le nombre important de paramètres à estimer, la nature et le nombre des données expérimentales que nous avons utilisées pour l'estimation de ces paramètres. Ce sont les raisons les plus évidentes que nous avons identifiées lors de ces travaux. Les travaux de G. Jia *et al.* évoquent ces raisons parmi les principales causes expliquant l'échec fréquent des techniques d'estimation de paramètres, à l'instar de : l'absence de données complètes, la mauvaise qualité des données et la difficulté de calcul pour résoudre les équations du modèle et les problèmes d'optimisation (Jia *et al.*, 2011). Un autre problème considéré dans cette étude est le manque d'identifiabilité complète des paramètres ; pour G. Jia *et al.*, seul un sous-ensemble de paramètres peut être déterminé à partir des données. L'utilisation de modèles différents qui n'utilisent pas les paramètres cinétiques des enzymes, pour la construction de cette voie, pourrait donner de meilleures prédictions. Nous envisagerons cette possibilité dans le **chapitre 3**.

En conclusion, l'étude d'une telle voie a permis la construction, pour la première fois, d'un modèle hybride se basant sur la méthode type modèle cinétique et intégrant une partie boîte-noire, dont l'apprentissage se fait sur un ensemble de données observées. Par ailleurs, nous pouvons extraire de ces expériences une méthodologie pour toute modélisation de voie métabolique, composée comme suit :

- Recueil d'informations à propos de la voie métabolique ;
- Première modélisation par un modèle simple ;
- Ajout de complexité dans le modèle (équation cinétique plus complexe, intervention d'inhibiteurs ou d'activateurs) ;
- Création de termes d'ajustement appropriés pour les réactions difficiles à modéliser ;
- Validation des modèles par confrontation à des données expérimentales.

Si un modèle cinétique a bien été bâti pour modéliser la voie basse de la glycolyse, il n'en reste pas moins que d'autres aspects de la modélisation de voie métabolique, pour la production de molécules, n'ont pas encore été abordés :

- La modélisation par des modèles basés sur l'utilisation de données (**chapitre 3**) ;
- L'évaluation des différents types de modélisation (**chapitre 3**) ;
- La reproductibilité de ces méthodes sur d'autres voies métaboliques (**chapitre 4**) ;

- L'influence des caractéristiques de ces voies métaboliques dans le choix du meilleur modèle à utiliser (**chapitre 4**).

Et cela, avant de s'initier à l'implémentation d'un système de contrôle sur le modèle de la voie métabolique au **chapitre 5**.

Nous avons vu dans ce chapitre comment développer un modèle cinétique pour la prédiction du flux en sortie d'une voie métabolique, en prenant en considération des informations sur les enzymes composant la voie (paramètres cinétiques, activité enzymatique) et les concentrations en substrats et cosubstrats. Dans le chapitre suivant, nous allons comparer la performance de divers modèles bâtis pour la voie basse de la glycolyse d'*Entamoeba histolytica*.

Chapitre 3

Étude comparative de trois types de modèles « boîte-blanche », « boîte-grise » et « boîte-noire »

Dans ce chapitre nous nous appliquons à modéliser le segment de la voie basse de la glycolyse par d'autres méthodes. Ces modèles sont construits pour permettre la prédiction du flux de sortie de la glycolyse chez le parasite *Entamoeba histolytica*. Une fois les modèles établis, nous entreprenons une comparaison de l'ensemble des modèles, afin d'évaluer leur performance.

Nous avons précédemment développé plusieurs modèles de type « modèle de connaissance », qui sont parfois difficiles à mettre en place à cause de la longueur des voies métaboliques à étudier, ou encore à cause du manque de connaissance de ces voies de synthèse. Cela nous amène donc à considérer d'autres techniques de modélisation pour la représentation des voies métaboliques. Parmi les modèles développés dans cette partie de notre étude, nous recensons :

- Les modèles « boîte-blanche » : ce sont les modèles qui tiennent compte de la cinétique de chaque enzyme contenue dans cette voie, ils décrivent de manière détaillée chaque réaction enzymatique. Ils sont au nombre de quatre et reprennent certains des modèles que nous avons développés dans le **chapitre 2.**
- Les modèles « boîte-noire » : ce sont des modèles bâtis avec des données expérimentales provenant toujours de l'équipe du Département de Biochimie de l'Institut National de Cardiologie Ignacio Chávez, au Mexique. De manière plus précise, nous développerons ici des modèles de type « Réseaux de Neurones Artificiels ».

- Un modèle « boîte-grise » qui est à l'interface des deux précédents modèles. La construction et l'architecture de ce modèle hybride ont été abordées en détail dans le chapitre précédent (**chapitre 2**).

Aussi, en parallèle de la construction et de l'évaluation de ces modèles de la glycolyse que nous allons présenter et évaluer, nous menons une analyse d'identification des points de contrôle du flux de la voie étudiée. En effet, cette étude complémentaire permettrait d'identifier les enzymes à optimiser ou à « changer », si nous souhaitons obtenir une production plus optimale ; des enzymes que l'on pourrait qualifier de « bottleneck », autrement dit de goulot d'étranglement dans la voie de production considérée. Si nous examinons l'aspect sanitaire ou thérapeutique de cette étude, l'identification de ces points de contrôle dans la voie de la glycolyse permet l'identification ou la confirmation de l'existence de nouvelles cibles thérapeutiques au sein de la voie de la glycolyse.

Dans ce chapitre, les modèles boîte-blanche et boîte-grise sont développés au moyen du même logiciel COPASI, utilisé au chapitre précédent. Les modèles boîte-noire, eux, sont développés sur RStudio. Ce logiciel, libre et gratuit, est un environnement de développement qui met à disposition de ses utilisateurs des outils et facilite l'écriture des scripts écrits en R (Wickham and Grolemund, 2017). Quant à R, il s'agit d'un langage de programmation très répandu dans plusieurs domaines : en chimie (Murrell and Wehrens, 2006; Setiawan, 2020), physique (Banas *et al.*, 2013; Pickering *et al.*, 2017; Amri *et al.*, 2016) ou en biologie (Layeghifard *et al.*, 2018; Grissa *et al.*, 2016; Terkelsen *et al.*, 2020). Les réseaux de neurones artificiels constitués dans ce chapitre sont faits à l'aide de deux packages de R : NeuralNet (Fritsch *et al.*, 2019) et Nnet (Ripley and Venables, 2009). Les modèles sont une fois de plus validés ou non en confrontant leurs prédictions avec des données observées et des données générées par un de nos modèles, cela dû à la faible quantité de données expérimentales dont nous disposons. Il s'ensuit une comparaison des modèles ainsi qu'une évaluation de leur fiabilité pour l'estimation du flux.

Les travaux détaillés dans ce chapitre analysent la fiabilité que possède chaque type de modèle cinétique possible (modèle de connaissance, modèle basé sur l'utilisation de données et modèle hybride) à modéliser une voie métabolique et à identifier les enzymes les plus importantes dans le contrôle du flux. Ils permettent aussi de nous remettre en question quant à l'utilisation des modèles boîte-blanche, robustes, mais qui peuvent parfois être difficilement applicables à d'autres voies métaboliques de synthèse ; en utilisant deux autres méthodes que sont les modèles boîte-noire et boîte-grise.

Ces travaux sont issus de l'article publié dans *Scientific Reports* (Lo-Thong et al., 2020).

Identification of flux checkpoints in a metabolic pathway through white-box, grey-box and black-box modeling approaches

Ophélie Lo-Thong, Philippe Charton, Xavier F. Cadet, Brigitte Grondin-Perez, Emma Saavedra, Cédric Damour and Frédéric Cadet

ABSTRACT

Metabolic pathway modeling plays an increasing role in drug design by allowing better understanding of the underlying regulation and controlling networks in the metabolism of living organisms. However, despite rapid progress in this area, pathway modeling can become a real nightmare for researchers, notably when few experimental data are available or when the pathway is highly complex. Here, three different approaches were developed to model the second part of glycolysis of *E. histolytica* as an application example, and have succeeded in predicting the final pathway flux: one including detailed kinetic information (white-box), another with an added adjustment term (grey-box) and the last one using an artificial neural network method (black-box). Afterwards, each model was used for metabolic control analysis and flux control coefficient determination. The first two enzymes of this pathway are identified as the key enzymes playing a role in flux control. This study revealed the significance of the three methods for building suitable models adjusted to the available data in the field of metabolic pathway modeling, and could be useful to biologists and modelers.

3.1. Introduction

Entamoeba histolytica is a protozoan parasite responsible for the development of amoebiasis in humans. This disease is a worldwide public health problem that causes over 100 000 deaths per year (Kantor et al., 2018). Indeed, a recent report estimates the prevalence of *E. histolytica* infection at 42% in Mexico, 41% in China and 34% in South Africa (Shirley et al., 2018). So far, no vaccine exists to prevent the infection, but patients who suffer from amoebiasis can be treated with different drugs such as metronidazole or tinidazole. However, intolerances to these treatments and potential appearance of drug resistance (Shirley et al., 2018; Upcroft and Upcroft, 2001; Duchêne, 2015; Samarawickrema, 1997) reveal the urgency of the situation and the need to find new therapies. Previous studies have focused on the identification of new drug targets in *E. histolytica* glycolysis (Saavedra *et al.*, 2005, 2007; Moreno-Sánchez, Encalada, *et al.*, 2008), since the parasite depends completely on glycolysis to produce ATP (Saavedra, Encalada, *et al.*, 2019).

While drug research and development is time consuming and expensive, the use of computational approaches might help to speed up the process. Lately, the combination of *in-vitro* reconstitution and *in-sili*co modeling of the glycolysis pathway in *E. histolytica* highlighted the possibility of using modeling for predicting flux and metabolite concentrations under given conditions (Saavedra et al., 2007) and for appraising the effect of the addition of alternative routes (Moreno-Sánchez, Encalada, *et al.*, 2008). Pathway modeling can be done through many statistical or knowledge driven approaches (Hou et al., 2016). The first one only uses experimental data to understand relationships between biological variables, whereas the second uses pathway information (metabolic reactions, thermodynamic and kinetic parameters) to design complete detailed metabolic pathway reconstructions. Artificial Neural Network (ANN) can be classified among the data-driven approaches and is based on the creation of a network whose structure and functioning are similar to those of a biological neural network (Lancashire et al., 2008). Traditionally, this method is employed to identify new biomarkers of diseases such as cancer (Lancashire et al., 2008) or to predict the bioavailability of drugs in patients (Dorronsoro et al., 2004; Thishya et al., 2018).

The recent model of *E. histolytica* glycolysis applies a knowledge-based method called metabolic network to each part of the pathway: the first part from glucose to dihydroxyacetone phosphate and the second part (figure 2.2) from 3-phosphoglycerate (3PG) to pyruvate (Pyr) (Moreno-Sánchez, Encalada, *et al.*, 2008). Interestingly, these studies found that 3-phoshoglycerate mutase (PGAM) was the main controlling factor in the second part of glycolysis, whereas pyruvate

phosphate dikinase (PPDK) exerted the lowest flux control. This result comes in conflict with previous research (Saavedra et al., 2005), which identified PGAM and PPDK as important flux control steps of amoebal glycolysis. This difference is explained by the use of inappropriate enzyme proportions in the *in-vitro* reconstitution experiments, not identical to those determined in amoebas, in the first study. Moreover, here our study is based on the experimental results of Moreno-Sanchez (Moreno-Sánchez, Encalada, *et al.*, 2008).

It should be noted that obtaining a solid knowledge-based model relies, as the name suggests, upon an advanced understanding of the cell system, including physiological metabolite concentrations and enzyme activities, kinetic parameters and the type of mechanism involved, as well as thermodynamic constants of the pathway reactions. However, this knowledge is often not available in the literature or is highly complex to model, as seen with the kinetic mechanism of PPDK (Moreno-Sánchez, Encalada, *et al.*, 2008; Varela-Gómez *et al.*, 2004).

Our analysis shows that the different models predict correctly the final flux values in the second part of *E. histolytica* glycolysis pathway. The ANN model presents great predictive and generalization abilities; however, its complexity, through high Akaike Information Criterion value (AIC), ranks it among the less satisfactory models. The COPASI models provide satisfactory predicted fluxes, as well as the ANN model, with a marked preference for the grey-box approach. Subsequently, the flux control coefficients of the enzymes (C_E^J) are calculated and allow the identification of the key enzymes involved in flux control (Fell, 1992; Saavedra, Gonzalez-Chavez, *et al.*, 2019; Moreno-Sánchez, Saavedra, *et al.*, 2008). Taken together, these models enable the construction of the pathway from experimental data and the determination of the main controlling enzymes in the system, revealing the relevance of both the traditional white-box approach and the novel grey- and black-box approach. Such approaches could be extended to further biological pathway modeling, as they provide models adapted to various backgrounds.

3.2. Materials and methods

3.2.1. Second part of glycolysis experimental data

Experimental data of PGAM, ENO and PPDK activities and pathway flux (J_{obs}) are obtained from plots of a previous study (Moreno-Sánchez, Encalada, *et al.*, 2008). The free online software WebPlotDigitizer (Version 4.1, https://automeris.io/WebPlotDigitizer/) is used to extract data from plots. These data are available in Tables A.1 and A.2 (*Appendix A*).

3.2.2. Artificial Neural Networks (ANNs)

ANNs functioning mimics that of biological neurons, the networks consist of many layers allowing input reception and processing and output delivery. This technique can be used for solving classification or regression problems (Puri *et al.*, 2016). To build the second part of glycolysis in ANNs, different types of software are employed: RStudio (Version 1.1.456), an open-source integrated development environment for R (Wickham and Grolemund, 2017) and two packages: NeuralNet (Version 1.44.2) and Nnet (Version 7.3-12) (Fritsch *et al.*, 2019; Ripley and Venables, 2009).

3.2.3. Complex Pathway SImulator (COPASI) metabolic networks

A first metabolic network of the studied pathway was kindly provided by the authors of a previous study (Moreno-Sánchez, Encalada, *et al.*, 2008). This model is developed on GEPASI (Mendes, 1997), an old software for metabolic pathway modeling, replaced by COPASI since 2002.

The second part of the glycolysis is also modeled by using the open source software called COPASI (Version 4.24) (Hoops *et al.*, 2006). This software is used for metabolic network design, analysis and optimization. The resulting metabolic networks are based on the use of enzyme properties (kinetic parameters and mechanism-based rate equations).

The ANN and COPASI models built during the present study are available in the Github repository, https://github.com/ophelielt/Lo-Thong_et_al._White-box_grey-box_and_black-box_pathway_modeling.git.

3.3. Methodology

3.3.1. Black- white- and grey-box approach procedure

To conduct the present study, a specific methodology, different from that envisaged in the original article (Moreno-Sánchez, Encalada, *et al.*, 2008), is defined (figure 3.1).

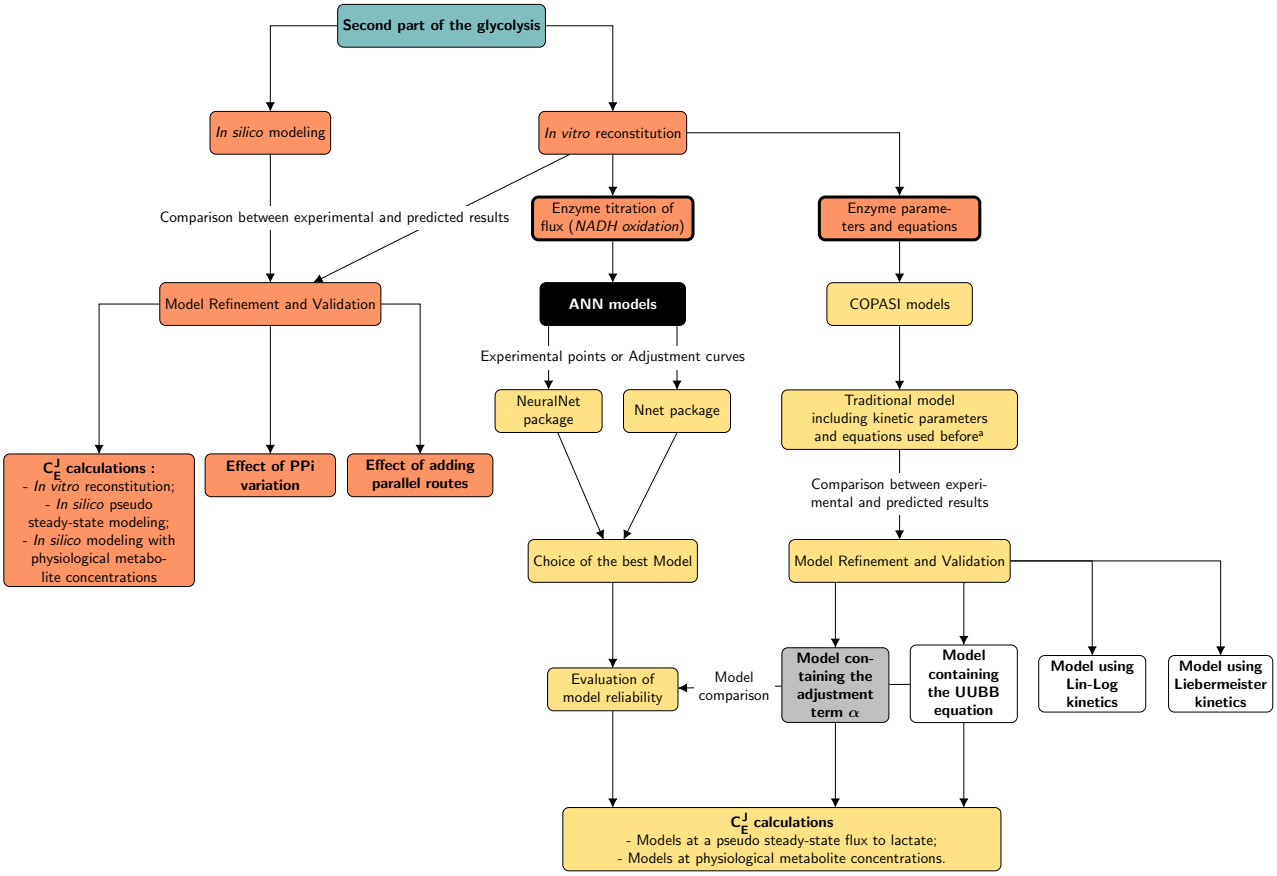

FIGURE 3.1 : *Study workflow.*
Moreno-Sanchez et al. methodology (Moreno-Sánchez, Encalada, et al., 2008) is represented in orange, whereas the methodology proposed here is represented in yellow. Boxes with a thick line indicate the experimental data used in this study; left box: the flux mentioned here refers to pathway flux titration by changing enzyme activities. The last boxes are the techniques used for a better understanding of the metabolic pathway. The five final models designed in this work are colored in black, white or grey. a See "Complex Pathway SImulator (COPASI) metabolic networks" part.

In the first case of the black-box approach, ANN models are built with the experimental data concerning the relationship of pathway flux versus enzyme activity in the pathway *in-vitro* reconstruction. Then, in the second and third case of the white- and grey-box approach, metabolic networks are built with enzyme parameters measured experimentally, and rate equations (Segel,

1975) according to the type of kinetic mechanism described for each enzyme. Once the models are designed, a comparison of their final flux and product concentrations is made. Also, for each approach, two different models are designed : one reaching a pseudo-steady-state flux through lactate and another at physiological metabolite concentrations. Subsequently, calculations of flux control coefficient for each of these models are made, allowing the determination of the main flux controlling enzyme.

3.3.2. Black-box approach

Artificial Neural Networks (ANNs) design

Typical feed-forward networks are designed and consist of three layers of neurons : an input layer, a single hidden layer and an output layer (figure 3.2).

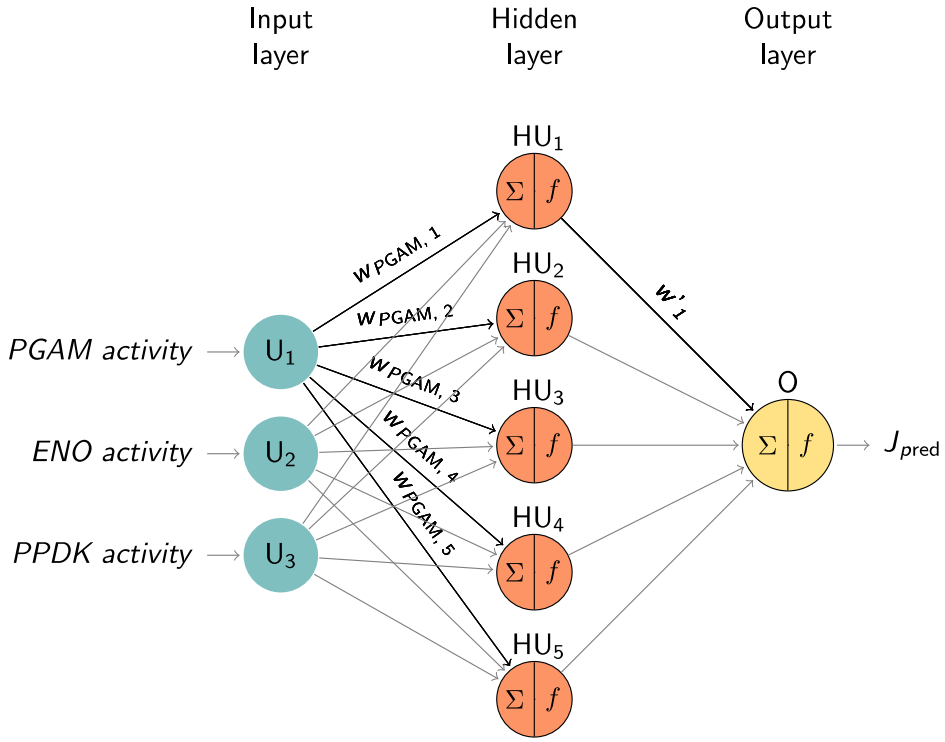

FIGURE 3.2 : *Structure of the ANN models.*
Each node represents an artificial neuron or unit. U_i, HU_j and O are, respectively, the input unit, the hidden unit and the output unit of the different layers; w_i and w'_j are the weights associated with each connection of the network between the input and the hidden layer for the first, and between the hidden and the output layer for the second. Only weights for the first unit (associated with PGAM) of the layers are labelled. Σ constitutes the weighted sum of the input and f constitutes the activation function applied in the unit.

Input data are connected to the neurons and weights (w_i and w'_j) are assigned to each connection. When input data are processed by the neuron, the latter computes the weighted sum of its inputs, then applies an activation function (*f*). The activation function makes it possible to convert input into output and decides whether the neuron is activated or not. There are several activation functions, including the non-linear activation functions : logistic (log) and hyperbolic tangent (tanh). If the resulting output is higher than the set threshold, the neuron is considered as being activated, otherwise not. Lastly, the hidden layer leads to the final output result, displayed in the output layer.

Optimization of ANNs is ensured through the back-propagation method (Rumelhart, David E., 1986) in the NeuralNet package and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (Battiti and Masulli, 1990) in the Nnet package. For detailed information on ANN functioning, see (Jain *et al.*, 1996). In the ANN models, the inputs are the activities of each enzyme (PGAM, ENO and PPDK) used in the *in-vitro* experiment (Table A.1, (Moreno-Sánchez, Encalada, *et al.*, 2008)), and the output is the predicted pathway flux (J_{pred}). Also, each weight in the ANN is assigned automatically by RStudio. Given the small amount of experimental data (Table A.1), ANN models are built with a training set made up of the complete Table A.1 or A.2 datasets (the data from the experimental dots or data from the fitting curves, respectively), then optimized through a Leave-One-Out cross validation (LOOcv) procedure. Then, since we needed a separate test set to prevent overfitting, the models are evaluated on a different test set generated by the grey-box COPASI model (Table A.3).

ANN selection and performance evaluation

The number of artificial neurons (or units) in the hidden layer is selected based on :

- The root-mean-square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \qquad (1)$$

with $\boldsymbol{Y_i}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ respectively the observed and predicted values, $\boldsymbol{n}$ the total number of values, and $\mathbf{i} = 1, 2\ldots n$;

- The mean absolute error (MAE) calculations:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \qquad (2)$$

with $|\ldots|$ symbolizing the absolute value;

- and a specific equation estimating a range of numbers of HUs (Schultz and Reitmann, 2018; Hagan *et al.*, 2014):

$$N_h = \frac{N_s}{\alpha^*(N_i + N_o)} \quad (3)$$

with $\mathbf{N_h}$ the number of HUs, $\mathbf{N_s}$ the number of samples in the training data, $\mathbf{N_i}$ the number of input units, $\mathbf{N_o}$the number of output units and $\boldsymbol{\alpha}$ an arbitrary scaling factor, usually 2-10.

RMSE and MAE are statistical metrics commonly used to evaluate the model performance (M. Vastrad, 2013; Cakit *et al.*, 2015; Küçükönder *et al.*, 2016; Chai and Draxler, 2014).

3.3.3. White-box approach

We discussed this part in the previous chapter, but here we briefly review the general idea.

Complex Pathway SImulator (COPASI) metabolic network design

The metabolic networks built in this study use the enzyme properties (kinetic parameters and kinetic rate equations), which are summarized in table 3.1 and table 3.2, and metabolite concentrations defined in table 3.3. Furthermore, several models are built using either V_{max} or k_{cat} and E and pseudo-steady state metabolite concentrations or physiological metabolite concentrations. All simulations are carried out during the first hour, as was done in the experimental procedure (Moreno-Sánchez, Encalada, *et al.*, 2008).

Enzyme	K_m[a]	K_i[a]	K_{eq}[a]	V_{max}[a]	k_{cat}[b]	E[c]
PGAM	473 (3PG) 106 (2PG)	173 (PPi)		Vf = 75 Vr = 67.24	kcat_f = 3,420 kcat_r = 3,066.14	$2.19*10^{-2}$
ENO	86.4 (2PG) 102 (PEP)	137 (PPi) 610 (3PG)		Vf = 328.5 Vr = 66.61	kcat_f = 8,820 kcat_r = 1,788.43	$3.72*10^{-2}$

PPDK	30 (PEP) 2 (AMP) 91 (PPi) 221 (Pyr) 597 (ATP) 1,342 (Pi)	0.73	Vf = 196.5 Vr = 12.28	kcat_f = 5,220 kcat_r = 326.22	$3.76*10^{-2}$

TABLE 3.1 : Kinetic parameters of the enzymes in the second part of the glycolysis.
Michaelis constants (K_m) and inhibitor constants (K_i) are in μM, maximum rates of the forward and reverse reactions (V_f and V_r) in mU, enzyme amounts (E) in nmol and k_{cat} of the forward and reverse reactions (k_{cat_f} and k_{cat_r}) in min^{-1}. K_{eq} is the equilibrium constant of the reaction. [a]Data taken from a previous study (Moreno-Sánchez, Encalada, *et al.*, 2008) and V_r were calculated from enzyme proportions (Saavedra *et al.*, 2007). [b]Data taken from a previous study (Saavedra *et al.*, 2005) and k_{cat_r} were calculated from V_r and E. [c]E were calculated from V_f and k_{cat_f} by using the equation: $E = \frac{V_f}{k_{cat\_f}}$.

Enzyme	**Kinetic equations[a]**
PGAM	$v = \frac{V_f \frac{[3PG]}{K_{m3PG}} - V_r \frac{[2PG]}{K_{m2PG}}}{1 + \frac{[3PG]}{K_{m3PG}} + \frac{[2PG]}{K_{m2PG}} + \frac{[PP_i]}{K_{iPP_i}}}$
ENO	$v = \frac{V_f \frac{[2PG]}{K_{m2PG}} - V_r \frac{[PEP]}{K_{mPEP}}}{1 + \frac{[2PG]}{K_{m2PG}} + \frac{[PEP]}{K_{mPEP}} + \frac{[PP_i]}{K_{iPP_i}} + \frac{[3PG]}{K_{i3PG}}}$
PPDK[b]	$v = \frac{V_f\left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}}\right)}{K_{mA}\mathrm{B} + K_{mB}\mathrm{A} + K_{mC}B + K_{mB}C + \frac{V_f K_{mQ}\mathrm{P}}{V_r K_{eq}} + \frac{V_f K_{mP}\mathrm{Q}}{V_r K_{eq}} + \frac{V_f K_{mQ}\mathrm{R}}{V_r K_{eq}} + \frac{V_f K_{mR}\mathrm{Q}}{V_r K_{eq}} + ABC + \frac{V_f \mathrm{PQR}}{V_r K_{eq}}}$

TABLE 3.2 : Kinetic equations of the enzymes in the second part of the glycolysis.
[a]In models using k_{cat} and E, V_f were replaced by $k_{cat\_f}*E$ and V_r were replaced by $k_{cat\_r}*E$. [b]A, B and C and K_{mA}, K_{mB} and K_{mC} are respectively the concentrations and $K_m$of the substrates PEP, AMP and PP_i; P, Q and R and K_{mP}, K_{mQ} and K_{mR} are the concentrations and K_m of the products Pyr, ATP, P_i.

Metabolite	Pseudo-steady state concentrations (in μM)[a]	Physiological concentrations (in μM)[b]
3PG	4,000	400
AMP	200	1,600
PPi	1,700	450
ATP	3,000	5,000
Pi	10,000	5,400

TABLE 3.3 : Metabolite concentrations used in the models.

[a]See Tables 1-2 of Ref. (Moreno-Sánchez, Encalada, *et al.*, 2008). [b]See Table 3 of Ref. (Moreno-Sánchez, Encalada, *et al.*, 2008).

As in the previous study, for establishing a quasi steady-state and calculating the flux control coefficients during modeling, a last reaction is added: Lac formation from Pyr (figure 2.2). The kinetic equation of LDH is $k \times [Pyr]$, with the rate constant $k = 2{,}000\, min^{-1}$, and the Lac concentration is fixed at 300 μM.

3.3.4. Metabolic network refinement and validation

To enhance the COPASI model predictions, changes to their contents are carried out. First of all, the PPDK kinetic equation is modified and a more accurate one describing the full rate equation is used, the Uni Uni Bi Bi Ping-Pong (UUBB) mechanism (Equation 4) as previously determined (Varela-Gómez *et al.*, 2004):

$$v_2 = \frac{V_f V_r \left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}} \right)}{\mathrm{D}} \qquad (4)$$

with the denominator $D = V_r K_{iB} K_C A + V_r K_C AB + V_r K_B AC + \frac{V_f}{K_{eq}} K_{iR} K_Q P + \frac{V_f}{K_{eq}} K_R PQ +$

$V_r K_{iB} \frac{K_C}{K_{iQ}} AQ + \frac{V_f}{K_{eq}} K_Q PR + \frac{V_f}{K_{eq}} K_P QR + \frac{V_f}{K_{eq}} K_Q PR + V_r K_A BC + \frac{V_f}{K_{eq}} \frac{(K_{iR} K_Q)}{K_{iC}} CP + V_r \frac{K_c}{K_{iQ}} ABQ +$

$\frac{V_f}{K_{eq}} \frac{K_R}{K_{iA}} APQ + \frac{V_f}{K_{eq}} \frac{K_P}{K_{CB}} BQR + V_r \frac{K_A}{K_{iP}} ACP + V_r \frac{K_A}{K_{iR}} BCR + \frac{V_f}{K_{eq}} \frac{K_P}{K_{iB}} BQR + V_r \frac{K_C}{(K_{iQ} K_{iiC})} ABCQ +$

$\frac{V_f}{K_{eq}} \frac{K_Q}{K_{iC}} CPR + \frac{V_f}{K_{eq}} \frac{K_Q}{(K_{iC} K_{iiC})} CPR + V_r \frac{K_A}{(K_{iR} K_{iiC})} BCQR + V_r ABC + \frac{V_f}{K_{eq}} PQR + \frac{V_f}{K_{eq}} \frac{(K_{iR} K_Q)}{K_{iA}} AP$

; **A**, **B** and **C** and **P**, Q and **R** are respectively the concentrations of the substrates PEP, AMP and PP_i and of the products Pyr, P_i and ATP of PPDK reaction ; $\boldsymbol{K}$ is the Michaelis constant; $\boldsymbol{K_i}$ and $\boldsymbol{K_{ii}}$ are respectively the dissociation constant of the substrate or product and the inhibitor constant that affects the intercept ($1/V_{max}$). The experimental and fitted constants are listed in Table 2.3 (**Chapter 2**).

Also, the estimation of kinetic parameters is made with COPASI Parameter Estimation task. With this task, a range of parameters is tested by COPASI, which predicts the final flux or the product concentrations and compares them to the experimental data. The process relies on the minimization of the cost function (Equation 5), i.e., the minimization of the error between the experimental values and the corresponding predicted values.

$$\mathrm{E(P)} = \sum_{i,j} \omega_j \,.\, (x_{i,j} - y_{i,j}(P))^2 \qquad (5)$$

with $\boldsymbol{E}$ the calculated error, **P** the tested parameter, $\boldsymbol{\omega_j}$ is the calculated weight for each experimental data column, $\boldsymbol{x_{i,j}}$ a point in the dataset and $\boldsymbol{y_{i,j}(P)}$ the corresponding predicted value; **i** and **j** are the rows and columns in the experimental dataset. The weight calculation method was the mean square: $\omega_j = \frac{1}{\bar{x^2}_j}$, with $\bar{x^2}_j$ the mean of squared data from one column. The software provides a list of optimization methods, to find optimized values for the estimated parameters[4].

Again, two types of estimations are carried out:

- one estimating one or several parameters with one target value and
- the other estimating one or several parameters with many target values.

The models obtained constitute the white-box approach, with known enzymatic parameters and equations.

3.3.5. Grey-box approach

This part was also developed in the previous chapter in the section « 2.2.4. Modeling the lower pathway of glycolysis using a hybrid model ».

[4] http://copasi.org/Support/User_Manual/Methods/Optimization_Methods/

In the specific case of the grey-box approach, to improve the COPASI model predictions, the kinetic equation of PPDK is changed to a ter-reactant reversible equation (Moreno-Sánchez, Encalada, *et al.*, 2008) which was modified as follows (Equation 6):

$$v = \frac{V_f\left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}}\right)}{K_{mA}\mathrm{B} + K_{mB}\mathrm{A} + K_{mC}\mathrm{B} + K_{mB}\mathrm{C} + \frac{V_f K_{mQ}\mathrm{P}}{V_r K_{eq}} + \frac{V_f K_{mP}\mathrm{Q}}{V_r K_{eq}} + \frac{V_f K_{mQ}\mathrm{R}}{V_r K_{eq}} + \frac{V_f K_{mR}\mathrm{Q}}{V_r K_{eq}} + \mathrm{ABC} + \frac{V_f \mathrm{PQR}}{V_r K_{eq}} + \alpha\left|V_f - V_{f0}\right|} \quad (6)$$

with the adjustment term $\boldsymbol{\alpha\,|\,V_f - V_{f0}\,|}$ in the denominator, $\boldsymbol{\alpha}$ is a defined number, $\boldsymbol{V_{f0}}$ is the PPDK maximum rate in the forward direction used in the *in-vitro* reconstitution and $\boldsymbol{V_f}$ is the PPDK maximum rate in the forward direction in the model.

This particular model was built because, although the previous model could predict fairly well the final flux when PGAM and ENO activities were varied, it overestimated the flux when PPDK activity was varied. However, the previous model predicted the flux well, with the enzyme parameters used in the *in-vitro* reconstitution. Therefore, an adjustment term should be added, in order to decrease PPDK rate with α. Also, as V_f of PPDK is equal to V_{f0} when PGAM's or ENO's activity is varied, α is multiplied by $V_f - V_{f0}$, so that the adjustment term to be zero when $V_f = V_{f0}$ and the flux predictions are not modified in these two cases mentioned above. Also, to ensure that the adjustment term is positive, we used the absolute value $\left|V_f - V_{f0}\right|$.

To determine the value of α, first a range of values from 0 to $4{*}10^6$ with steps of 10^6 is assessed. Then the range and the steps are reduced, from 10^6 to 1, until we obtain better results for RMSE, and coefficient of determination (R^2) between the predicted and experimental data. The equation for R^2 is given below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (7)$$

with $\boldsymbol{Y_i}$ and $\boldsymbol{\hat{Y}_i}$ respectively the observed and predicted values, $\boldsymbol{n}$ being the total number of values and $\mathbf{i} = 1, 2 \ldots n$.

It is important to note that this parameter α has no biological significance and is determined by a data-driven learning method, hence the name "grey-box" for this model.

3.3.6. Model comparison

To compare accuracy of the models, RMSE, R^2 and AIC are assessed for the experimental dataset (Table A.2). The same statistical metrics are used to evaluate their generalization ability with the test dataset (Table A.3).

AIC measures the quality of the model by taking into account its complexity. Additionally, as the ratio "number of data-number of parameters" is less than 40, a corrected AIC is calculated as follows (Fritsch *et al.*, 2019; Panchal *et al.*, 2010):

$$\mathrm{AIC} = 2{*}\mathrm{k} + \mathrm{n}{*}\ln\left(\frac{\mathrm{SSE}}{n}\right) + \frac{2{*}k{*}(k+1)}{n-k-1} \qquad (8)$$

with **k** being the number of parameters, **SSE** the Sum of Square Errors and $\boldsymbol{n}$ the number of data.

Furthermore, to assess the generalization ability of the models, a comparison of RMSE, R^2 and MAE is made on the previous test set (Table A.3).

3.3.7. Flux control analysis

For purposes of analyzing the pathway flux control and identifying the key enzymes involved in the flux control, the flux control coefficient of each enzyme (C_E^J) is calculated with each model (ANNs and metabolic networks). This measure, generally used in Metabolic Control Analysis (MCA), allows us to assess quantitatively the impact of the enzyme on the pathway flux (Fell, 1992; Saavedra, Gonzalez-Chavez, *et al.*, 2019; Moreno-Sánchez, Saavedra, *et al.*, 2008). Here, C_E^J is determined in an analytical way using the formula mentioned below (Equation 9):

$$C_E^J = \frac{\partial J}{\partial x} * \frac{x_0}{J_0} \qquad (9)$$

where $\boldsymbol{J}$ is the flux and $\boldsymbol{x}$ is either the enzyme activity in the case of ANNs or the rate of the reaction catalyzed by the enzyme in the case of metabolic networks (COPASI), multiplied by a scalar factor $\frac{\boldsymbol{x_0}}{\boldsymbol{J_0}}$ which represents the reference values of enzyme activity/reaction rate and pathway flux.

3.4. Application and results

3.4.1. ANN modeling of the second part of glycolysis

First, we model the second part of *E. histolytica* glycolysis using the black-box modeling approach with ANN models and the first experimental dataset (Table A.1, figure 3.3 A, B) or the second experimental dataset (Table A.2, figure 3.3 C-F). For the first dataset, the evaluation of RMSE in cross-validation (cvRMSE) and MAE in cross-validation (cvMAE) shows a fluctuation of the error values when the number of HUs is varied and allows the identification of the best ANN model, presenting the lowest cvRMSE and cvMAE values. Also, the calculation of N_h gives a maximum value of 4 (α =2), making it possible to identify the best model, regarding cvRMSE and cvMAE, with a number of HU equal to 1 (figure 3.3 A). By comparing the ANN predicted fluxes with the experimental ones, we observe that this model can predict rather well the flux of the pathway for the training set, especially at high values of flux (figure 3.3 B), and even if the calculated errors remain high (cvRMSE= 4.23 nmol·min^{-1}, cvMAE= 2.78 nmol·min^{-1}). The prediction of the test set shows that the model predicts the flux better when PGAM or ENO activity is varied, than when PPDK's activity is varied. This can be explained by the small experimental data number in the training set, which is derived from experimentally controlled conditions. We built other ANN models with the NeuralNet package and tanh activation function and Nnet package, but the predictions are less good than those of previous models, with lower R^2 in cross-validation (cvR^2) and respectively, cvRMSE = 4.47 nmol·min^{-1} and cvMAE=2.84 nmol·min^{-1}, for the first one and cvRMSE=4.56 nmol·min^{-1} and cvMAE = 2.66 nmol·min^{-1} for the second one (figure 3.4).

Afterwards, we built another ANN model, this time using the second dataset, corresponding to the data from the fitting curves obtained from the experimental points in the first dataset. From the two packages used, we notice that, with NeuralNet and tanh activation function, it is easier to identify the optimal number of HUs, which is 18, but this is not the case with the Nnet package, where the models with 22 and 23 HUs present a better cvMAE or a better cvRMSE (figure 3.3 C, D). As RMSE is the most used model selection criterion of both, we use 23 HUs for the second model with the Nnet package. The comparison of these two models shows their ability to simulate the metabolic pathway, with better results for the Nnet model (figure 3.3 E, F). Also, the calculation of N_h gives a maximum value of 23 (α =2); thus, both models comply with the limit set by the equation.

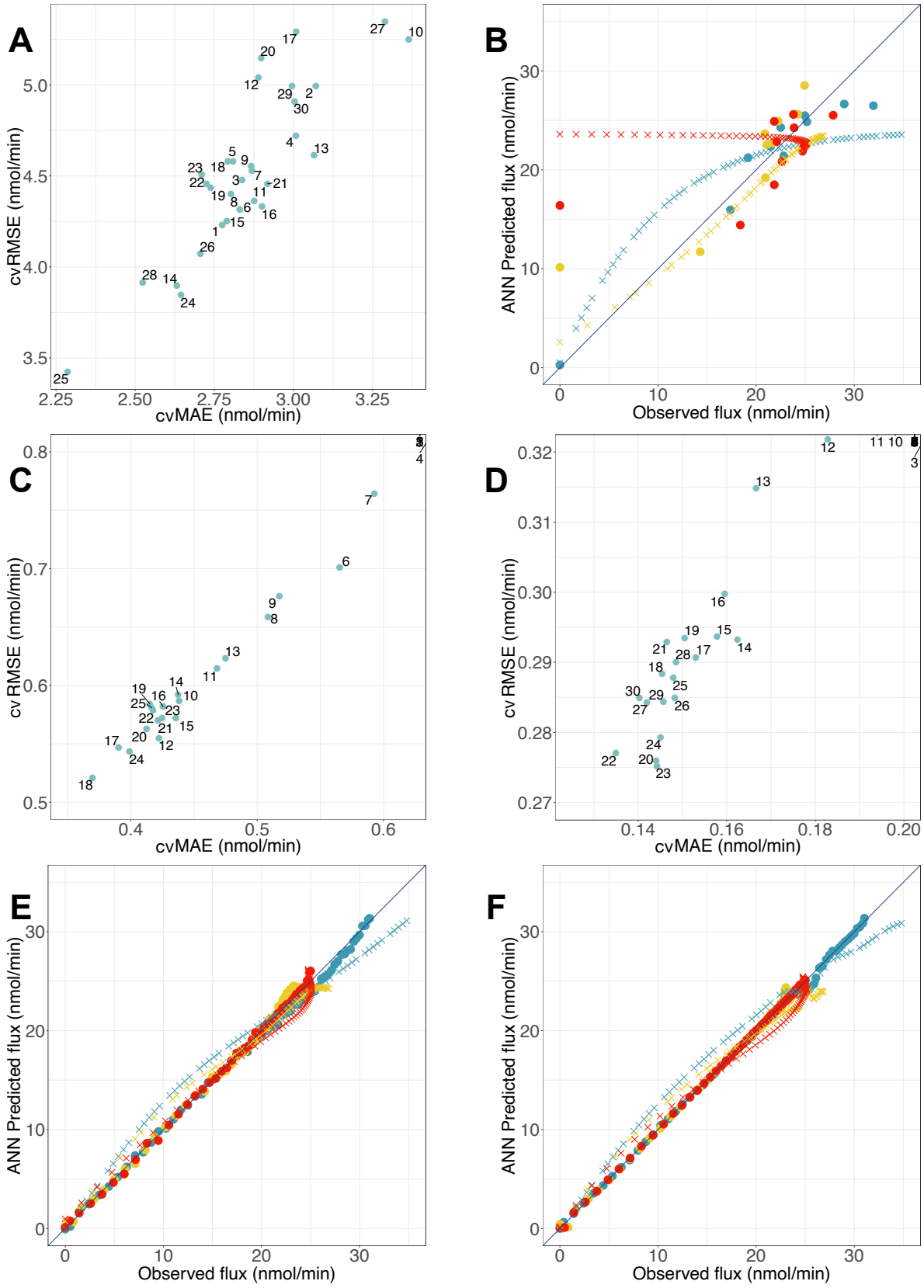

FIGURE 3.3 : *ANN model selections and flux predictions.*

(A) *cvRMSE and cvMAE for the first dataset and using NeuralNet package and log activation function. The numbers represent the number of HUs.* ***(B)*** *Flux prediction with the best ANN model with 1 HU. Training: cvRMSE= 4.23 nmol·min^{-1}, cvMAE= 2.78 nmol·min^{-1}, cvR^2= 0.71 and Test: RMSE=1.56 nmol·min^{-1}, MAE= 1.24 nmol·min^{-1}, R^2= 0.97.* ***(C, D)*** *cvRMSE and cvMAE for the second dataset and using NeuralNet*

package and tanh activation function ***(C)*** *or Nnet package and log activation function* ***(D)****. The numbers represent the number of HUs.* ***(E)*** *Flux prediction with the best ANN model using NeuralNet, tanh activation function and 18 HUs for the training set (circles) and test set (crosses). Training: cvRMSE= 0.52 nmol·min^{-1}, cvMAE= 0.37 nmol·min^{-1}, cvR^2= 1 and Test: RMSE=1.61 nmol·min^{-1}, MAE= 1.37 nmol·min^{-1}, R^2= 0.98.* ***(F)*** *Flux prediction with the best ANN model using Nnet, log activation function and 23 HUs for the training set (circles) and test set (crosses). Training: cvRMSE= 0.28 nmol·min^{-1}, cvMAE= 0.13 nmol·min^{-1}, cvR^2= 1 and Test: RMSE=1.69 nmol·min^{-1}, MAE= 1.47 nmol·min^{-1}, R^2= 0.98. Colored circles/ crosses refer to the various levels of enzyme activity: PGAM (blue), ENO (yellow) or PPDK (red) for the training/test set.*

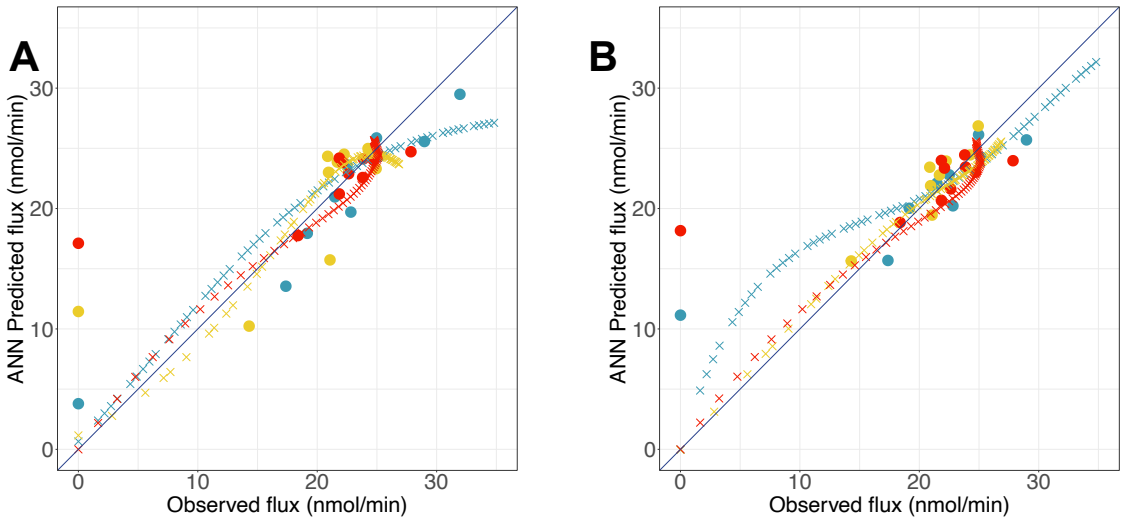

FIGURE 3.4 : *Flux predicted by ANN models with the first and third dataset.*
(A) *Flux prediction with NeuralNet and tanh activation function (4 HUs) with the first dataset. Train: cvRMSE= 4.47 nmol·min^{-1}, cvMAE= 2.84 nmol·min^{-1}, cvR^2= 0.68 and Test: RMSE=2.01 nmol·min-1, MAE= 1.52 nmol·min^{-1}, R^2= 0.95.* ***(B)*** *Flux predicted with Nnet (2 HUs) with the first dataset. Train: cvRMSE= 4.56 nmol·min^{-1}, cvMAE= 2.66 nmol·min^{-1}, cvR^2= 0.67 and Test: RMSE=2.43 nmol·min^{-1}, MAE= 1.76 nmol·min^{-1}, R^2= 0.94. Circle colors refer to the varied enzyme activity: PGAM (blue), ENO (yellow) or PPDK (red).*

However, in order to select the best model and ensure that it is not too specific to our second dataset, we used the test set from the most performing COPASI model (Table A.3), and predicted the final flux with our two ANN models. The NeuralNet model produced better results, with RMSE= 1.61 nmol·min^{-1} and MAE=1.37 nmol·min^{-1}, compared to the Nnet model. These results suggest that this novel black-box approach, using ANN, is relevant for constructing metabolic pathways from experimental data, with better predictions when working with bigger datasets, whether it be with NeuralNet or Nnet package.

3.4.2. Design of metabolic network with the white-box approach

After the modeling phase using the black-box method approach, we focused on the white-box approach and designed mechanistic models with COPASI. The first COPASI model we used was that of Moreno-Sanchez (Moreno-Sánchez, Encalada, *et al.*, 2008); although it was created in GEPASI, we were able to work with this model on COPASI (figure 3.5 A-C).

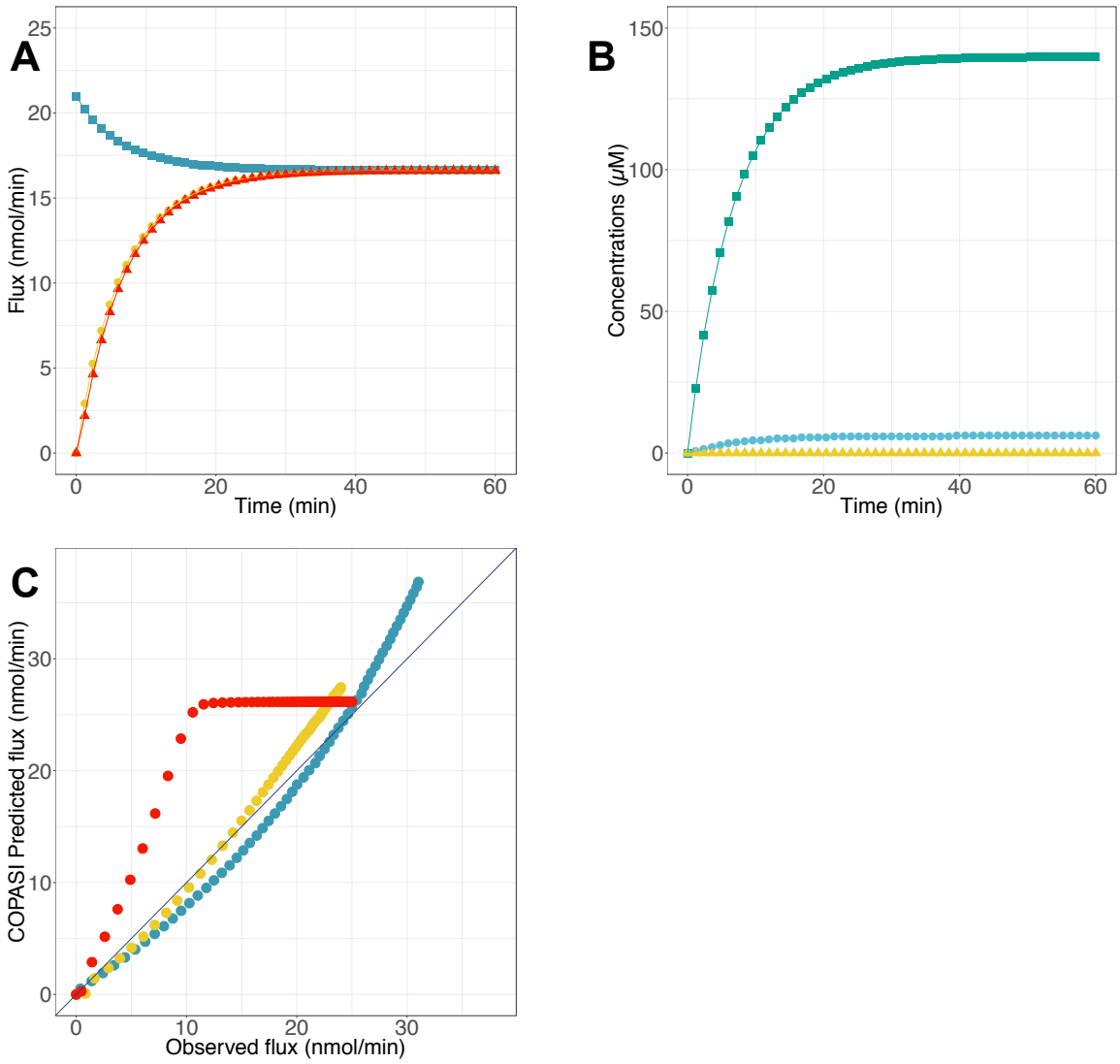

FIGURE 3.5 : *Flux and metabolite concentration predictions with the Moreno-Sanchez model (Moreno-Sánchez, Encalada, et al., 2008) using COPASI software.*

***(A)** PGAM (blue squares), ENO (yellow circles) and PPDK (red triangles) fluxes predicted as function of time. **(B)** Predicted concentrations of 2PG (green), PEP (blue) and Pyr (yellow). **(C)** Flux predicted by the model. RMSE= 4.33 nmol·min^{-1}, MAE= 3.17 nmol·min^{-1}, R^2= 0.85. Circle colors refer to the varied enzyme activity: PGAM (blue), ENO (yellow) or PPDK (red).*

The steady-state flux predicted with this model converged around 16.6 nmol·min^{-1} for the three enzymes, with a flux that decreased for PGAM and increased for ENO and PPDK during simulation time (figure 3.5 A). This result was lower than the experimentally measured result (27 nmol·min^{-1}) (Moreno-Sánchez, Encalada, *et al.*, 2008). As for the prediction of metabolite concentrations, after one hour simulation time, 2PG was at 139.78 μM, PEP at 6.08 μM and Pyr at $8.31*10^{-3}$ μM (figure 3.5 B). Here also the predicted concentrations were higher than the experimentally measured results, with a concentration of 2PG at 58±29 μM and PEP at 37±16 μM (Pyr experimental concentration was not available) in the previous work (Moreno-Sánchez, Encalada, *et al.*, 2008). Furthermore, analysis of the predicted flux when enzyme activities were varied showed quite good prediction of the flux for PGAM and PEP, but not for PPDK, which showed RMSE of 4.33 nmol·min^{-1} (figure 3.5 C).

The results of this first model clearly indicate that the studied metabolic pathway can be modeled with COPASI as a biochemical network using different kinetic parameters and equations, but it needs to be fine-tuned to be more accurate in terms of predictions. The primary modification made in this model concerned the V_{max} values and the metabolite concentrations. Indeed, we replaced these values with those used in the experimental conditions at a pseudo steady-state (see table 3.1, table 3.2, table 3.3 and figure 3.6 A-C). These changes have the effect of increasing the predicted fluxes and metabolite concentrations, in particular with a flux of 25.2 nmol·min^{-1} closer to the experimental value (figure 3.6 A). As for the metabolite concentrations, they were still higher than those measured experimentally (figure 3.6 B). The comparison between the predicted and observed fluxes revealed an enhancement of the predictive capability of our model with RMSE= 3.39 nmol·min^{-1} and R^2= 0.88 (figure 3.6 C), emphasizing the importance of using appropriate parameters in the model.

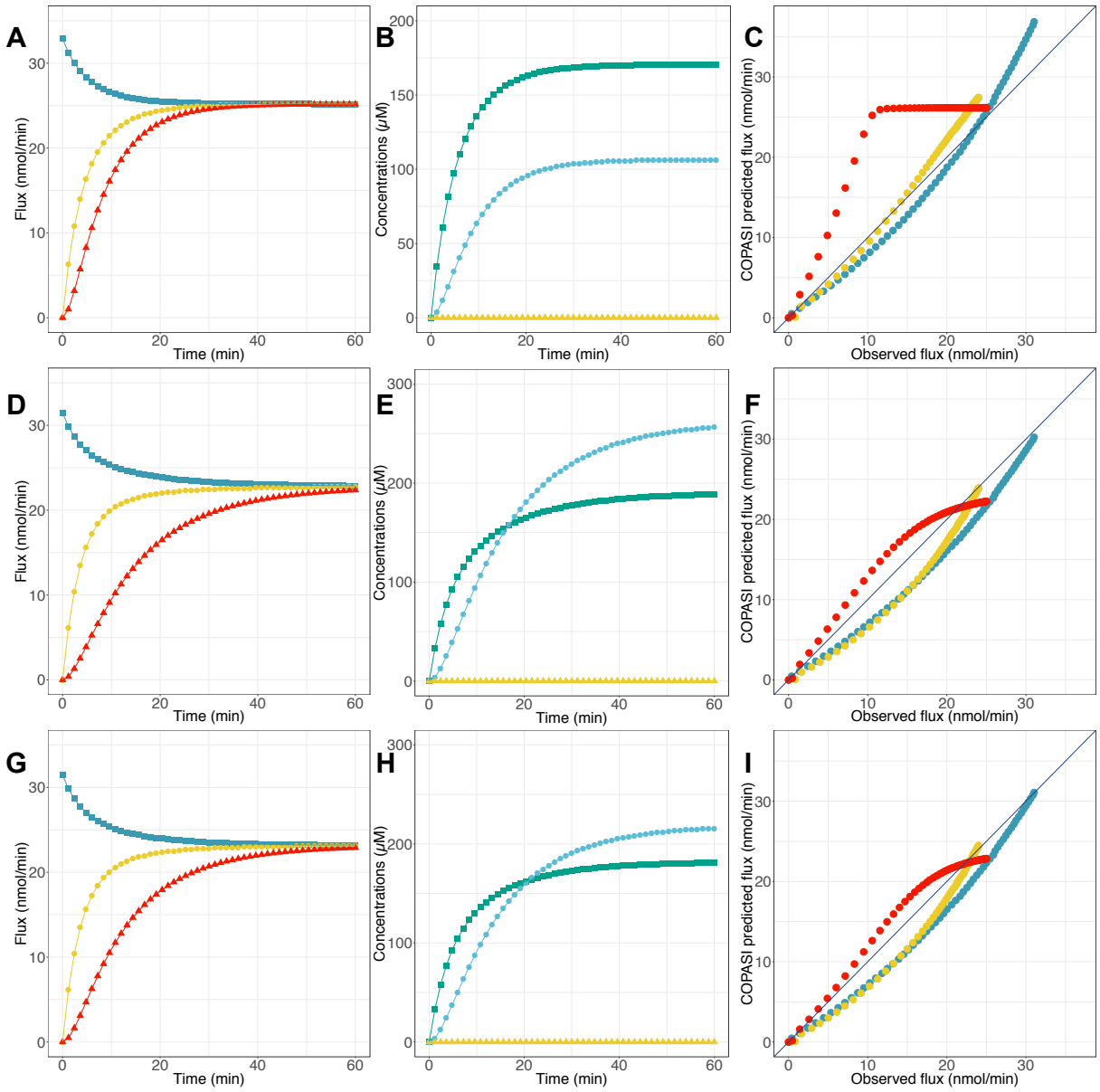

FIGURE 3.6 : *Flux and metabolite concentration predictions with COPASI models.*
(A, D, G) *PGAM (blue squares), ENO (yellow circles) and PPDK (red triangles) flux predicted as function of time with the adjusted Moreno-Sanchez model* ***(A)****, the model containing UUBB equation* ***(D)*** *and the improved model containing UUBB equation* ***(G)****.* ***(B, E, H)*** *2PG (green), PEP (blue) and Pyr (yellow) concentration predicted with the adjusted Moreno-Sanchez model* ***(B)****, the model containing UUBB equation* ***(E)*** *and the improved model containing UUBB equation* ***(H)****.* ***(C, F, I)*** *Flux predictions by the adjusted Moreno-Sanchez model* ***(C)****, the model containing UUBB equation* ***(F)*** *and the improved model containing UUBB equation* ***(I)****. Circle colors refer to the various levels of enzyme activity: PGAM (blue), ENO (yellow) or PPDK (red).*

However, this second model presents a poorer ability to predict the flux when PPDK activity is varied. For this reason, we decided to improve it by modifying the PPDK kinetic equation only and replace the equation used in the preceding models with the more precise Uni Uni Bi Bi kinetic

equation defined by Varela-Gómez et al. (Varela-Gómez *et al.*, 2004) (figure 3.6 D-F). As some kinetic parameters (K_i and K_{ii}) were not characterized experimentally, they were arbitrarily fixed at 1000 μM (see Table 2.3 in **Chapter 2**). This last model yielded a slight decline of reaction fluxes to around 22 nmol·min^{-1} and higher metabolite concentrations than experimentally determined (figure 3.6 D-E). Interestingly, we noted an improvement of flux predictions when enzyme activities were varied (RMSE= 2.43 nmol·min^{-1} and R^2= 0.94), in particular in the case of PPDK activity variation (figure 3.6 F). Therefore, this second attempt to refine the COPASI model revealed that beyond the use of appropriate parameters, our model has to include precise kinetic equations to be more efficient.

As we said before, some parameters are not yet defined experimentally; therefore, we use COPASI Parameter Estimation task to estimate these kinetic parameters. The best results are obtained with the Particle Swarm optimization method, with a cost function of 771.135; the optimized values of K_i and K_{ii} are presented in Table 2.4 (**Chapter 2**). It is worth noting that the cost function value remains high, suggesting a failure of COPASI to estimate parameters better. This could be due to the high number of values to be parameterized and the low number of experimental data. Besides, these parameterized values have no physiological meaning, since they are in the molar range, and could be explained by the negligible impact of the parameterization with COPASI. Simulations run for one hour and fluxes and concentrations are analyzed again (figure 3.6 G-I). We notice no significant change between the initial model and the optimized one. For the most part, the fluxes are increased: PGAM flux is at 23.4 nmol·min^{-1} and ENO flux at 22.9 nmol·min^{-1}, except for PPDK flux which is at 21.3 nmol·min^{-1}, while metabolite concentrations are greater than their experimental values (figure 3.6 G-H). In general, we notice a minor enhancement of flux predictions with this optimized model (figure 3.6 I). These findings suggest that the white-box modeling approach, through COPASI modeling, stands as a conventional method of choice to build consistent *in-silico* models of metabolic pathways and this, despite the fact that, in our case, metabolite concentrations are poorly predicted even after the parameterization of the kinetic constants.

Besides, other approximative models, with lin-log approximation kinetics and Liebermeister kinetics, could have been evaluated (Visser and Heijnen, 2003; Liebermeister *et al.*, 2010). Consequently, we built a model including the approximative lin-log equation (see modeling details in the legend of figure 3.7). Despite simplifying the rate equation by using lin-log kinetics, the model gives results comparable to the previous white-box model, with RMSE=4.8 nmol·min^{-1} and R^2= 0.78 (figure 3.7 C). Another model using the simpler modular rate law from Liebermeister (Liebermeister *et al.*, 2010) is built (see modeling details in the legend of figure 3.8). This model has

the immediate effect of simplifying the rate equation for PPDK and allows good prediction of flux (26 nmol·min^{-1}) in the experimental conditions (figure 3.8 A). However, results show that metabolite concentrations are still overestimated and the model presents a lower predictive capacity compared to the previous models, with RMSE= 4.03 nmol·min^{-1} and R^2= 0.87 (figure 3.8 B, C). Both models, with lin-log approximation kinetics or Liebermeister kinetics, display the same dynamics, with better flux predictions when PGAM's or ENO's activity is varied than when PPDK's activity is varied. Together, these results reveal that there is some kinetic aspects of PPDK kinetics that are not completely modeled by these different mechanistic approaches.

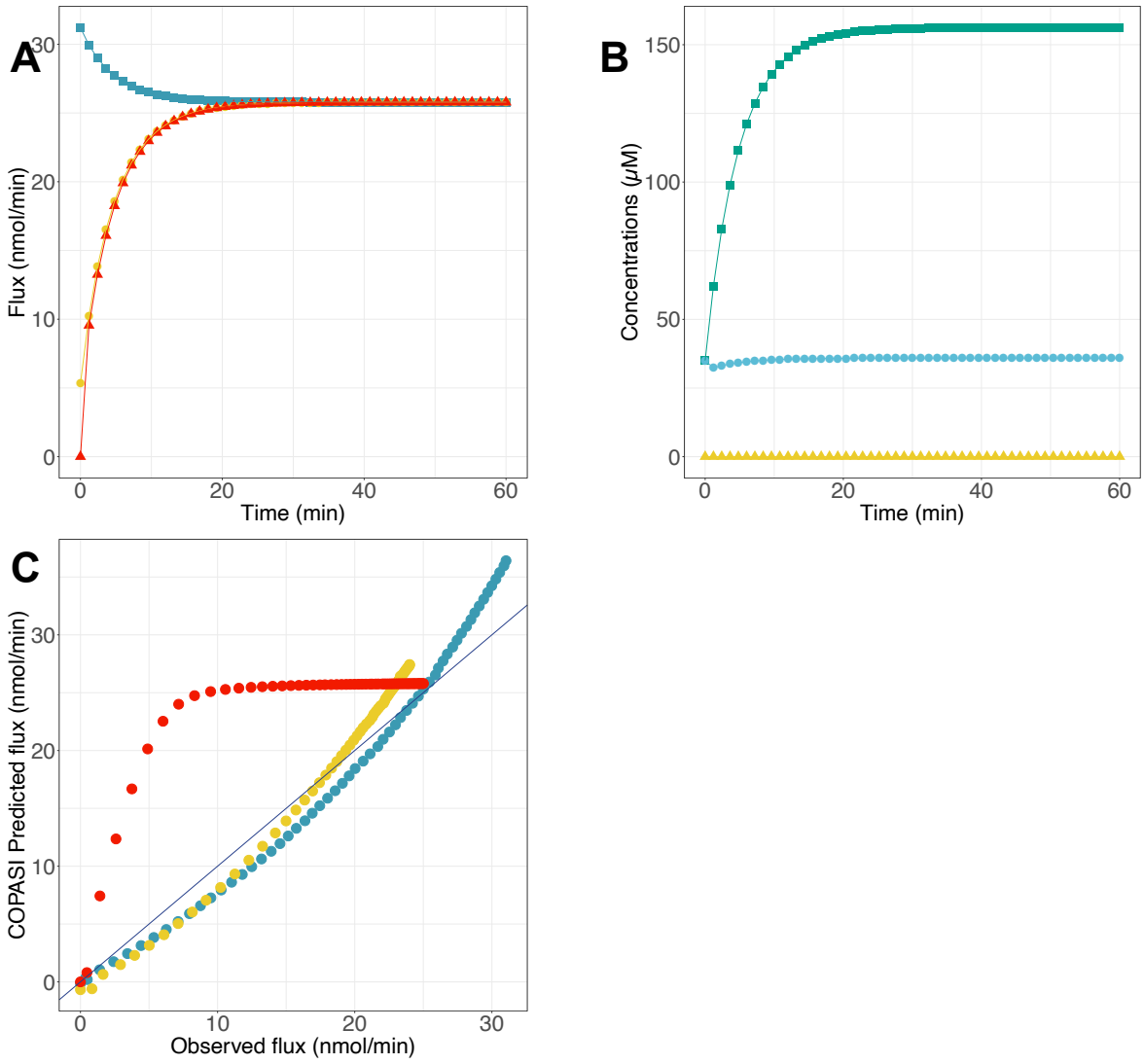

FIGURE 3.7 : *Flux and metabolite concentration predictions with the lin-log approximation kinetics using COPASI software.*

Kinetic parameters used are from table 3.1 and kinetic equations for PGAM and ENO from table 3.2. Lin-log rate equation used for PPDK:

$$v = V_{max} \bullet (1 + \varepsilon_A \bullet ln\left(\frac{[A]}{[A]_{ss}}\right) + \varepsilon_P \bullet ln\left(\frac{[P]}{[P]_{ss}}\right) + \varepsilon_I \bullet ln\left(\frac{[I]}{[I]_{ss}}\right)),$$

with $\boldsymbol{V_{max}}$*, the maximum rates of each reaction;* $\boldsymbol{\varepsilon_A}$*,* $\boldsymbol{\varepsilon_P}$ *and* $\boldsymbol{\varepsilon_I}$ *the elasticities for substrate (A), product (P) and inhibitor (I) and* $[\boldsymbol{A}]_{\boldsymbol{ss}}$*,* $[\boldsymbol{P}]_{\boldsymbol{ss}}$ *and* $[\boldsymbol{I}]_{\boldsymbol{ss}}$ *the steady-state concentrations of substrate, product and inhibitor. Concerning the parameters used here:* V_{max} *are from table 3.1, elasticities (*ε_i*) are estimated with Model 6 and steady-state concentrations are taken from Model 6.* ***(A)*** *PGAM (blue squares), ENO (yellow circles) and PPDK (red triangles) fluxes predicted as function of time.* ***(B)*** *Predicted concentrations of 2PG (green), PEP (blue) and Pyr (yellow).* ***(C)*** *Flux predicted by the model. RMSE= 4.8 nmol·min*$^{-1}$*, MAE= 3.3 nmol·min*$^{-1}$*, R*^{2}*= 0.78. Circle colors refer to the varied enzyme activity: PGAM (blue), ENO (yellow) or PPDK (red).*

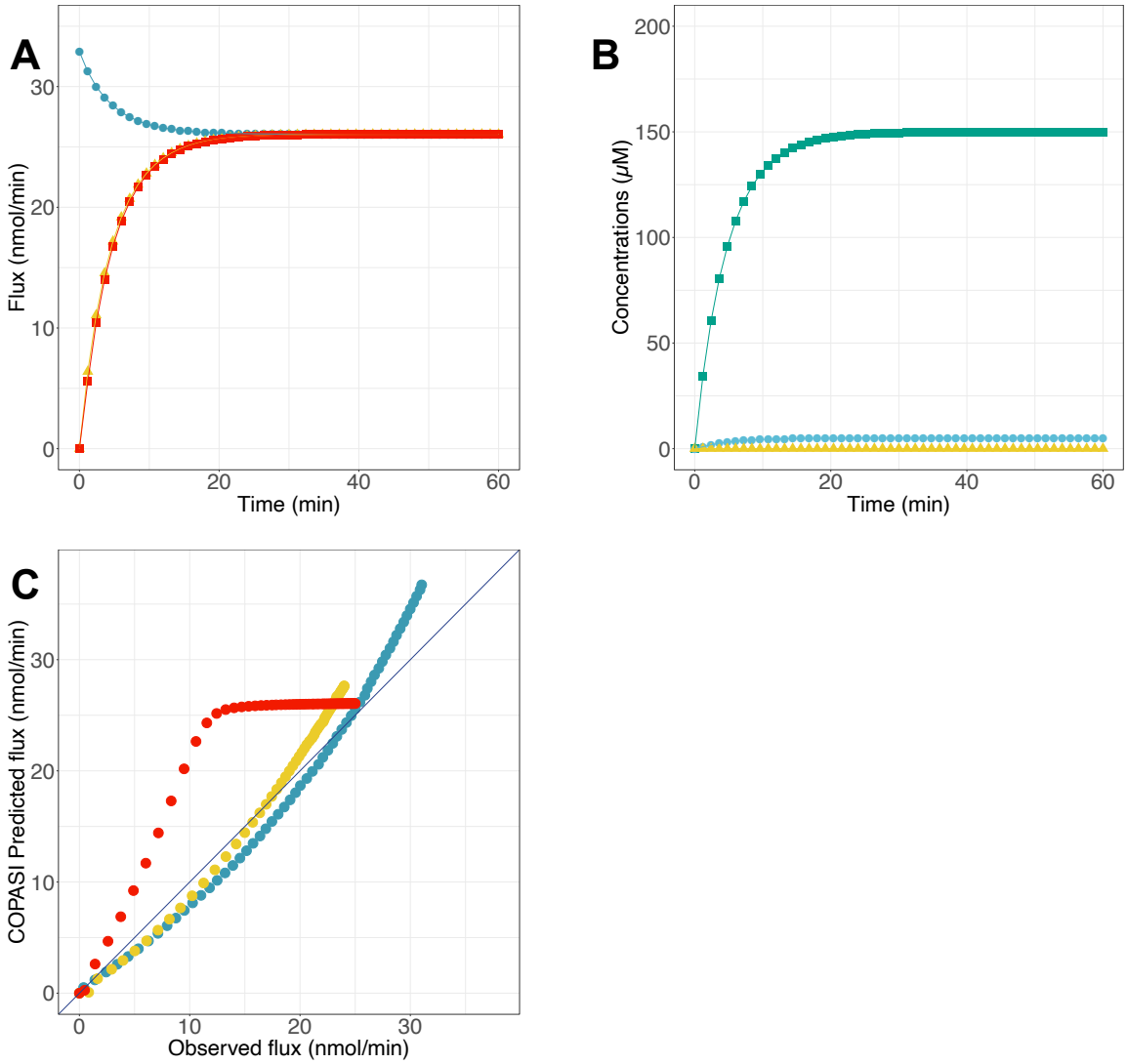

FIGURE 3.8 : *Flux and metabolite concentration predictions with the modular rate law from Liebermeister using COPASI software.*

Kinetic parameters used are from table 3.1 and kinetic equations for PGAM and ENO from table 3.2. Modular rate law equation used for PPDK:

$$v = \frac{V_f \bullet \frac{PEP}{K_{mPEP}} \bullet \frac{AMP}{K_{mAMP}} \bullet \frac{PPi}{K_{mPPi}} - V_r \bullet \frac{Pyr}{K_{mPyr}} \bullet \frac{ATP}{K_{mATP}} \bullet \frac{Pi}{K_{mPi}}}{\left(1 + \frac{PEP}{K_{mPEP}}\right) \bullet \left(1 + \frac{AMP}{K_{mAMP}}\right) \bullet \left(1 + \frac{PPi}{K_{mPPi}}\right) + \left(1 + \frac{Pyr}{K_{mPyr}}\right) \bullet \left(1 + \frac{ATP}{K_{mATP}}\right) \bullet \left(1 + \frac{Pi}{K_{mPi}}\right) - 1}$$

(A) *PGAM (blue squares), ENO (yellow circles) and PPDK (red triangles) fluxes predicted as function of time.* ***(B)*** *Predicted concentrations of 2PG (green), PEP (blue) and Pyr (yellow).* ***(C)*** *Flux predicted by the model. RMSE= 4.03 nmol·min^{-1}, MAE= 2.99 nmol·min^{-1}, R^2= 0.87. Circle colors refer to the varied enzyme activity: PGAM (blue), ENO (yellow) or PPDK (red).*

3.4.3. The grey-box modeling approach

Based on our previous experiences, the major hurdle in the second part of glycolysis modeling is the third reaction catalyzed by PPDK. Then, we investigate the use of a novel approach called the grey-box modeling approach, consisting of using an adjustment term ($\alpha \left| V_f - V_{f0} \right|$) in the kinetic equation of PPDK.

This part was developed in the previous chapter in the section « 2.3.2. Improvement of flux prediction by the hybrid grey-box model ».

3.4.4. Model comparison and reliability

Following the design of the second part of glycolysis using three modeling approaches, we assess the reliability of each approach and proceed to their comparison. Also, for an easier understanding of the following results, the properties of each model are summarized in table 3.4.

Model[a]	Name	Specificity[b]	Number of parameters	Based on…
0	Moreno-Sanchez model	See (Moreno-Sánchez, Encalada, *et al.*, 2008)	20	Experimental

1	Adjusted Moreno-Sanchez model	Respects the experimental conditions at a pseudo steady-state	20	Experimental kinetic data
2	ANN model (NeuralNet, log, HU=1)	Only uses the experimental dots	6	Enzyme activities and final flux data
3	ANN model (NeuralNet, tanh, HUs=18)	Uses data from the fitting curves	91	
4	UUBB model	Use of UUBB equation for PPDK[c]	29	
5	UUBB model optimized	Uses the UUBB equation for PPDK with optimized parameters[c]	29	Experimental and fitted kinetic data
6	Model with an added adjustment term	PPDK equation with an added adjustment term[c]	21	Experimental kinetic data + adjustment term

TABLE 3.4 : List of the main properties of each model.

[a] Only the best models from each approach are kept. [b] For a complete description of the modeling process, see the Methodology section. [c] Respect of pseudo steady-state experimental conditions.

By comparing the predicted fluxes to their experimental values, we found that all models, from Models 1-6, worked well for predicting the final flux when activity of PGAM varies (figure 3.9 A). When ENO activity is varied, we notice that Model 2 does not perform well, particularly for the low values, for which the model overestimates the final flux (figure 3.9 B). Besides, for these two enzymes we note that Models 1, 4 and 5 from the white-box approach and Model 6 from the grey-box approach underestimate the flux when activity of PGAM or ENO is varied, with a gap that seems smaller in the case of the grey-box approach. As expected, dots from Model 3 are practically aligned with the first bisector, suggesting an almost perfect flux prediction with this model (figure 3.9 A, B). Lastly, the variation of PPDK activity shows the greatest effect on model prediction. We observe that Model 2, as well as Model 1, are the two models that have the most difficulty in predicting flux under these conditions (figure 3.9 C). Indeed, they overestimate the flux when PPDK activity is varied; this was also the case for Models 4 and 5, but with a smaller difference between the predicted and observed values. In contrast, fluxes are closely predicted with Models 3,

5 and 6. These results indicate that these models are suitable to simulate our studied metabolic pathway and that we can count on their reliability for the analysis of the flux in the second part of glycolysis, at least for an overall flux ranging from 0 to 30 nmol·min^{-1}.

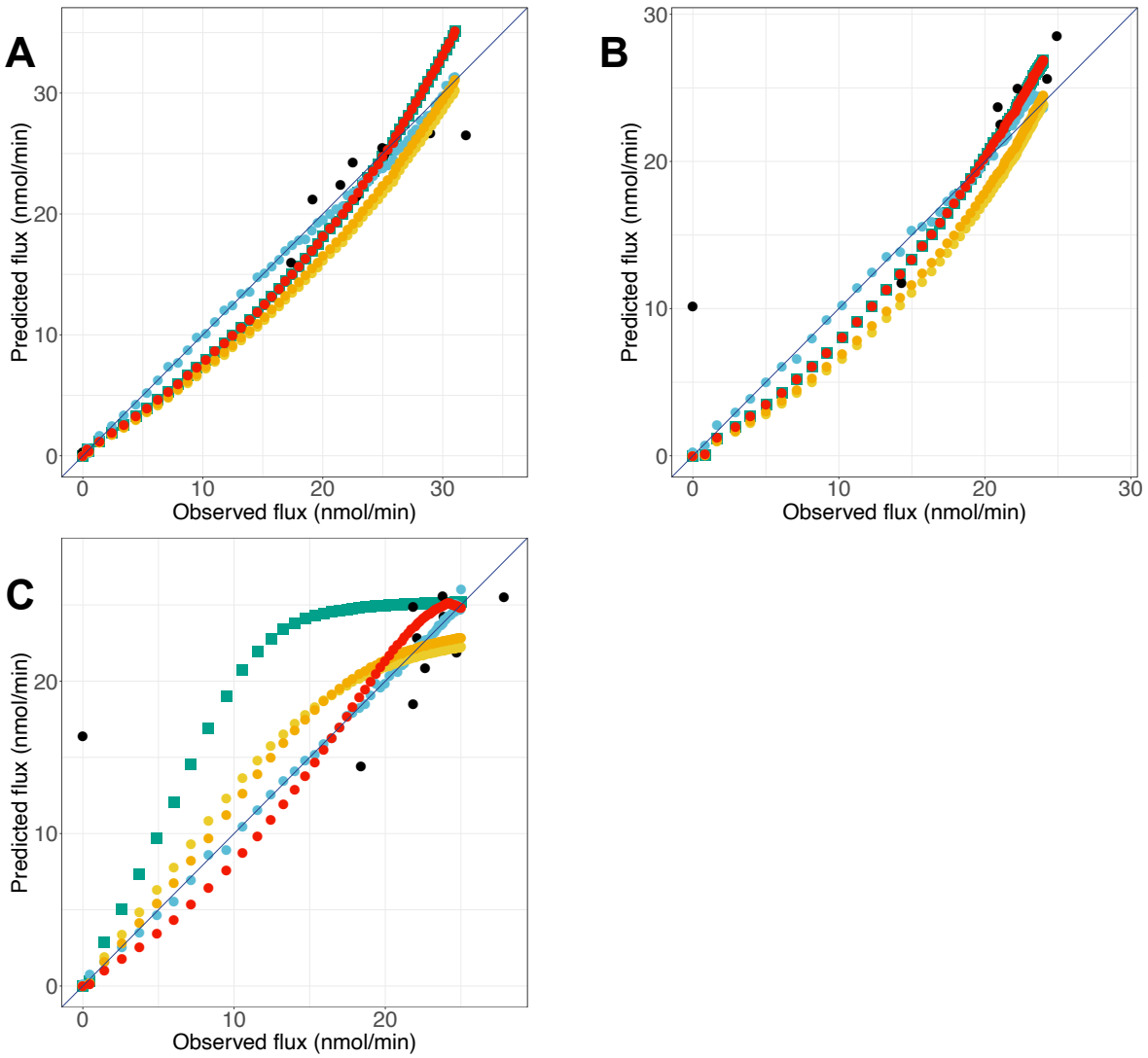

FIGURE 3.9 : *Comparison of flux predictions and experimental flux for all models.*
Flux predictions by the model, when PGAM activity ***(A)****, ENO activity* ***(B)*** *or PPDK activity* ***(C)*** *is varied. Colors refer to the model used: Model 1 (green squares), Model 2 (black circles), Model 3 (blue circles), Model 4 (yellow circles), Model 5 (orange circles) and Model 6 (red circles).*

The analysis of the statistics for each model reinforced the results obtained before (table 3.5). Indeed, all models exhibited a fairly low RMSE under 3 nmol·min^{-1} and a high R^2, around 0.98, when PGAM activity was varied. When ENO activity was varied, almost all models predicted the flux with a good RMSE under 3 nmol·min^{-1} and R^2 above 0.97, except for Model 2. However, when PPDK activity was varied, Models 0, 1 and 2 showed the weakest results, with RMSE above 5 nmol·min^{-1} and a R^2 under 0.9. Only the three models mentioned above (Models 3, 5 and 6) yielded

good results with a low RMSE and a high R^2 value. These results corroborated those obtained earlier. Interestingly, the calculation of AIC allows the establishment of a ranking of models (from the best to less good): Model 2 > 3 > 6 > 5 > 4 > 1 > 0 (table 3.5). Model 2, which has the lowest AIC, proved to be a poor model for flux prediction. Conversely, Model 3, that gives the best results in terms of RMSE, MAE and R^2 presents a good AIC. We also notice that the second-best model in flux prediction (Model 6) also presents a low AIC value.

Model	R^2	RMSE	MAE	AIC	PGAM		ENO		PPDK	
					R^2	RMSE	R^2	RMSE	R^2	RMSE
Model 0	0.85	4.33	3.17	584.74	0.98	2.42	1	2.41	0.71	6.75
Model 1	0.88	3.39	2.48	494.39	0.98	2.02	0.98	1.78	0.8	5.27
Model 2[a]	0.71	4.23	2.78	99.5	0.94	2.19	0.78	4.02	0.41	5.71
Model 3[a]	1	0.52	0.37	124.21	1	0.62	1	0.62	1	0.22
Model 4	0.94	2.43	2.1	396.72	0.98	2.96	0.97	2.3	0.94	1.94
Model 5	0.95	2.06	1.7	336.05	0.98	2.59	0.98	1.89	0.96	1.6
Model 6	0.98	1.71	1.47	244.71	0.98	2.02	0.98	1.78	0.99	1.22

TABLE 3.5 : Comparative table of statistical metrics of each model for the training set (Table A.2). RMSE and MAE are in $nmol \cdot min^{-1}$. [a] For these models, (cv)RMSE and (cv)R^2 are calculated.

Subsequently, in order to evaluate the generalization ability of our models, we predict the flux with the test set (table 3.6). Many models do not have an adequate ability of generalization; nevertheless, Model 6 from the grey-box approach stands out from the others. Indeed, it is the only model able to predict the flux very well from new data, regardless of the enzymatic activity that is varied. Model 0 and 1 can predict the flux well, except when PPDK activity is varied. Also, AIC calculations identify Model 6 as the best one to generalize (AIC=- 486.7), since Model 3 presents higher RMSE, MAE and AIC value (AIC=539.06). These results confirm the reliability of the three approaches for the analysis of the flux in the second part of glycolysis, with a preference for Model 6, which offers the best compromise between precision and complexity.

Model	R^2	RMSE	MAE	AIC	PGAM		ENO		PPDK	
					R^2	RMSE	R^2	RMSE	R^2	RMSE
Model 0	0.86	3.71	2.15	527.32	1	0.88	0.99	1.42	0.59	6.26
Model 1	0.89	2.78	1.06	421.76	1	0.02	1	0.11	0.72	4.87

Model 2	0.52	5.54	3.89	642.46	0.99	4.04	1	5.71	0.99	4.17
Model 3	0.98	1.61	1.37	539.06	0.98	2.12	0.99	1.32	0.98	1.26
Model 4	0.96	2.73	2.51	439.52	1	2.68	1	2.89	0.93	2.63
Model 5	0.97	2.19	2	357.26	1	2.17	1	2.3	0.95	2.08
Model 6	1	0.23	0.13	-486.7	1	0.02	1	0.11	1	0.39

TABLE 3.6 : Comparative table of statistical metrics of each model for the test set (Table A.3).

3.4.5. Identification of the main controlling enzymes of the pathway

After establishing three types of models for the considered metabolic pathway, we determined the enzyme C_E^J with each model. These coefficients are calculated at a pseudo steady-state flux to Lac (table 3.7) or at physiological metabolite concentrations (table 3.8) at the reference or basal level of enzyme activity of 75 mU PGAM, 328.5 mU ENO and 196.5 mU PPDK. Each C_E^J provides a quantitative measurement of the enzyme effect on the pathway flux. The closer the coefficient is to 1, the higher the enzyme impact on the flux. Thus, this coefficient differs from the concept of rate-limiting enzyme, which is commonly defined as the enzyme which catalyzes the slowest step in the pathway and corresponds to a qualitative evaluation of the enzyme impact on the pathway flux (Fell, 1992; Saavedra, Gonzalez-Chavez, *et al.*, 2019; Moreno-Sánchez, Saavedra, *et al.*, 2008).

Model	PGAM	ENO	PPDK
Experimentally determined (Moreno-Sánchez, Encalada, *et al.*, 2008)	0.72	0.11	0.13
Model 0	0.79	0.21	0.0025
Model 1	0.75	0.21	0.04
Model 2[a]	0.4	0.33	0.22
Model 3[a]	0.61	0.12	0.25
Model 4	0.70	0.2	0.1
Model 5	0.71	0.2	0.09
Model 6	0.75	0.21	0.002

TABLE 3.7 : Flux control coefficient determination. [a] For these models, C_E^J are determined manually.

Model	PGAM	ENO	PPDK
Moreno-Sanchez model (Moreno-Sánchez, Encalada, *et al.*, 2008)	0.77	0.18	0.05
Adjusted Moreno-Sanchez model	0.77	0.18	0.05
UUBB model	0.92	0.58	$9.55*10^{-3}$
UUBB model optimized	0.92	0.57	$5.36*10^{-3}$
Model with the added adjustment term ($\alpha = 3.09*10^{6}$)	0.77	0.18	0.05

TABLE 3.8 : Flux control coefficient determination for models at physiological metabolite concentrations

As we can see, at a pseudo steady-state flux to Lac, the enzyme that exerted the greatest control on the final flux is PGAM (0.65±0.2), then ENO (0.18±0.04) and PPDK (0.07±0.1) which showed the least control on the flux (table 3.7). The predicted values by the different models are within the same interval as those experimentally determined by pathway reconstitution (Moreno-Sánchez, Encalada, *et al.*, 2008). Similar results were obtained with all models at physiological metabolite concentrations (table 3.8). From these findings, we can conclude that the main controlling enzymes of the second part of glycolysis in *E. histolytica* are PGAM and, to a lesser extent, ENO and PPDK exert low or no control over the pathway flux.

In addition, we varied the enzyme activity from 0 to 400 mU and observed the final flux during the first hour of simulation using the COPASI model with the adjustment term (figure 3.10). When PGAM was varied, the flux went from 0 to 90.93 nmol·min^{-1} (figure 3.10 A) and when ENO was varied, the flux went to 26.26 nmol·min^{-1}. By contrast, PPDK activity variation did not affect the final pathway flux very much, which went to 24.13 nmol·min^{-1} at 400 mU of PPDK. These results were consistent with previous C_E^J calculations showing that PGAM and ENO are indeed the two main controlling enzymes of the pathway.

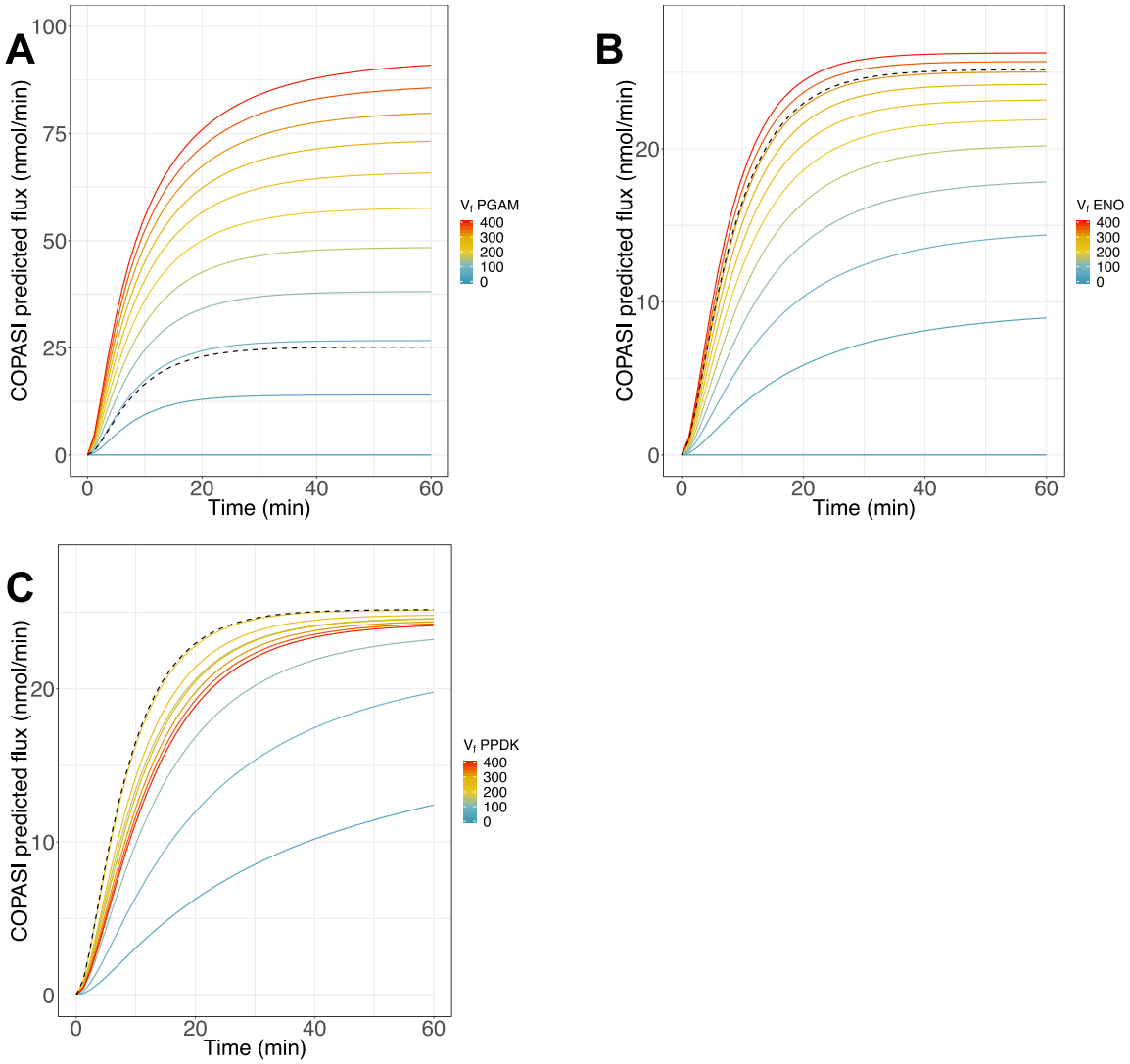

FIGURE 3.10 : *Effect of enzyme variation on the pathway flux.*
Pathway flux predicted with the model with the added adjustment term, when PGAM activity ***(A)****, ENO activity* ***(B)*** *or PPDK activity* ***(C)*** *is varied. Dotted curves: fluxes obtained at the quasi steady-state to Lac.* V_f *of PGAM, ENO and PPDK are in mU.*

3.5. Discussion

3.5.1. Relevance of the white- grey- and black-box approach for the modeling of metabolic pathways

In this work, we model the second part of the glycolysis pathway of *E. histolytica* using three approaches: the white-, grey- and black-box approach, and we highlight their ability to predict the final flux. Many comparative studies are made in other fields to evaluate the relevance of using either of the three methods, and point out that the method depends on the problems encountered (Gernaey *et al.*, 2004; Li and Wen, 2014; Arendt *et al.*, 2018). In the case of energy model building, Li and Wen showed that simplified grey-box models are better as practical building models, compared to white-box models that require numerous parameters (Li and Wen, 2014). In another study, the black-box models outperformed the other two models for the modeling of thermal simulation in a particular environment (Arendt *et al.*, 2018).

Here, the first approach is based on the use of kinetic parameters and equations and is related to the widely used method known as kinetics-based (or dynamic) modeling for industrial applications such as the production of molecules of interest, development of *de novo* synthesis pathways or understanding of microorganism metabolism (Huang and Buekens, 2001; Liu *et al.*, 2009; Petroll *et al.*, 2019). This method can provide accurate predictions; however, it requires numerous parameters and good knowledge of mechanistic rate laws; hence the need to develop new strategies of modeling when we do not have access to this information (Almquist *et al.*, 2014; Saa and Nielsen, 2016).

Despite the use of a more complex kinetic equation in the kinetic models, the results were not satisfactory; consequently, we used a simplified kinetic equation with an adjustment term in the grey-box approach. This is the first time this method is applied to enhance performance of a metabolic pathway kinetic model. In other studies found in the literature, the unknown kinetic constants are parameterized or the kinetic equations can be approximated (Costa *et al.*, 2016; Rohwer, 2012; del Rosario *et al.*, 2008; Liebermeister *et al.*, 2010). The present approach has some major advantages as it needs less parameters than the white-box approach, and it uses simplified kinetic equations that are biochemically plausible.

Finally, we used a novel black-box approach and built an ANN model with experimental data. As previously mentioned, ANN is generally used in biology to solve classification problems, for example, to classify lung carcinomas (Brougham *et al.*, 2011), but it has rarely been used to model a

metabolic pathway (Mendes and Kell, 1996; Voit and Almeida, 2004; Antoniewicz *et al.*, 2006; Naushad *et al.*, 2016). A recent study applied a similar technique to model the first part of glycolysis, and showed the success of this technique for predicting the flux (Ajjolli Nagaraja *et al.*, 2019). This last approach is characterized by its rapidity; however, it requires a large number of experimental data to be sufficiently effective.

Together, the approaches we describe here may be beneficial for modeling other metabolic pathways, depending on background information including "raw" experimental data, kinetic parameters and kinetic equations.

3.5.2. Factors impacting model performance

During this study, we relied on three main statistical metrics (RMSE, MAE and AIC) to evaluate model performance. The results revealed that different criteria are important and impact the value of these metrics and thus the model performance itself. Among these criteria, we identified the size of the dataset, but also the choice of the activation function (log or tanh) and the number of HU in our ANN models. Indeed, having a large number of high-quality datasets is essential to obtain a good ANN, and one challenge here would be to avoid over-fitting (Oyetunde *et al.*, 2018; Angermueller *et al.*, 2016). Other studies bring out the importance of the size of the input sequence during the analysis of the DNA sequence, to increase model performance (Zhou and Troyanskaya, 2015). Also, they reveal the relevance of neural network architecture, proposing the design of multi-task neural networks with multiple output variables (Dahl *et al.*, 2014). These factors raise new questions about the use of ANN to model metabolic pathways, and can be subjected to further investigation concerning the number of inputs and outputs to include in our model, to make it more efficient.

As we said earlier, to predict accurate results COPASI models need extensive data, such as kinetic parameters and equations. Our results reveal the impact of the kinetic equation on the final flux prediction. The impact of the kinetic equation on the model predictions depends on the complexity of the model and on the flux control coefficient of the enzyme. When the enzyme has high $\mathbf{C}_{\mathrm{E}}^{\mathrm{J}}$, variations of its rate equation or small variations in the kinetic constants or V_{max} greatly impact the predicted pathway flux (*e.g.* PGAM in figure 3.10). In contrast, rate equation variations of a low controlling enzyme (such as PPDK) have less impact on the flux. It would be interesting to test in the models the influence of the lack of regulatory feedback on the enzyme that has the highest control, as was done in the Moreno-Sanchez et al. study (Moreno-Sánchez, Encalada, *et al.*, 2008) focusing on PGAM. As was described in that paper, the lack of those regulatory effects

renders the predictive power of the model ineffective. Therefore, regulatory properties on high controlling enzymes can drastically modify the model predictions. Furthermore, the question of which kinetic equation to use in the pathway remains a real topic in research today. Kim et al. review all kinetic rate expressions used in the kinetic model, from mechanistic expressions (Michaelis-Menten and Hill rate laws equations) to approximate kinetic equations (lin-log kinetics, modular rate laws…) (Kim *et al.*, 2018). These approximate kinetic equations have the advantage of simplifying the modeling, but they cannot help with estimating the parameters. Moreover, particular attention is given to the kinetic parameters that need to be as close as possible to *in-vivo* kinetics. This can be done during enzyme analysis by bringing the *in-vitro* conditions closer to the *in-vivo* conditions (Kim *et al.*, 2018). Therefore, the consideration of these different factors may impact the process of model design but also the upstream research that is done to study metabolic pathways in a particular organism.

3.5.3. Possible model optimizations

Although we have almost accurate prediction results, we can consider additional improvements of the different models. Actually, as this analysis is only made on the second part of glycolysis, it could be envisioned to merge it with the first part of this metabolic pathway to investigate the changes in terms of C_E^J and pathway flux control, and then compare the results to the previous ones, where the parts were modeled separately (Moreno-Sánchez, Encalada, *et al.*, 2008). It would be interesting to have a detailed kinetic model of glycolysis in *E. histolytica* combined with other major metabolic pathways (glycogen metabolism, pentose phosphate pathway) (Saavedra *et al.*, 2007), to highlight the need to inhibit or not the main controlling enzymes identified here, as was done for cancer cells (Marín-Hernández *et al.*, 2016). Also, the addition of genetic-level regulations could help to better understand parasite metabolism, as is done for *E. coli* (Khodayari and Maranas, 2016). However, in order to do this, we still need experimental data on gene expression and regulation in the parasite under conditions of infection.

Also, another way to optimize the models could be by parameter estimation of the unknown kinetic parameters in the UUBB equation. Here, we tried to estimate these parameters, defined first arbitrarily, but the parameter estimation results in very little improvement of the flux prediction with the use of the new estimated values. Actually, parameterization of kinetic constants can provide a mathematical solution to the problem with unrealistic values likely to be physiologically unlikely. Hence, the importance of performing parameter estimation with constraints, within intervals that may be possible in enzymes and may have physiological meaning (*e.g.* K_m or K_i

values not surpassing the lower mM interval). This emphasizes again the need for more experimental data concerning the PPDK mechanism in *in-vivo* conditions. Additionally, in kinetic models, parameters can be determined in two ways, as we have done, either one at a time or collectively; the only difference being that some parameters are often set to measured values (Almquist *et al.*, 2014; Kim *et al.*, 2018). We can also consider the use of different parameter estimation techniques. As demonstrated in a previous work, kinetic parameters can be estimated with the flux balance analysis constraint-based modeling approach, by integrating multi-omics data in the model (fluxomic, proteomic and metabolomics data) (Cotten and Reed, 2013). Consequently, additional work needs to be done involving this part of the modeling, to improve our white-box model using a UUBB equation; it would also be interesting to integrate the data from the grey-box approach into the next parameter estimation procedure.

3.5.4. Biological insights

With the MCA method (C_E^J) and with all models, we identified PGAM as the main controlling enzymes of the second part of glycolysis in this parasite, with a slight contribution of ENO. These results are supported by other studies conducted on this particular pathway (Saavedra *et al.*, 2007; Moreno-Sánchez, Encalada, *et al.*, 2008). Furthermore, it has been found by elasticity analysis, another experimental approach of MCA, that the group of enzymes from PP_i-dependent phosphofructokinase to PPDK controls about 0.2-0.28 of the pathway flux of amoebal glycolysis (Pineda *et al.*, 2015). Within this pathway section, PGAM is the enzyme with the lowest activity in the cell (Saavedra *et al.*, 2007), which may contribute to the better control observed. Additionally, novel enzyme inhibitors were recently identified and tested *in-vitro* (Othman *et al.*, 2017; Stephen *et al.*, 2008). Therefore, these models may be an interesting subject of future research in which the inhibitor effect on the flux can be assessed.

3.6. Conclusion

Be it for the purpose of designing new valuable enzymatic pathways for industrial-scale production of molecules of interest or designing new efficient drugs, metabolic pathway modeling remains a great challenge today (Rajasethupathy *et al.*, 2005; Eriksen *et al.*, 2014; Church and Regis, 2012). Different techniques of modeling exist, including kinetic modeling, based on the use of kinetic parameters and equations that are not necessarily known or experimentally measured. Moreover, several machine learning-based methods are emerging for analysis of metabolic pathway modeling (Cuperlovic-Culf, 2018; Costello and Martin, 2018).

In this study, our objective was to compare three different modeling approaches to model metabolic pathways and identify the main controlling enzymes of the pathway. To this end, we used an application example (lower part of glycolysis of a parasite) and obtained:

- The white-box approach, with the use of all known kinetic information about PGAM, ENO and PPDK. This method gave better results after the modification of the PPDK kinetic equation from ter-reactant reversible equation to UUBB equation (Training: R^2 = 0.95, RMSE = 2.06 nmol·min^{-1} and MAE = 1.7 nmol·min^{-1} and AIC= 336.05 and Test: R^2 = 0.97, RMSE = 2.19 nmol·min^{-1} and MAE = 2 nmol·min^{-1}).
- The grey-box approach, with the kinetic equation with an added adjustment term for PPDK; this model was the best of our models (Training: R^2 = 0.98, RMSE = 1.71 nmol·min^{-1}, MAE = 1.47 nmol·min^{-1} and AIC = 244.71 and Test: R^2 = 1, RMSE = 0.23 nmol·min^{-1} and MAE = 0.13 nmol·min^{-1}).
- The black-box method, using the ANN method to predict the pathway flux. This model presents a low capacity of generalization since its high AIC (539.06) makes it one of the least preferred models here. Nonetheless, the speed and the low cost of this method make it interesting to develop. The model had a good predictive ability with Training: cvR^2 = 1, cvRMSE = 0.52 nmol·min^{-1}, cvMAE= 0.37 nmol·min^{-1} and AIC=124.21 and Test: R^2= 0.98, RMSE=1.61 nmol·min^{-1} and MAE = 1.37 nmol·min^{-1}.

Also, all these models identified the same enzymes as the main controlling enzymes of the pathway: PGAM and ENO, PPDK not having much influence on the flux in *E. histolytica*.

Despite the need for further improvement, these models showed the relevance of the different methods for their future application in the field of metabolic pathway modeling and drug design, for *in-silico* design starting from various backgrounds.

3.7. Discussion et conclusion du chapitre

Dans ce chapitre, nous avons comparé différents modèles de la voie basse de la glycolyse chez l'amibe *E. histolytica* au niveau de leur prédiction du flux final de la voie. Parmi ces modèles, nous retrouvons : des modèles cinétiques (« *white-box* »), des réseaux de neurones (« *black-box* ») et le modèle hybride boîte-grise développé au **chapitre 2** (« grey-box »).

Afin de déterminer le meilleur modèle de la voie métabolique étudiée, nous avons comparé différentes techniques. Parmi elles, nous retrouvons : i) les modèles basés sur la connaissance de type modèle cinétique (dont certains ont été développés et présentés au **chapitre 2**) portant le nom de *modèle boîte-blanche*, ii) les modèles basés sur l'utilisation de données de type réseaux de neurones artificiels portant le nom de *modèle boîte-noire* et iii) Le modèle hybride présenté au **chapitre 2** portant le nom de *modèle boîte-grise*. Suite à cela, la fiabilité de ces modèles est examinée au moyen d'un ensemble de données expérimentales et générées. Nous avons mené en parallèle une identification des enzymes contrôlant le flux en sortie de la voie, par la détermination de leur coefficient de contrôle de flux.

La modélisation de cette voie métabolique par le biais de ces différentes méthodes a mis en évidence leur efficacité à prédire le flux final à partir de données de départ différentes. En effet, la diversité des outils informatiques existants, nous donne la capacité de nous adapter en fonction des données que nous possédons au commencement d'une étude sur une voie métabolique précise. Avec l'aide des données cinétiques des enzymes couplées aux informations générales sur la voie (concentrations initiales en substrats/cosubstrats/enzymes), nous avons pu créer des modèles boîte-blanche. Dans notre cas, ces modèles présentent une bonne capacité de prédiction, mais une faible capacité de généralisation. Aussi, nous avons émis l'hypothèse, au chapitre précédent, que les modèles de type « boîte-noire » qui n'utilisent pas les paramètres cinétiques des enzymes, pourraient donner de meilleures prédictions. Et avec l'aide de données expérimentales, ces modèles boîte-noire ont été bâtis et donnent les meilleurs résultats de prédiction du flux. Ce qui est intéressant pour la suite, puisque ces modèles d'apprentissage automatique sont peu utilisés pour prédire le flux d'une voie métabolique (Antoniewicz *et al.*, 2006; Wu *et al.*, 2016; Ajjolli Nagaraja *et al.*, 2019; Zhou *et al.*, 2020). Malheureusement, pour l'étude de cette voie métabolique, ils ne présentent pas de bonne capacité de généralisation lors de cette étude. Sachant que nous avons peu de données expérimentales, nous ne pouvons conclure réellement sur la supériorité de ce modèle sur les autres pour prédire le flux d'une voie. En dernier lieu, nous sommes revenus sur l'analyse de notre modèle boîte-grise, qui a été construit à la fois à partir des connaissances que

l'on avait sur la voie métabolique, mais aussi à partir de données expérimentales. Ce modèle est celui qui nous offre les meilleurs résultats en termes de généralisation. Aussi, si chacun de ces modèles offre globalement une bonne prédiction du flux, le modèle boîte-grise se démarque des autres, en offrant un bon compromis entre précision et complexité.

Cette étude a confirmé parallèlement les enzymes jouant un rôle dans le contrôle du flux de la voie étudiée (Moreno-Sánchez, Encalada, *et al.*, 2008). En effet, les coefficients de contrôle de flux ont été calculés pour chacun des modèles et ont été ensuite comparés aux coefficients issus de la reconstitution *in-vitro*. Ainsi, PGAM et ENO sont identifiées comme les enzymes les plus importantes dans le contrôle du flux de la partie basse de la glycolyse. Il serait intéressant d'ajouter à ces modèles la partie haute de la voie, afin d'établir les enzymes contrôlant le flux sur toute la voie. Nous avons remarqué que les différents modèles arrivaient plus ou moins bien à obtenir les mêmes coefficients de contrôle obtenus en *in-vitro*. Les meilleurs modèles pour l'identification de ces enzymes sont les modèles boîte-blanche et le modèle boîte-grise. Nous pouvons supposer que ces meilleurs résultats obtenus avec ces modèles sont dus à l'intégration de données cinétiques « précises » sur les enzymes ; contrairement aux modèles boîte-noire qui ne contiennent que des séries d'activités enzymatiques reliées au flux final mesuré. Ces modèles de type boîte-noire sont, effectivement, les moins adéquats pour l'identification des enzymes contrôlant la voie basse de la glycolyse de ce parasite. Cela viendrait confirmer le fait que plus un modèle est détaillé, plus il prédit avec précision des informations de la voie étudiée. Si nous menons notre réflexion sur le côté thérapeutique de ces travaux, nous pouvons en déduire que les enzymes PGAM et ENO feraient de meilleures cibles thérapeutiques, par rapport à PPDK, qui fait l'objet de plusieurs recherches dans ce domaine (Eubank and Reeves, 1982; Saavedra *et al.*, 2004; Stephen *et al.*, 2008; Othman *et al.*, 2017). Dans une optique de production de molécules, l'identification des enzymes contrôlant le flux final de la voie est une étape très importante ; car elle nous permet de :

- Identifier les étapes clés du système de production ;
- Cibler les modifications génétiques à effectuer pour améliorer l'activité de l'enzyme qui contrôle la voie ;
- Ajouter des réactions supplémentaires au système afin de « consommer » les inhibiteurs des réactions catalysées par ces mêmes enzymes.

Pour clore ce chapitre, notre étude comparative met en valeur la compétence de trois types de modèles (modèles cinétiques, modèles réseaux de neurones artificiels et modèle boîte-grise) à représenter une voie métabolique à partir de données différentes. Même si certains sont encore à améliorer, un modèle est resté en tête de liste, tant au niveau de sa précision et de sa complexité, le

modèle boîte-grise. Nous prenons donc la décision de continuer avec ce modèle pour l'implémentation d'un système de contrôle au sein d'une voie métabolique, avec comme objectif final la production de molécules.

Bien que les modèles d'apprentissage automatique n'aient pas donné les meilleurs résultats en matière de complexité et de capacité de généralisation, il n'en demeure pas moins que leurs résultats en apprentissage (avec le « training set ») ont surpassé ceux des autres modèles. Ils représentent donc des techniques avec un fort potentiel pour la modélisation de voie métabolique. De plus, il a été constaté que peu d'études abordent la comparaison de techniques d'apprentissage automatique pour ce type de problématique (Zhou *et al.*, 2020). Ainsi, il serait intéressant de focaliser notre étude sur l'utilisation de ces modèles, appliqués, cette fois, à différents ensembles de données beaucoup plus importants.

Dans ce chapitre nous avons vu comment bâtir des modèles de voies métaboliques dotés d'une bonne performance générale et adaptés aux données initiales que nous avions. L'évaluation de l'efficacité de ces modèles a légitimé notre choix final, portant sur le modèle boîte-grise. Mais au vu de la capacité des modèles d'apprentissage automatique, nous suggérons, au chapitre suivant, de construire cette fois plusieurs modèles différents de Machine-Learning (ML) afin de décrire trois voies métaboliques différentes : (i) la voie basse de la glycolyse du parasite *E. histolytica*, (ii) la voie de détoxification du peroxyde chez le parasite *Trypanosoma cruzi* (González-Chávez et al., 2015) et (iii) le processus de fermentation de la pénicilline à l'échelle industrielle de *Penicillium chrysogenum* (Goldrick *et al.*, 2015).

Chapitre 4

Modélisation de voies métaboliques par des méthodes de Machine-Learning

Ce chapitre est consacré à la modélisation de trois voies métaboliques par des méthodes d'apprentissage automatique (« *Machine Learning* » ou ML). Différents modèles sont établis pour différentes raisons : 1) évaluer la capacité de ce type de méthode à prédire le flux final d'une voie ; 2) examiner la capacité d'une méthode ML donnée à être appliquée à diverses voies métaboliques et 3) déterminer les critères de sélection d'une méthode optimale de modélisation selon les données initiales.

Comme nous l'avons précisé dans le **chapitre 1**, nous procédons ici à la modélisation de plusieurs voies métaboliques, uniquement par le biais de méthodes d'apprentissage automatique. Les trois voies métaboliques auxquelles nous nous intéressons dans cette partie sont :

- La voie basse de la glycolyse du parasite *Entamoeba histolytica* (Moreno-Sánchez *et al.*, 2008) : l'une des principales voies de synthèse d'énergie chez ce parasite ; que nous avons, par ailleurs, étudiée aux chapitres précédents (**chapitre 2 et 3**) ;
- La voie de détoxification du peroxyde chez le parasite *Trypanosoma cruzi* (González-Chávez et al., 2015) : responsable de l'élimination de peroxyde présent dans le microorganisme ;
- Le voie métabolique de fermentation du glucose en pénicilline chez le champignon *Penicillium chrysogenum* (Goldrick *et al.*, 2015) : utilisée au niveau industriel pour la production de pénicilline.

Plusieurs techniques d'apprentissage automatique sont utilisées et peuvent être catégorisées en deux groupes : les méthodes linéaires et celles qui sont non-linéaires. Une revue récente a examiné les diverses possibilités d'utilisation de méthodes ML (linéaires ou non-linéaires) pour augmenter les flux des voies métaboliques (Zhou *et al.*, 2020). Malheureusement, cette étude n'a pu statuer sur l'efficacité ou non des méthodes de ML dans l'augmentation du flux des voies métaboliques, à cause de la rareté des exemples de comparaison de techniques entre elles. Il en est de même pour ce qui est de la problématique qui nous intéresse ici : la prédiction de flux. Nous nous intéresserons alors à cet aspect lors de ces travaux. Par ailleurs, nous n'avons trouvé aucune étude traitant de la distinction entre les modèles linéaires ou non-linéaires pour modéliser une voie métabolique visant à la production de molécule. Ces recherches auront alors pour vocation de donner une première réponse à cette problématique. Ainsi, nous pourrons à la fois évaluer la véritable capacité de ces modèles basés sur les données à prédire le flux et sélectionner les algorithmes les plus adéquats pour le faire.

Les travaux menés dans ce chapitre peuvent être regroupés en trois étapes:

- La première consiste à rassembler les données expérimentales sur chaque voie et à les analyser. Cette analyse servira à une première caractérisation des données et plus globalement de la voie de synthèse étudiée. Suite à cela, des modifications pourront être apportées à l'ensemble de données de départ. Ces modifications consistent en l'addition de données supplémentaires, générées par un modèle robuste de la voie, à l'ensemble de données initiales. Les modèles utilisés pour générer de nouvelles données sont des modèles boîte-grise. Le premier, modélisant la voie basse de la glycolyse, a été développé dans les chapitres qui précèdent celui-ci et le second, représentant la voie de détoxification du peroxyde, a été créé dans cette partie de l'étude, à l'aide du logiciel COPASI.

- La seconde étape consiste à créer les modèles ML sur RStudio, en utilisant un package très connu : caret (Kuhn, 2020). De nombreux modèles sont ainsi construits : allant de la régression par les moindres carrés partiels (« *Partial Least Squares* » ou PLS) aux forêts aléatoires (« *Random forests* » ou RF) en passant par les réseaux de neurones artificiels.

- Une fois les modèles construits, nous passons par une dernière étape d'évaluation de la fiabilité de ces modèles.

Contrairement aux modèles cinétiques qui peuvent parfois être lourds à mettre en place, les modèles d'apprentissage automatique requièrent uniquement un ensemble de données pour

modéliser une voie d'intérêt. Aussi, même si elles sont décrites comme étant des méthodes robustes dans d'autres domaines (Mitchell, 2014; Tami *et al.*, 2019; Ambrosen *et al.*, 2020; Gerdes, 2021), la question reste en suspens pour notre problématique. Ces approches sont très prometteuses, d'où l'intérêt de les considérer dans notre travail sur la modélisation de voie métabolique pour la production de molécule.

Les travaux présentés ci-dessous ont été soumis au journal *Scientific Reports*.

Non-linearity of metabolic pathways critically influences the choice of machine learning model.

Ophélie Lo-Thong-Viramoutou, Philippe Charton, Xavier F. Cadet, Brigitte Grondin-Perez, Emma Saavedra, Cédric Damour and Frédéric Cadet

ABSTRACT

The use of machine learning (ML) in life sciences has gained wide interest over the past years, as it speeds up the development of high performing models. Important modeling tools in biology have proven their worth for pathway design, such as mechanistic models and metabolic networks, as they allow better understanding of mechanisms involved in the functioning of organisms. However, little has been done on the use of ML to model metabolic pathways, and the degree of non-linearity associated with them is not clear. Here, we report the construction of different metabolic pathways with several linear and nonlinear ML models. Different types of data are used; they lead to the prediction of important biological data, such as pathway flux and final product concentration. A comparison reveals that the data features impact model performance and highlight the effectiveness of nonlinear models (*e.g.*, QRF: RMSE=0.021 nmol·min^{-1} and R^2=1 versus Bayesian GLM: RMSE=1.379 nmol·min^{-1} R^2=0.823). It turns out that the greater the degree of non-linearity of the pathway, the better suited a nonlinear model will be. Therefore, a decision-making support for pathway modeling is established. These findings generally support the hypothesis that nonlinear aspects predominate within the metabolic pathways. This must be taken into account when devising possible applications of these pathways for the identification of biomarkers of diseases (*e.g.*, infections, cancer, neurodegenerative diseases) or the optimization of industrial production processes.

4.1. Introduction

Machine learning (ML) holds an increasingly prominent place in the field of biology. Indeed, it can lead to better results and has a large range of applications including: drug design using machine leaning algorithms such as the support vector machine (SVM) algorithm to perform structure-activity relationship analysis (Burbidge *et al.*, 2001; Réda *et al.*, 2020; Hartwell *et al.*, 1999); directed protein evolution and enzyme function prediction (Li *et al.*, 2018; Wu *et al.*, 2019); reconstruction of biological routes (Kotera *et al.*, 2013; Baranwal *et al.*, 2020) or modeling and optimization of metabolic pathways (Zhang *et al.*, 2019; Kim *et al.*, 2020). With regard to the latter topic, several methods have been developed to analyze complex biological systems (figure 4.1):

- The **knowledge-based model** including kinetic models (Chance, 1943; Sel'Kov, 1968; Curto *et al.*, 1997; Hatzimanikatis *et al.*, 1998; Curto *et al.*, 1998; Visser and Heijnen, 2003; Liebermeister *et al.*, 2010) and metabolic flux analysis methods (Fell and Small, 1986; Stephanopoulos, 1999);
- The **data-based model** including ML algorithms and ensemble learning (Zelezniak *et al.*, 2018; Oyetunde *et al.*, 2019; Ajjolli Nagaraja *et al.*, 2019);
- The **hybrid model** including combinations of models or modified preceding methods (Cascante *et al.*, 2002; Morgan and Rhodes, 2002).

Although, these analyses are conducted on metabolic pathways, few of them are used to predict their fluxes. Among these few works on metabolic fluxes, it is interesting to highlight those of Ajjolli Nagaraja *et al.* on which we collaborated (Ajjolli Nagaraja et al., 2019). For the present work, the method of greatest interest is the data-based model and more precisely, ML. In fact, ML abounds in various methods and is a promising and growing approach that could greatly help to improve existing models, integrate multi-omics data and save researchers' time. Also, a distinction can be made between ML methods: some are linear (ridge and lasso regression, multivariate adaptive regression spline…) and others are nonlinear (artificial neural network, k-nearest neighbors, decision tree…). However, it has not yet been investigated whether linear or nonlinear methods are more efficient in predicting pathway fluxes, and how to choose the appropriate one.

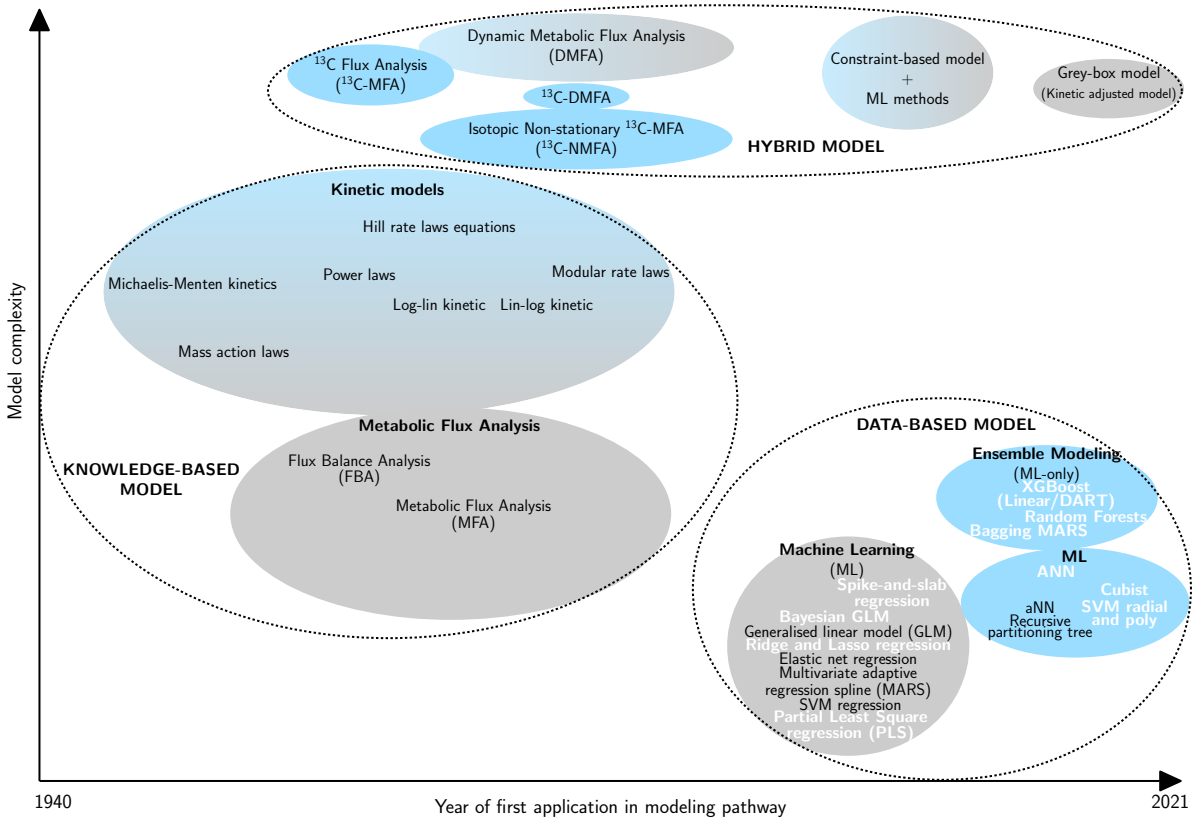

FIGURE 4.1 : *Classification of metabolic pathway modeling methods according to their complexity and the year of first application in this field.*
The ellipse size is proportional to the occurrence of the method for pathway modeling in the literature. Linear methods are represented in grey and nonlinear ones are in blue. Methods in bold and white are those evaluated in this study.

Therefore, this study aims to elucidate the most appropriate methods to model three distinct metabolic pathways by designing and comparing five linear and eight nonlinear machine learning-based methods (figure 4.2):

- The lower part of *Entamoeba histolytica* glycolysis, one of the major metabolic pathways of the parasite (Pineda *et al.*, 2015; Muller *et al.*, 2012; Moreno-Sánchez *et al.*, 2008) , through the use of a recently developed model (Lo-Thong *et al.*, 2020);
- The peroxide detoxification pathway of *Trypanosoma cruzi* (González-Chávez *et al.*, 2015, 2019);
- The industrial-scale penicillin fermentation process of *Penicillium chrysogenum* (Goldrick *et al.*, 2015).

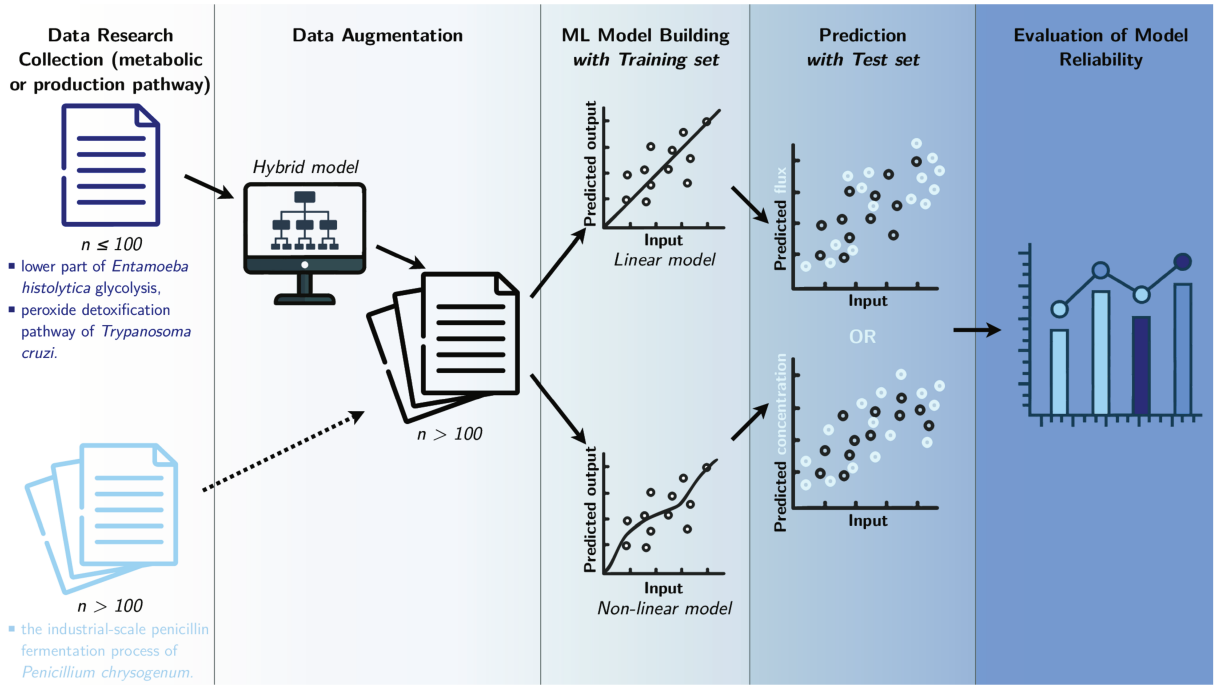

FIGURE 4.2 : *Study workflow.*

Data from three different metabolic pathways are collected and used to build data-based models. Datasets that contain a small amount of data (n) go through a process of data augmentation, before being separated into two sets: training set and test set. Then, in order to predict the final flux or final product concentration, multiple ML models are built with the training set, while the test set is used to assess the final models. The resulting predictions are compared in a last step to evaluate model reliability.

Although these machine-learning approaches have been used to model metabolic pathways, few studies have focused on their usefulness in predicting flux (Ajjolli Nagaraja *et al.*, 2019; Wu *et al.*, 2016).

Creating an efficient ML model depends on the availability of a large amount of experimental data (Schmidt *et al.*, 2019; L'Heureux *et al.*, 2017). The measurement of fluxes is cumbersome to carry out experimentally and hinders the possibility of having massive data. Because of the scarcity of these large experimental datasets in the literature, the methodology employed here consists of applying data augmentation to the first two pathways by using hybrid models (figure 4.2). These hybrid models, called grey-box models, often predict better results than knowledge-based models or data-based models (Lo-Thong *et al.*, 2020; Wei *et al.*, 2018; Pintelas *et al.*, 2020); in this study, the grey-box models consist of metabolic networks that include an adjustment term in one or more kinetic equations.

In this work, models are based both on experimental datasets and predicted data coming from the previous grey-box model. Here, we show that random forest models are the most effective, with a high predictive capacity starting from predicted and experimental enzyme activities or

experimental parameters collected from a bioreactor. Also, two other models stand out as good ways to predict the flux or the final product concentration: XGBoost Linear and Cubist models. This shows the importance of using a nonlinear model to design metabolic pathways. Based on these findings, we propose a means of decision support for researchers who wish to use machine learning techniques as a starting or a complementary method for modeling and for improving existing biological pathway models. By greatly increasing the quality of the outputs (flux prediction), machine learning opens the way to better drug target identification within a pathway, efficient disease modeling at molecular level and more efficient optimization for industrial production of metabolites.

4.2. Methods

4.2.1. Experimental procedures

The lower part of glycolysis is reconstituted *in-vitro* in a reaction assay medium described in our previous work (Moreno-Sánchez *et al.*, 2008), containing different recombinant enzymes (PGAM, ENO and PPDK). The reaction was started by adding 3PG (4 mM). An additional reaction is added, the formation of lactate with lactate dehydrogenase (figure 4.3), in order to follow the flux of the overall pathway by following the rate of NADH oxidation, for more details, see our works (Moreno-Sánchez *et al.*, 2008).

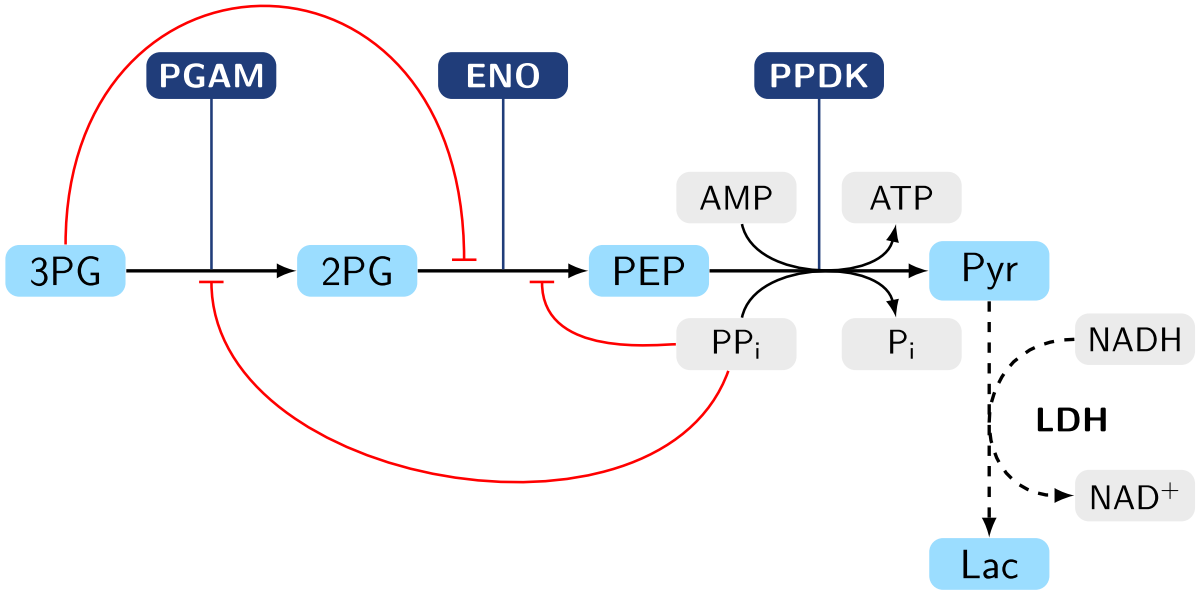

FIGURE 4.3 : *Lower part of E. histolytica glycolysis pathway. Formation of pyruvate (Pyr) from 3-phosphoglycerate (3PG).*
The L-lactate (Lac) formation (dashed lines) is not part of the natural pathway; however, lactate dehydrogenase (LDH) has been added in order to experimentally follow the final flux and establish a quasi-steady-state to Lac (Moreno-Sánchez et al., 2008). Metabolite inhibitions are represented in red. PGAM, 3-phosphoglycerate mutase; 2PG, 2-phosphoglycerate; ENO, enolase; PEP, phosphoenolpyruvate; PPDK, pyruvate phosphate dikinase. This pathway is also shown in ***chapter 2*** *in figure 2.2.*

Concerning the peroxide detoxification pathway (figure 4.4), each enzyme was individually titrated, while keeping the other parameters in the *in-vitro* system constant. The pathway flux was determined in parallel by observing NADPH oxidation, see González-Chávez *et al.* in our previous work for more information (González-Chávez *et al.*, 2015).

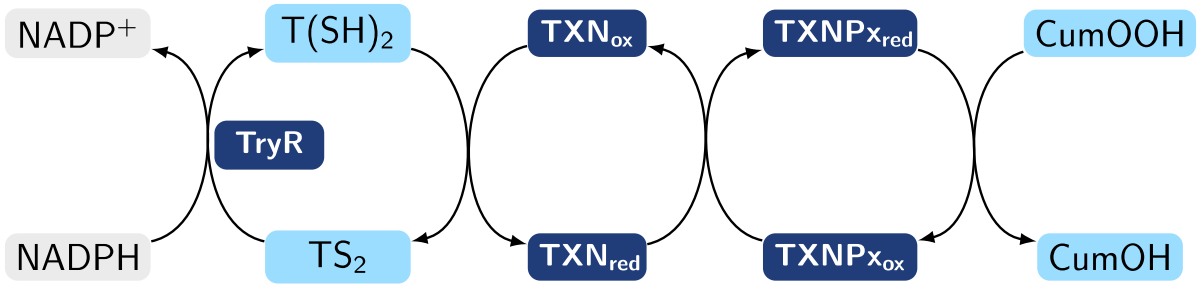

FIGURE 4.4 : *Tryparedoxin-dependent hydroperoxide detoxification pathway in Trypanosoma cruzi (González-Chávez et al., 2015).*
Reduction of cumene hydroperoxide (CumOOH) is assessed here. TryR, trypanothione reductase; $T(SH)_2$, trypanothione; TS_2, trypanothione disulfide; $TXN_{ox/red}$, oxidized/reduced tryparedoxin; TXNPxox/red, oxidized/reduced tryparedoxin peroxidase.

Finally, the experimental procedures that were followed to obtain penicillin production data are described in the studies of Goldrick *et al.* (Goldrick *et al.*, 2015).

4.2.2. Lower part of glycolysis datasets

Two datasets are constructed here by applying data augmentation, using a grey-box model detailed in one of the following sections. For the first one, an exploration around the experimental data flux (43±10 nmol·min^{-1}) from R. Moreno-Sánchez *et al.* at pH 6 is conducted (Moreno-Sánchez *et al.*, 2008). In fact, a sample of 2,000 normally distributed enzymatic balances was generated with the *sample* function on RStudio and resulted in a predicted flux between 0-53 nmol·min^{-1} with the grey-box model. The term balance refers to a set of concentrations of the enzymes involved in the cascade of reactions. The second cascade is made up of experimental and predicted (grey-box model) data of PGAM, ENO and PPDK activities and pathway flux (J). The experimental data are obtained from plots of a previous study (Moreno-Sánchez *et al.*, 2008), only the dots were used, while the predicted data are obtained with the grey-box model developed in our previous work (Lo-Thong *et al.*, 2020), by varying each enzyme activity from 0 to 1000 mU with a step of 25 mU. These datasets are shown in Tables A.4 and A.5 respectively (*Appendix A*).

4.2.3. Peroxide detoxification datasets

The second studied pathway consisted first of 58 experimental enzymatic balances and their corresponding flux. After applying data augmentation by using a grey-box model of this pathway, a bigger dataset of 1,671 data was obtained. As with the previous dataset, a combination of data

normally distributed is generated with the *sample* function on RStudio, resulting in a predicted flux ranging from 0 to 11.46 nmol·min^{-1}. The new dataset is a mix of the previous experimental data and new predicted data of enzyme activities (TryR, TXN and TXNPx); final flux and is shown in Table B.1 (*Appendix B*).

4.2.4. The grey-box models

*The detailed construction of this model was done in **chapter 2**; nevertheless, we briefly revisit this part, in order to highlight the creation of the grey-box model for the peroxide detoxification pathway.*

The two following pathways are modeled with an open-source software called COPASI (Version 4.24) (Hoops *et al.*, 2006): the second part of glycolysis and the peroxide detoxification pathway. This software is used for metabolic network design, analysis and optimization. The first grey-box model, representing the lower part of glycolysis, is taken from our previous work (Lo-Thong *et al.*, 2020). It is based on the use of enzyme properties, including kinetic parameters and kinetic equations. To enhance the flux predictions, we suggested adding an adjustment term to the PPDK kinetic equation. The whole process concerning the composition of this term is explained in our work (see Methods part of **Chapter 2**).

The second grey-box model represents the peroxide detoxification pathway and is built specifically for this study. It contains kinetic parameters and equations of three enzymes: TryR, TXN and TXNPx (table 4.1).

Enz.	Kinetic equations
TryR[a]	$\dfrac{V_f\frac{AB}{K_{mA}K_{mB}} - V_r\frac{PQ}{K_{mP}K_{mQ}} + \alpha(V_f - V_{f0})}{1 + \frac{A}{K_{mA}} + \frac{B}{K_{mB}} + \frac{P}{K_{mP}} + \frac{Q}{K_{mQ}} + \frac{AB}{K_{mA}K_{mB}} + \frac{AP}{K_{mA}K_{mP}} + \frac{BQ}{K_{mB}K_{mQ}} + \frac{PQ}{K_{mP}K_{mQ}} + \frac{ABP}{K_{mA}K_{mB}K_{mP}} + \frac{BPQ}{K_{mB}K_{mP}K_{mQ}}}$
TXN[b]	$\dfrac{V_f\left(AB - \frac{PQ}{K_{eq}}\right)}{\mathrm{AB} + K_{mB}A + K_{mA}B\left(1 + \frac{Q}{K_{iQ}}\right) + \frac{V_f}{V_r K_{eq}}\left[K_{mQ}P\left(1 + \frac{A}{K_{iA}}\right) + Q\left(K_{mP} + P\right)\right]}$
TXNPx[c]	$\dfrac{V_f\left[CumOOH\right][TXN_{red}] + \beta\left(V_f - V_{f0}\right)}{K_{mTXN_{red}}\left[CumOOH\right] + K_{mCumOOH}[TXN_{red}] + \left[CumOOH\right][TXN_{red}]}$

TABLE 4.1 : Kinetic equations used in the grey-box model of the peroxide detoxification pathway (González-Chávez *et al.*, 2015).

[a] A, B and K_{mA}, K_{mB} are respectively the concentrations and K_m of the substrates NADPH and TS_2; P, Q and K_{mP}, K_{mQ} are the concentrations and K_m of the products $NADP^+$ and $T(SH)_2$; $\alpha(V_f - V_{f0})$ is

the adjustment term with α, a defined number, V_{f0}, TryR maximum rate in the forward direction used in the *in-vitro* reconstitution and V_f is TryR maximum rate in the forward direction in the model. [b] A, B and K_{mA}, K_{mB} are respectively the concentrations and K_m of the substrates $T(SH)_2$ and TXN_{ox}; P, Q and K_{mP}, K_{mQ} are the concentrations and K_m of the products TS_2 and TXN_{red}. [c] $\beta(V_f - V_{f0})$ is the adjustment term with β, a defined number, V_{f0}, TXNPx maximum rate in the forward direction used in the *in-vitro* reconstitution and V_f is TXNPx maximum rate in the forward direction in the model.

Also, we proposed to add two adjustment terms in TryR and TXNPx equations to improve flux predictions (table 4.1). These are determined in the same way as the terms used for the glycolysis pathway. In fact, a first model was provided by González-Chávez et al. (González-Chávez *et al.*, 2019) and could predict the final flux quite well when TryR and TXN activities were varied. However, it overestimated the flux when TryR activity was varied and underestimated it when TXNPx activity was varied. Therefore, we suggest adding a first adjustment term $\alpha(V_f - V_{f0})$ in order to increase TryR rate and a second adjustment term $\beta(V_f - V_{f0})$ to decrease TXNPx rate. In these adjustment terms, α and β are defined numbers selected as the best for flux prediction from a tested range, V_f is TryR (or TXNPx) maximum rate in the forward direction in the model and V_{f0} TryR (or TXNPx) maximum rate in the forward direction used in the *in-vitro* reconstitution. Also, as V_f of TryR (or TXNPx) is equal to V_{f0} when TXN's/TXNPx's (or TryR's/TXN's) activity is varied, we multiplied α (or β) by $V_f - V_{f0}$, so that the adjustment term would be zero when $V_f = V_{f0}$ and the flux predictions are not modified in these cases mentioned above.

Also, residual values are determined to evaluate how accurate the grey-box model is, and calculated as follows: $e = y - \hat{y}$, where $\boldsymbol{e}$ is the residual, $\boldsymbol{y}$ is the observed value and $\hat{\boldsymbol{y}}$ the corresponding predicted value.

4.2.5. Data augmentation

For the datasets with less than 100 data, a process called data augmentation is performed. It consists of using models that accurately predict the experimental data. Two different grey-box models are used in this study for the lower part of glycolysis pathway, retrieved from our previous study (Lo-Thong *et al.*, 2020), and for the peroxide detoxification pathway (built for the present work).

4.2.6. Dataset analysis

A brief analysis of the datasets is performed, including an examination of data distribution and the calculation of linear correlations between the input and output variables.

4.2.7. Machine learning models building and selection

To model the metabolic pathway, different machine learning models are developed on RStudio (Version 1.2.5001), with the help of Classification And Regression Training (caret, Version 6.0-86) (Kuhn, 2020).

The datasets are split into 80/20 for the training and test sets, and a k-fold cross-validation (with k= 10 for Dataset 1-2 and k=3 for Dataset 3) is performed on the models with the training set.

After this, the best models are selected based on:

- The root-mean-square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \qquad (1)$$

  with $\boldsymbol{Y_i}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ being respectively the observed and predicted values, $\boldsymbol{n}$ being the total number of values and $\mathbf{i} = 1, 2 \ldots n$;

- The coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (2)$$

  with $\boldsymbol{Y_i}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ respectively the observed and predicted values, $\boldsymbol{n}$ being the total number of values and $\mathbf{i} = 1, 2 \ldots n$.

Also, a calculator was used for modeling the metabolic pathways, which has the following characteristics: cluster 2x Intel Xeon E5-2630v4 Broadwell-EP @ 2.20GHz 10 cores, 8x 16GB of RAM, 2400MHz, DDR4, ECC.

The custom codes for the data analysis used in this study are available from the corresponding author in the Github repository: https://github.com/ophelielt/Lo-Thong_et_al._Non-linearity_of_metabolic_pathways_influences_the_choice_of_ML.git.

4.3. Results

As previously mentioned, ML models could have different applications in biology, including the identification of biomarker, i.e., a valuable, quantitative component (metabolites, proteins, enzymes…), within a metabolic pathway for health purposes (diseases diagnosis, treatment) or the optimization of a valuable production pathway. Therefore, we have targeted three different datasets based on these two applications. The first one concerns the lower part of glycolysis in *Entamoeba* histolytica (figure 4.3) and contains a set of enzyme activities for which the final flux has been measured (Moreno-Sánchez *et al.*, 2008). The second pathway is the tryparedoxin-dependent hydroperoxide detoxification pathway in *Trypanosoma cruzi* (figure 4.2), which provides the same type of data as in the previous dataset (González-Chávez *et al.*, 2015). It is important to consider how essential these two previous pathways are, as they play a significant role in the survival of these parasites. Given the small size of the experimental dataset, we use two grey-box models (one developed before (Lo-Thong *et al.*, 2020) and the other developed in this study), to generate a larger dataset for these two pathways (Dataset 1 and 2) before building the ML models (figure 4.2).

The last metabolic pathway modeled here is the penicillin fermentation process in *Penicillium chrysogenum* (figure 4.5). This dataset did not need to be enlarged (Dataset 3), and we used it to build different ML models (figure 4.2).

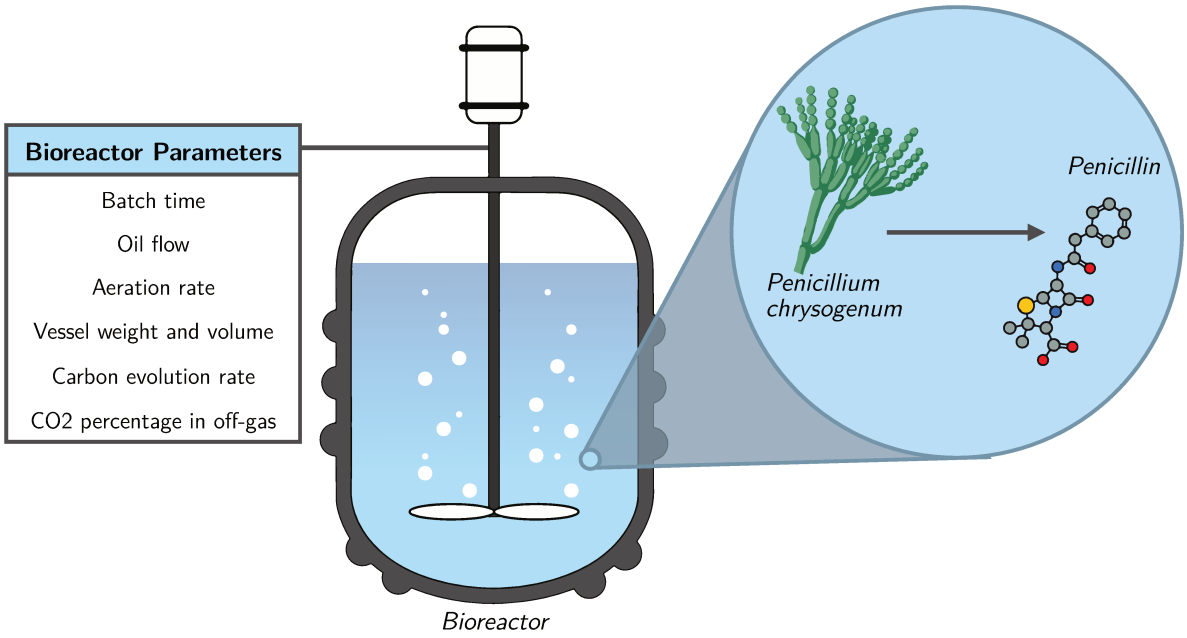

FIGURE 4.5 : *Simplified representation of the industrial-scale penicillin fermentation process of Penicillium chrysogenum.*
The bioreactor parameters represented here are those that will be of interest in this study. See a more detailed scheme of the bioreactor in the work of Goldrick et al. (Goldrick et al., 2015).

4.3.1. Example 1: The lower part of *Entamoeba histolytica* glycolysis

The grey-box model allows the building of huge datasets

Since the amount of experimental data is limited, the first step here is to build a robust model to generate more data.

As explained in the Methods section, the grey-box model developed on COPASI in our previous work contains all kinetic parameters and kinetic equations of PGAM, ENO and PPDK (Lo-Thong *et al.*, 2020). In order to improve the flux prediction, the first two enzymes employ the Michaelis-Menten reversible rate equation, whereas the third employs a modified termolecular reaction reversible rate equation including an adjustment term in the denominator (table 4.2).

Enz.	Kinetic equations
PGAM	$\dfrac{V_f \frac{[3PG]}{K_{m3PG}} - V_r \frac{[2PG]}{K_{m2PG}}}{1 + \frac{[3PG]}{K_{m3PG}} + \frac{[2PG]}{K_{m2PG}} + \frac{[PP_i]}{K_{iPP_i}}}$
ENO	$\dfrac{V_f \frac{[2PG]}{K_{m2PG}} - V_r \frac{[PEP]}{K_{mPEP}}}{1 + \frac{[2PG]}{K_{m2PG}} + \frac{[PEP]}{K_{mPEP}} + \frac{[PP_i]}{K_{iPP_i}} + \frac{[3PG]}{K_{i3PG}}}$
PPDK[a]	$\dfrac{V_f\left(\mathrm{ABC} - \frac{\mathrm{PQR}}{K_{eq}}\right)}{K_{mA}\mathrm{B} + K_{mB}\mathrm{A} + K_{mC}\mathrm{B} + K_{mB}\mathrm{C} + \frac{V_f K_{mQ}\mathrm{P}}{V_r K_{eq}} + \frac{V_f K_{mP}\mathrm{Q}}{V_r K_{eq}} + \frac{V_f K_{mQ}\mathrm{R}}{V_r K_{eq}} + \frac{V_f K_{mR}\mathrm{Q}}{V_r K_{eq}} + \mathrm{ABC} + \frac{V_f \mathrm{PQR}}{V_r K_{eq}} + \alpha\left\lvert V_f - V_{f0}\right\rvert}$

TABLE 4.2 : Kinetic equations used in the grey-box model of the lower part of glycolysis (Lo-Thong *et al.*, 2020).

K_m is the Michaelis constant of the enzyme; K_i is the inhibitor constant; V_f and V_r are maximum rates of the forward and reverse reactions; K_{eq} is the equilibrium constant of the reaction. [a] A, B and C and K_{mA}, K_{mB} and K_{mC} are respectively the concentrations and K_m of the substrates PEP, AMP and PP_i; P, Q and R and K_{mP}, K_{mQ} and K_{mR} are the concentrations and K_m of the products Pyr, ATP, P_i; $\alpha\left|V_f - V_{f0}\right|$ is the adjustment term with α, a defined number, V_{f0}, PPDK maximum rate in the forward direction used in the *in-vitro* reconstitution and V_f is PPDK maximum rate in the forward direction in the model.

The resulting fluxes show good reliability of the model to predict the final experimental flux (R^2 ≈ 0.95 and RMSE=1.993 nmol·min^{-1}), even when enzyme activities are varied (figure 4.6 A-C).

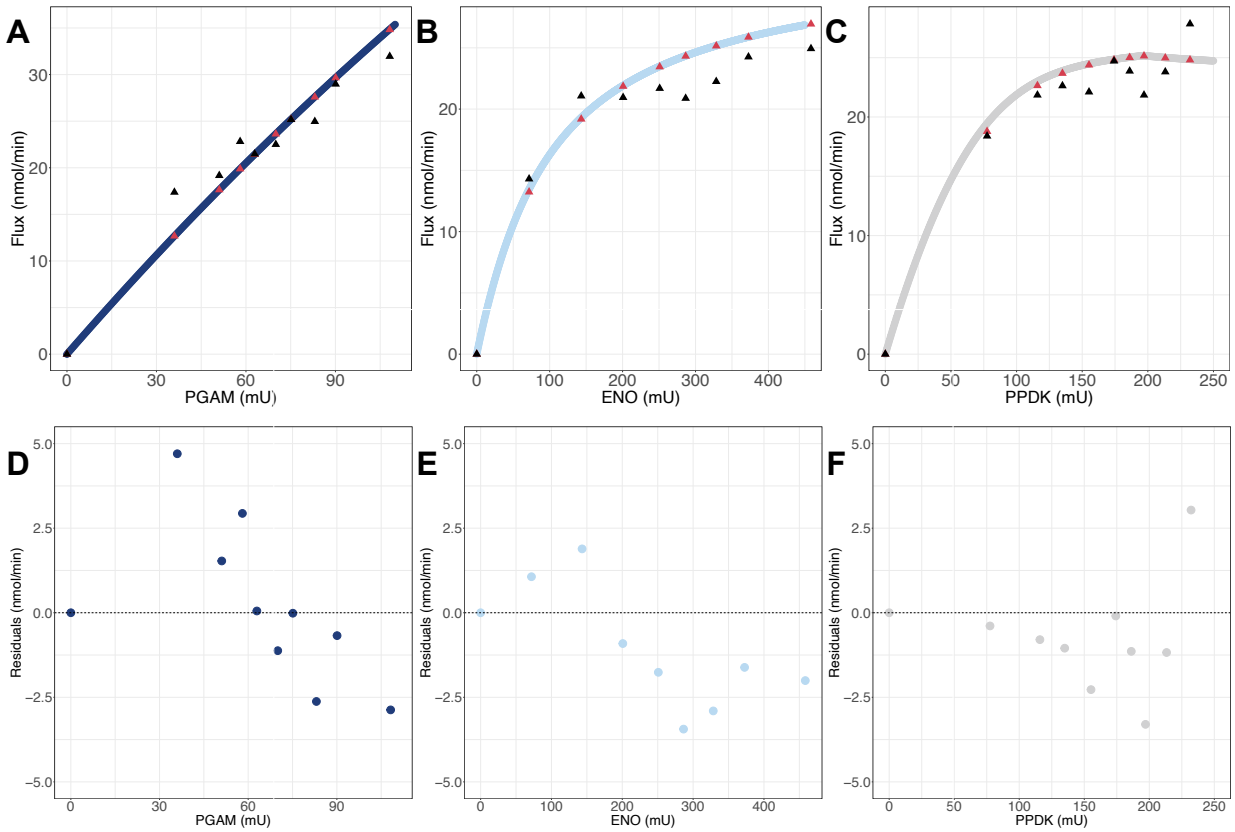

FIGURE 4.6 : *Flux predictions with the grey-box model.*
(A, B, C) *Flux variation according to PGAM* ***(A)****, ENO* ***(B)*** *or PPDK* ***(C)*** *activity. Data are taken from Table A.4 (Appendix A). Experimental fluxes are in black triangles, their corresponding predicted flux values are in red triangles and the predicted fluxes are in dark blue for PGAM, light blue for ENO and grey for PPDK.* ***(D, E, F)****. Residuals for the experimental values versus PGAM* ***(D)****, ENO* ***(E)*** *or PPDK* ***(F)****.*

The calculation of residuals shows a defined pattern that is the same for PGAM and ENO. It reveals a general trend of the model to underestimate the flux for low enzyme activity values, and overestimate it for high enzyme activity values (figure 4.6 D and E). Concerning PPDK, the grey-box model tends instead to underestimate the final flux when the enzyme activity is varied, with an exception for the last point (at 232.13 mU), which is overestimated (figure 4.6 F). The model is quite accurate to predict the pathway flux and presents low residuals between -3.4-4.7 nmol·min^{-1}.

The next step of this work consists of using the *in-silico* model for generating larger datasets, a process we call data augmentation. The first new dataset contains 2,000 enzyme balances evolving around the experimental ones (see Table A.4 in *Appendix A* and figure 4.7). The term balance refers to a set of concentrations of the enzymes involved in the cascade of reactions. The predicted final

fluxes vary between 0 and 60.84 nmol·min^{-1}; the distribution of the other data from the first dataset is shown in figure 4.8.

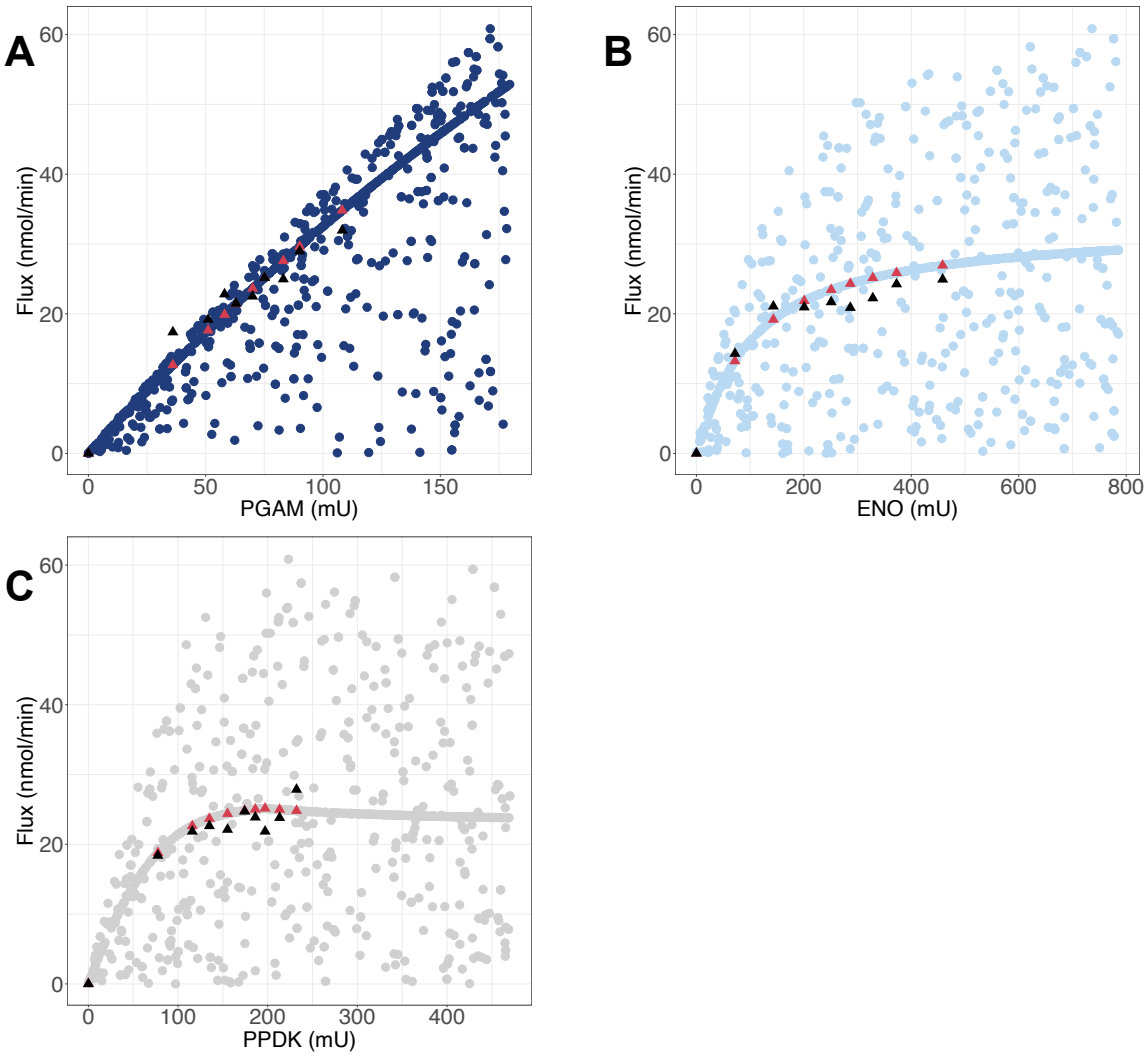

FIGURE 4.7 : *Flux predicted for the first dataset (2,000 data).*

(A, B, C) *Flux variation according to PGAM* ***(A)****, ENO* ***(B)*** *or PPDK* ***(C)*** *activity. Predicted flux from the experimental dataset are in red diamonds and the corresponding experimental flux values are in black diamonds.*

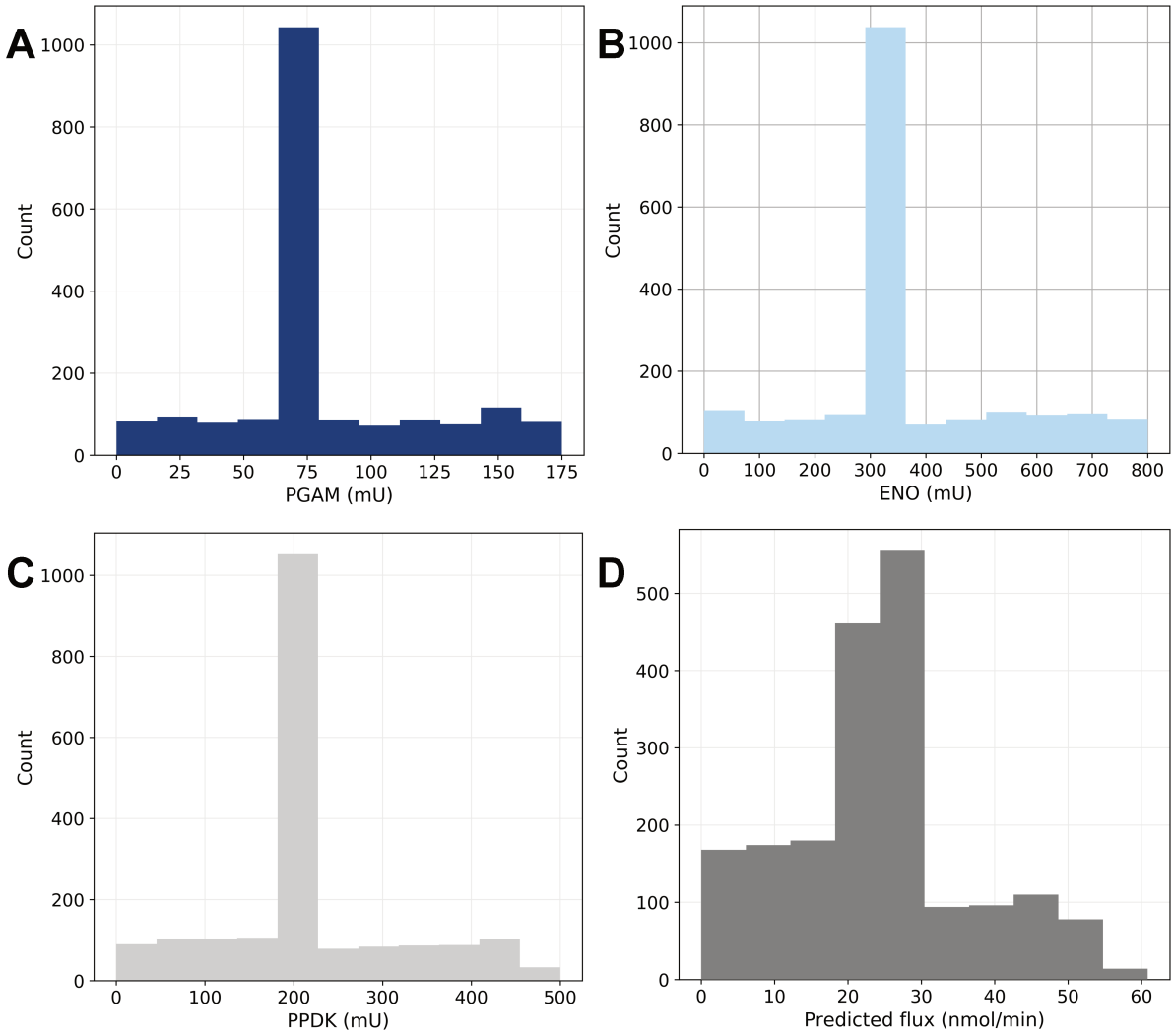

FIGURE 4.8 : *Histogram of the first dataset distribution.*
***(A, B, C)** Variation of PGAM **(A)**, ENO **(B)** or PPDK **(C)** activity. **(D)** Variation of the final flux predicted by the grey-box model.*

In fact, the predicted fluxes count with the highest representation are within the experimental data of the reconstituted pathway (Moreno-Sánchez *et al.*, 2008) and *in-vivo* pathway fluxes in live parasites (Pineda *et al.*, 2015). In order to compare the models, a second dataset (Dataset 1) is generated and includes 68,950 data for which all enzyme activity is varied between 0 and 1,000 mU (see table 4.3 and Table A.5 in *Appendix A*). The final fluxes are then predicted and fluctuate between 0 and 215.45 nmol·min^{-1}; additional information is provided in table 4.3 and figure 4.9.

	PGAM	ENO	PPDK	J_{pred}
Data count	68,950	68,950	68,950	68,950
Mean	499.82	499.92	499.87	83.46
Standard deviation	295.87	295.78	295.82	55.32

Minimum value	0	0	0	0
25%	250	250	250	35.998
50%	500	500	500	80.23
75%	750	750	750	127.6
Maximum value	1,000.0	1,000.0	1,000.0	215.45

TABLE 4.3 : Description of the new generated dataset (Dataset 1).

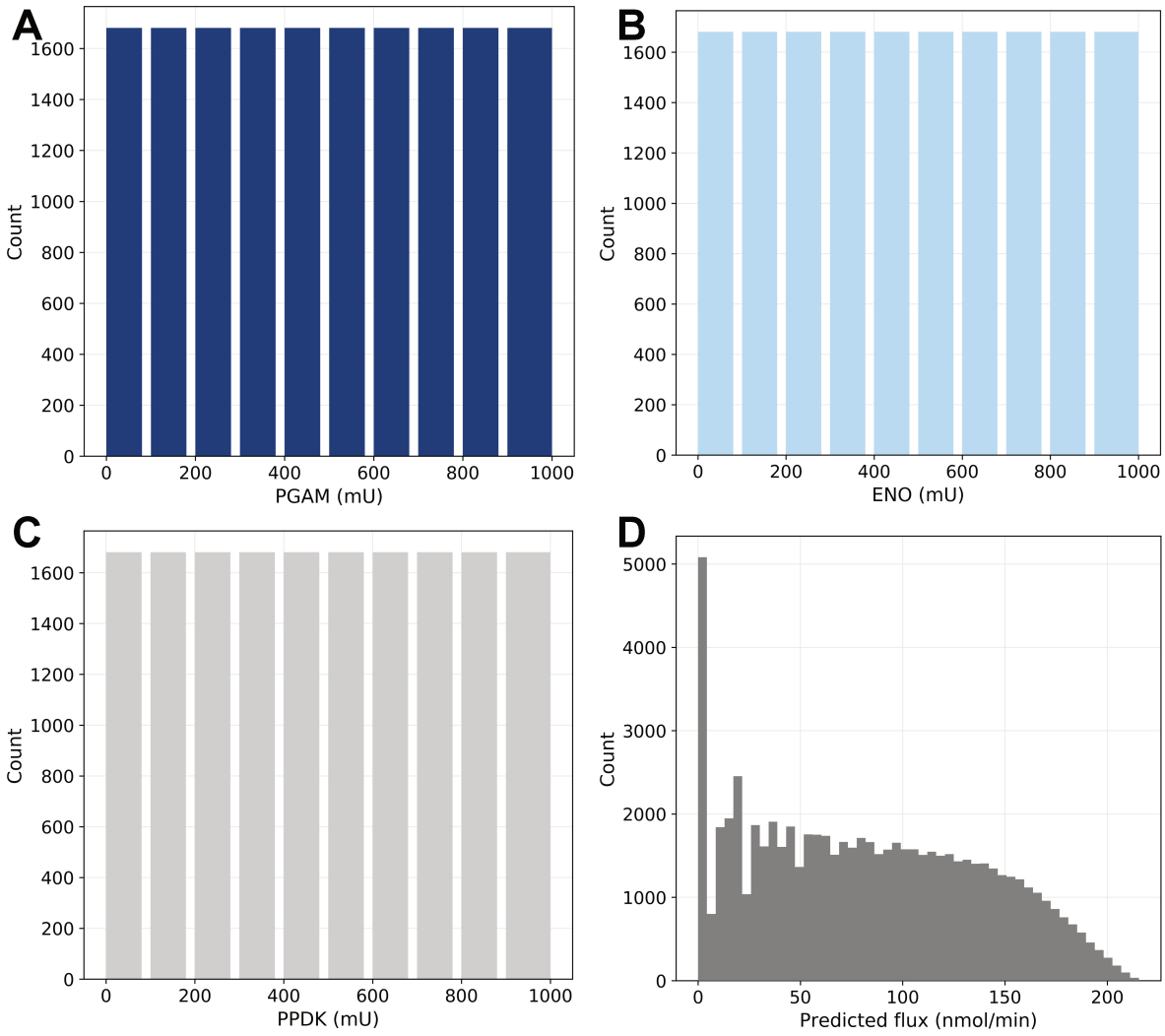

FIGURE 4.9 : *Histogram of the distribution of Dataset 1 (68,950 data).*
***(A, B, C)** Variation of PGAM **(A)**, ENO **(B)** or PPDK **(C)** activity. **(D)** Variation of the final flux predicted by the grey-box model. The physiological in-vivo fluxes are around 50 nmol·min^{-1}.*

We then plotted the final flux in function of the enzyme activity for the largest dataset (Table A.5 in *Appendix A*) and obtained the same type of curve as we did previously (figure 4.7 and figure 4.10). Indeed, variations of PGAM activity have a great impact on the final flux, while those of ENO and PPDK have a lesser impact on the pathway flux (figure 4.10).

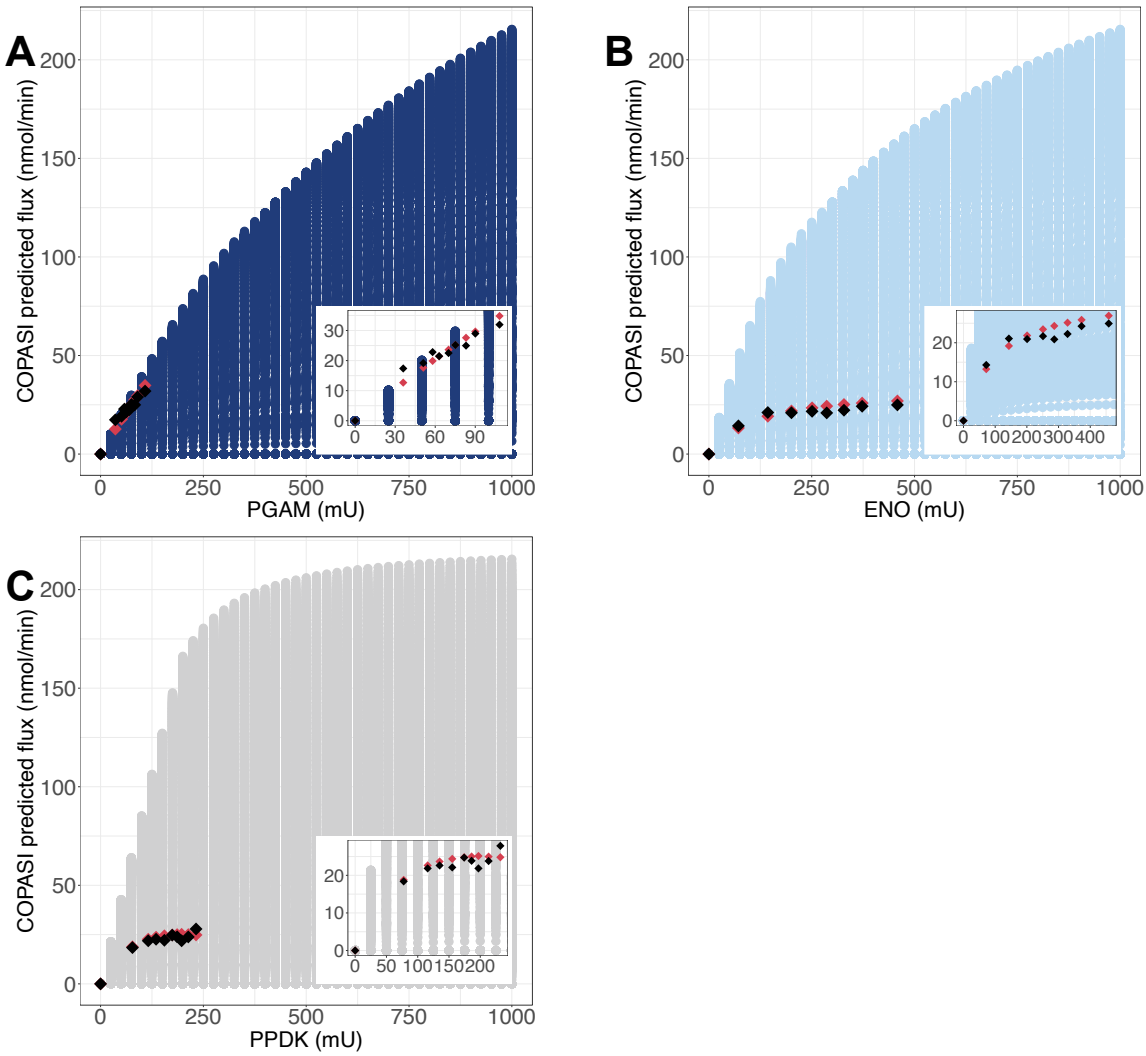

FIGURE 4.10 : *Flux predicted for Dataset 1 (Table A.5 in Appendix A).*
(A, B, C) *Flux variation according to PGAM* ***(A)****, ENO* ***(B)*** *or PPDK* ***(C)*** *activity. Predicted flux from the experimental dataset are in red diamonds and the corresponding experimental flux values are in black diamonds.*

It should also be noticed that the experimental fluxes are in the lower part of the predicted flux values. The insets show a gap between the experimental flux values and the dataset flux values; this difference is due to the intervals between two values, used in the two cases, with the interval being smaller for the experimental dots (7-85 mU) than for the predicted data (25 mU). Following this initial analysis of the data, we assessed the correlation between the various variables. The table of correlation shows that the enzymes and the final flux are correlated to varying degrees, with the highest correlation coefficient for PGAM, followed by ENO, and the lowest coefficient for PPDK (table 4.4). These linear correlation coefficients provide insight into the degree of non-linearity of this metabolic pathway. Indeed, even if the mean value of the correlations is above 0.5 (table 4.4), we observe a weak linear correlation for many ranges of enzyme activity (figure 4.11) when one of the enzymes is varied over the three, for example for PPDK when PGAM varies between 0-625 mU

and ENO between 0-1000 mU (figure 4.11 C). These results indicate significant non-linearity in the metabolic pathway, particularly for PPDK and ENO. In addition, these results lead to the same conclusions as those from flux control coefficient calculations (Lo-Thong *et al.*, 2020): the enzyme exerting the greatest flux control is PGAM, followed by ENO, and PPDK has the weakest control of the pathway flux.

Good quality augmented datasets having been generated; they are used to test different ML approaches in the following section.

	J_{pred}
PGAM	0.90
ENO	0.85
PPDK	0.53

TABLE 4.4 : Table of mean linear correlations between the enzyme activities and the predicted final flux (J_{pred}) for Dataset 1.

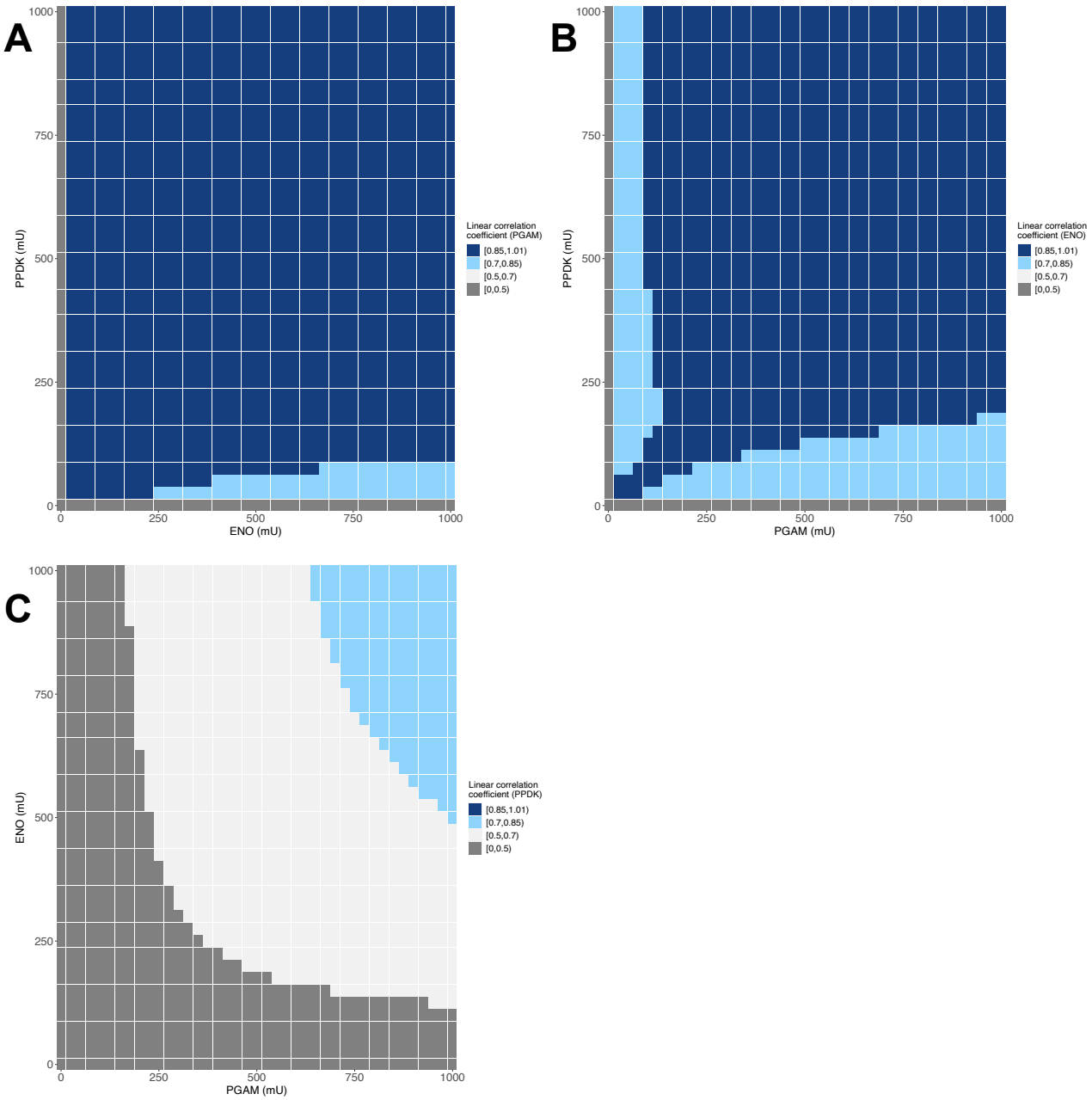

FIGURE 4.11 : *Evolution of linear correlation coefficient for each enzyme of Dataset 1 (68,950 data).* ***(A, B, C)*** *Variation of PGAM* ***(A)****, ENO* ***(B)*** *or PPDK* ***(C)*** *correlation coefficient between enzymes activities and the predicted final flux.*

Nonlinear Machine learning methods for Metabolic pathway modeling outperform Rborist

Based on the preceding data, we also investigate whether we can build a good predictive model by using linear and nonlinear ML methods. In a previous study, we used Artificial Neural Networks (ANN) to predict the flux (Lo-Thong *et al.*, 2020). Here, only one ANN model is developed and proves to be one of the best models obtained (table 4.5 and figure 4.12 E).

	Dataset 1				Dataset 2				Dataset 3			
	Training set		Test set		Training set		Test set		Training set		Test set	
Model	cvRMSE	cvR²	RMSE	R²	cvRMSE	cvR²	RMSE	R²	cvRMSE	cvR²	RMSE	R²
QRF (RF)	0.565	1	0.22	1	0.181	0.997	0.021	1	0.824	0.993	0.105	1
XGBoost Linear	0.486	1	0.403	1	0.156	0.997	0.023	1	1.351	0.982	1.094	0.988
Cubist	0.213	1	0.158	1	0.13	0.998	0.059	1	1.214	0.985	1.223	0.985
Rborist (RF)	0.69	1	0.416	1	0.195	0.996	0.078	0.999	1.03	0.989	0.617	0.996
ANN	2.787	0.997	2.7	0.998	0.133	0.998	0.098	0.999	1.924	0.962	1.9	0.964
SVM Radial	3.361	0.996	3.368	0.996	0.329	0.99	0.243	0.995	1.895	0.964	1.899	0.964
SVM Poly	9.484	0.971	9.351	0.971	0.473	0.979	0.409	0.985	2.102	0.955	2.11	0.955
bagEarth GCV (bagging MARS)	21.008	0.856	20.603	0.859	0.949	0.917	0.967	0.913	2.384	0.942	2.417	0.941
Bayesian GLM	30.643	0.693	30.365	0.694	1.44	0.805	1.379	0.823	3.522	0.874	3.582	0.87
Spike-and-slab	30.643	0.693	30.365	0.694	1.44	0.805	1.379	0.823	3.522	0.874	3.582	0.87
Ridge	30.643	0.693	30.365	0.694	1.439	0.804	1.381	0.823	3.522	0.874	3.582	0.87
Lasso	30.954	0.693	30.676	0.694	1.461	0.803	1.407	0.821	3.526	0.874	3.585	0.87
PLS	30.643	0.693	30.365	0.694	1.584	0.763	1.55	0.777	4.046	0.834	4.124	0.828

TABLE 4.5 : Summary table of statistical measurements for each predictive model.

RF: Random Forest. RMSE are in nmol·min^{-1}. Colors refer to: linear models (grey) and nonlinear models (blue). Models in bold are the top five models for all datasets. Dataset 1 corresponds to the lower part of *Entamoeba histolytica* glycolysis; Dataset 2 to the peroxide detoxification pathway of *Trypanosoma cruzi* and Dataset 3 to the industrial-scale penicillin fermentation process of *Penicillium chrysogenum*.

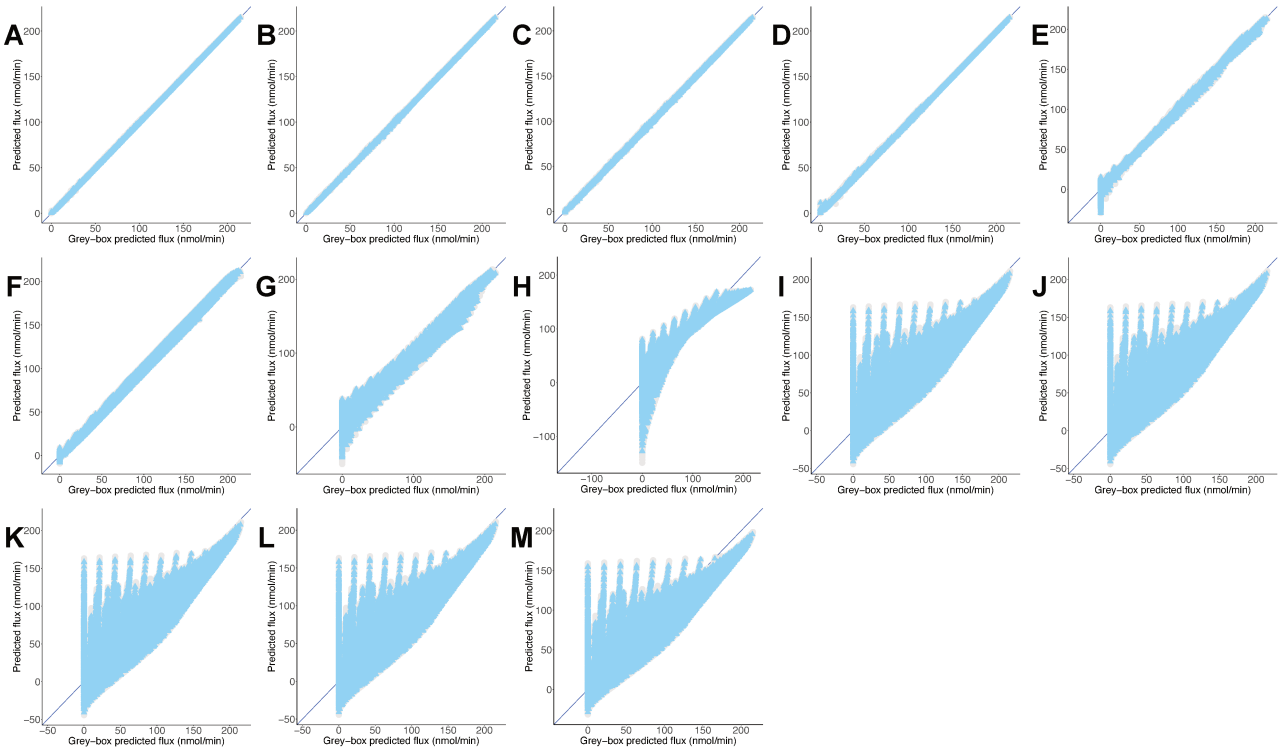

FIGURE 4.12 : *Predictions of a mix of experimental and grey-box predicted flux by different predictive models.*

(A, B, C, D) *Flux from Dataset 1 (Table A.5 in Appendix A) predicted by the Cubist* ***(A)****, QRF* ***(B)****, XGBoost Linear* ***(C)****, Rborist* ***(D)****, ANN* ***(E)****, SVM Radial* ***(F)****, SVM Poly* ***(G)****, bagEarth GCV* ***(H)****, PLS* ***(I)****, Bayesian GLM* ***(J)****, Ridge* ***(K)****, Spike-and-slab* ***(L)*** *and Lasso* ***(M)*** *models. Grey circles: training set and blue triangles: test set. See Table 4.5 for the statistical measurements of each model.*

Among the designed models and for the first dataset (Table A.5 in *Appendix A*), the random forest models stand out, with better flux prediction for the training set with the model built with Rborist package: cvRMSE=0.886 nmol·min^{-1} and cvR²=0.995, than the QRF model: cvRMSE=1.005 nmol·min^{-1} and cvR²=0.993 (table 4.6 and figure 4.13 B, D).

	Training set		Test set	
Model	**cvRMSE**	**cvR²**	**RMSE**	**R²**
XGBoost Linear	0.898	0.994	0.051	1
QRF (RF)	1.005	0.993	0.09	1
Cubist	0.481	0.998	0.28	1
Rborist (RF)	0.886	0.995	0.31	1
ANN	0.666	0.997	0.637	0.998
SVM Radial	1.267	0.99	0.999	0.995
SVM Poly	1.446	0.986	1.282	0.991

bagEarth GCV (bagging MARS)	3.44	0.921	3.233	0.941
Ridge	6.044	0.748	5.882	0.8
Bayesian GLM	6.045	0.748	5.882	0.8
Spike-and-slab	6.045	0.748	5.882	0.8
Lasso	6.108	0.746	5.969	0.802
PLS	8.604	0.497	9.074	0.521

TABLE 4.6 : Summary table of statistical measurements for each predictive model for Table A.4 in *Appendix A* (2,000 data).

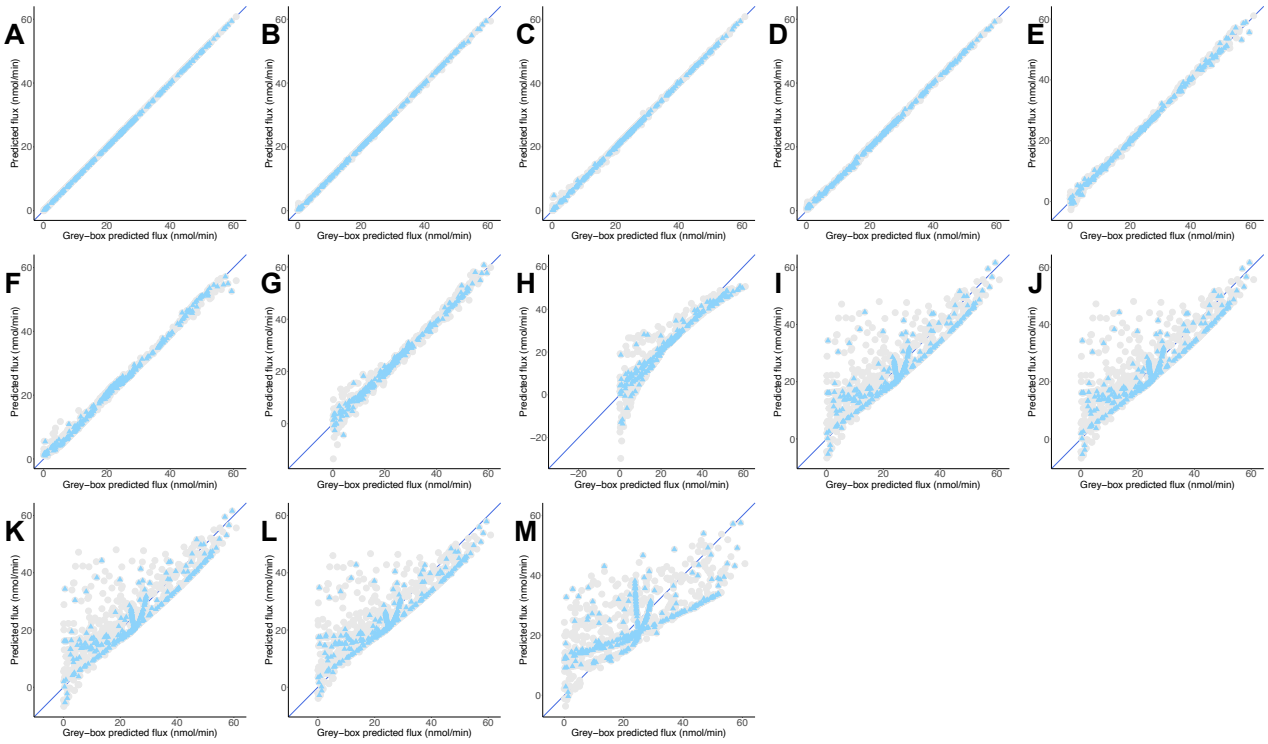

FIGURE 4.13 : *Predictions of mix of experimental and grey-box predicted flux by different predictive models.* ***(A-M)*** *Flux from Table A.4 (Appendix A) predicted by the XGBoost Linear* ***(A)****, QRF* ***(B)****, Cubist* ***(C)****, Rborist* ***(D)****, ANN* ***(E)****, SVM Radial* ***(F)****, SVM Poly* ***(G)****, bagEarth GCV* ***(H)****, Ridge* ***(I)****, Bayesian GLM* ***(J)****, Spike-and-slab* ***(K)****, Lasso* ***(L)*** *and PLS* ***(M)*** *models. Grey circles: training set, and blue triangles: test set. See Table 4.6 for the statistical measurements of each model.*

As for the test set, the QRF model outperforms the Rborist model, with RMSE=0.09 nmol·min^{-1} and R^2=1. Another good model, also nonlinear, is the XGBoost Linear method, with cvRMSE=0.898 nmol·min^{-1} and cvR^2=0.994 (table 4.6 and figure 4.13 A). Moreover, the results obtained with Bayesian GLM, Lasso, Ridge, Spike-and-slab, and the PLS model indicate that a linear model is not really adequate to describe this metabolic pathway. In fact, the PLS model gives the highest value

for cvRMSE and the lowest value for cvR² (table 4.6); also, we can see that the flux predictions are not very good (figure 4.13 M). For the second dataset (Table A.5 in *Appendix A*), we obtained almost the same results: first with the Cubist model (cvRMSE=0.213 nmol·min^{-1} and R²=1), then the two random forest models (table 4.5). This time, better results are obtained with the QRF model: cvRMSE=0.565 nmol·min^{-1} and R² = 1, than with the Rborist model: cvRMSE=0.69 nmol·min^{-1} and R²=1 for the training set (table 4.5 and figure 4.12 A-B and D). The XGBoost Linear method also gives good flux predictions, with cvRMSE=0.486 nmol·min^{-1} and cvR²=1 (table 4.5 and figure 4.12 C).

For the same reasons stated above, all linear models show poor results in predicting flux starting from enzyme activities, and are therefore not adequate to model the lower part of glycolysis here (figure 4.12 H-M). Overall and for Dataset 1, the Cubist model has the best generalization capability, with a lower RMSE=0.158 nmol·min^{-1} and a higher R²=1 for the test set (table 4.5). These results show that the nonlinear models, such as random forests, Cubist and XGBoost Linear, are able to indicate the final flux of the pathway by using the predicted data.

4.3.2. Example 2: The peroxide detoxification pathway of *Trypanosoma cruzi*

An *ad hoc* grey-box model allows data augmentation of enzyme activities and flux

We look at modeling the second metabolic pathway, which can also be used for drug design purposes. In the grey-box model developed here around this second dataset, the first and third enzymes employ a modified kinetic equation including two different adjustment terms: α=23 and *ß*=8 (table 4.1).

The determination of these parameters is detailed in the Methods section. We obtained a relatively good model of flux prediction (R²≈0.67 and RMSE=4.668 nmol·min^{-1}) when enzyme activities are varied (figure 4.14). However, the model still overestimates the flux when TryR activity is varied and when TXNPx activity is higher than 698.35 mU. The new dataset contains 1,671 enzyme balances evolving around the experimental ones (Dataset 2, see Table B.1 in *Appendix B*).

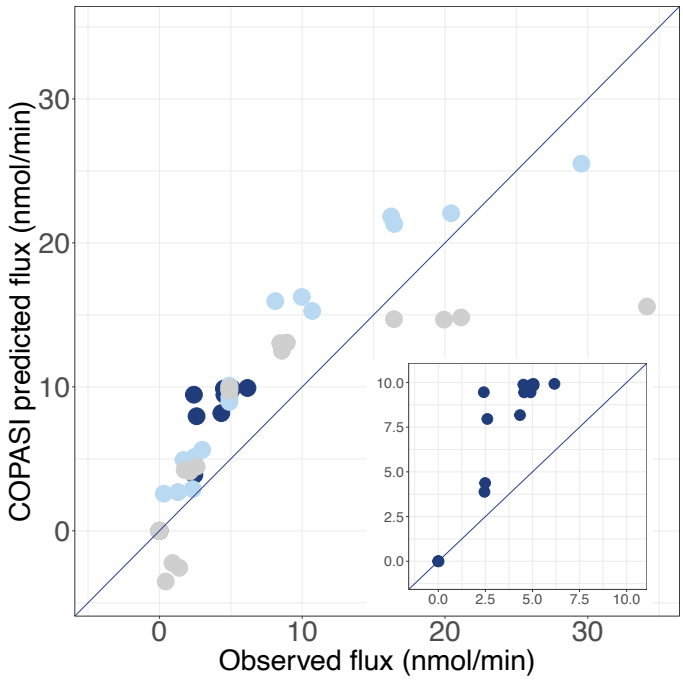

FIGURE 4.14 : *Flux predictions by the grey-box model.*
Circle colors refer to the various levels of enzyme activity: TryR (dark blue), TXN (light blue) or TXNPx (grey). Inset: flux predicted when TryR activity is varied.

The predicted final fluxes vary between 0 and 11.46 nmol·min^{-1}; the dataset's distribution is shown in figure 4.15 and in table 4.7. It is important to note that we could not go below 16.1 mU and 57.6 mU for TryR and TXNPx activity. The reason is that the grey-box model is not able to predict the flux below these values. Also, an analysis of the correlation between the different variable shows that TXN has the highest correlation coefficient, followed by TXNPx and lastly TryR (figure 4.16 A). Here, these linear correlation coefficients point out the predominantly linear character of this metabolic pathway, when each or all enzyme activities is varied. Nevertheless, the nonlinear aspect of the peroxide detoxification pathway is certainly not to be negligeable, since the coefficient average, when all enzyme activities are varied, is lower than 0.5. These results support those obtained by González-Chávez *et al.*, which demonstrate that TXN and TXNPx exert the greatest control on the pathway's flux, while TryR exerts very little control on the flux (González-Chávez *et al.*, 2015, 2019).

The augmented dataset is now used to test different ML approaches, as described in the following section.

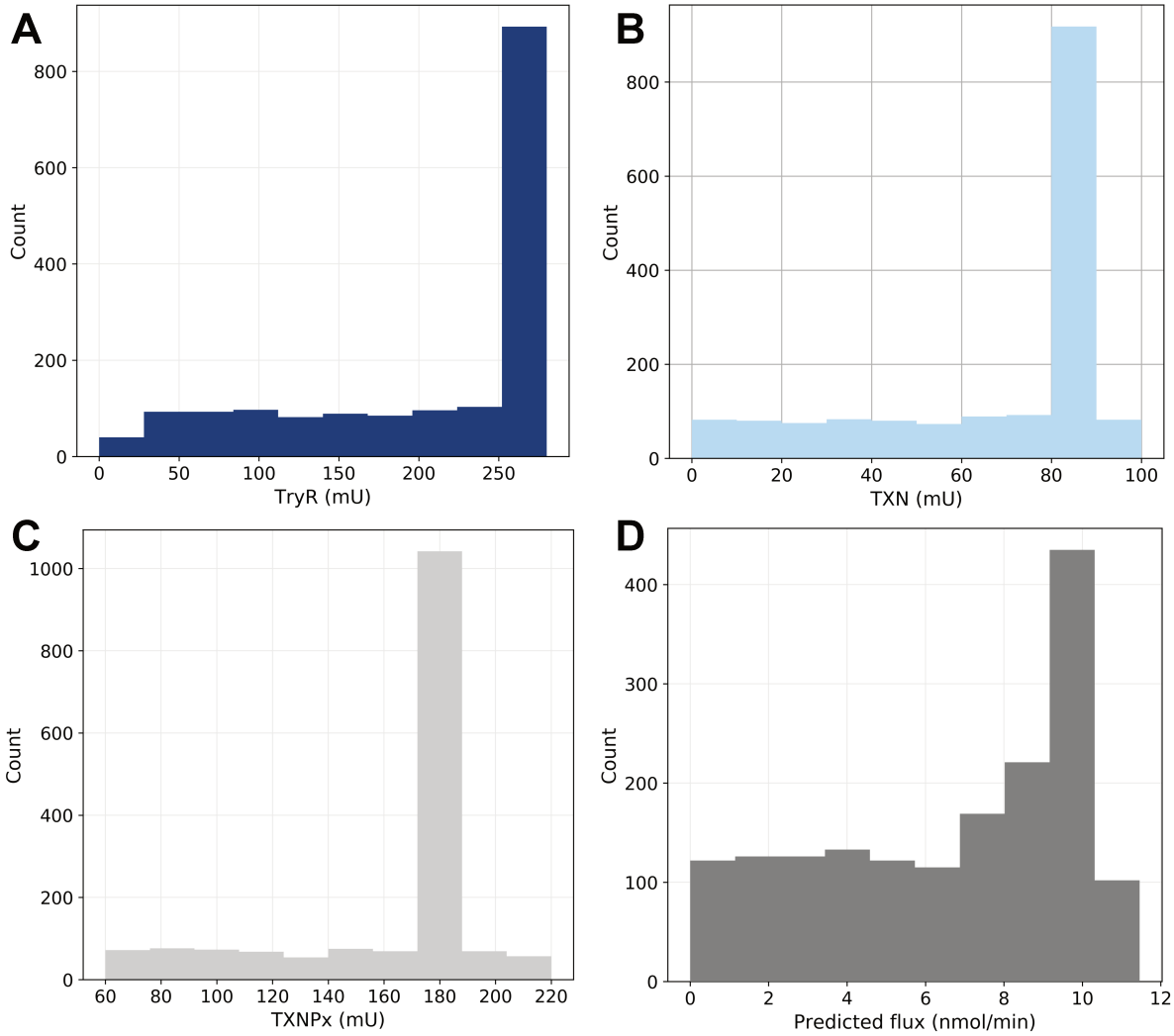

FIGURE 4.15 : *Histogram of the distribution of Dataset 2.*
(A, B, C) *Variation of TryR* ***(A)****, TXN* ***(B)*** *or TXNPx* ***(C)*** *activity.* ***(D)*** *Variation of the final flux predicted by the grey-box model.*

	TryR	TXN	TXNPx	J_{pred}
Data count	1,671	1,671	1,671	1,671
Mean	203.88	69.84	161.09	6.6
Standard deviation	79.8	27.7	37.15	3.25
Minimum value	16.1	0	57.6	0
25%	143.75	52.95	151.7	3.82
50%	264	88	179	7.58
75%	264	88	179	9.58
Maximum value	264	102	220	11.46

TABLE 4.7 : Description of the new generated dataset (Dataset 2).

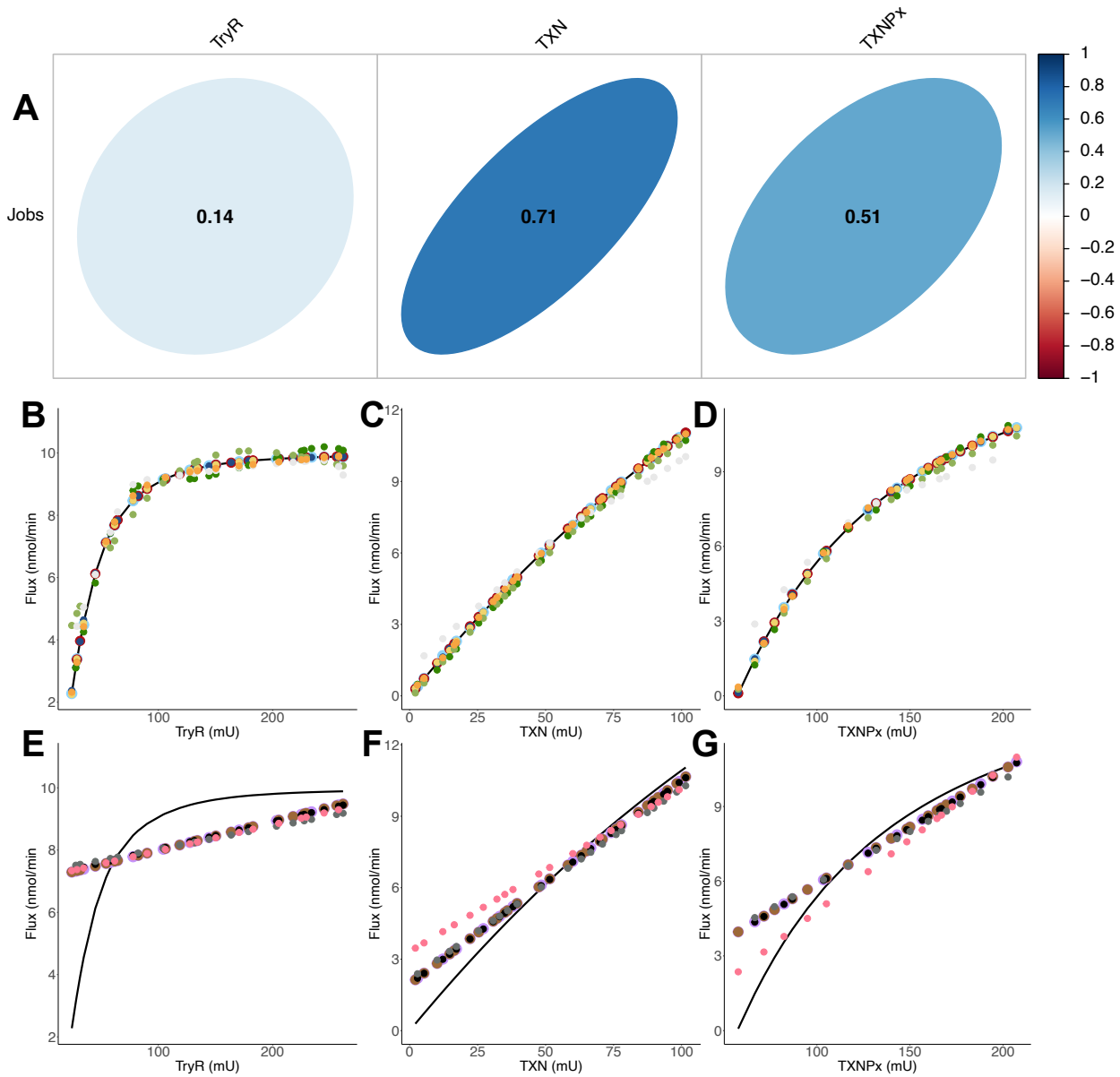

FIGURE 4.16 : *Predictions of final flux by different predictive models.*

(A) *Linear correlation between enzyme activities (inputs) and the flux (output) of Dataset 2 (Table B.1 in Appendix B) when all enzyme activities are varied. Correlation coefficients are also calculated when only one enzyme activity is varied: 0.76 (TryR), 0.998 (TXN) and 0.97 (TXNPx). A perfect circle means that there is no linear correlation between the variables, while a straight line means that there is a perfect linear correlation between the variables.* ***(B-G)*** *Flux variation as a function of the enzymatic activity of TryR* ***(B, E)****, TXN* ***(C, F)*** *and TXNPx* ***(D, G)****. Colored circles refer to predicted data from: QRF (dark blue), XGBoost Linear (light blue), Cubist (red), Rborist (yellow), ANN (orange), bagEarth GCV (light grey), SVM Poly (light green), SVM Radial (dark green), Bayesian GLM (purple), Spike-and-slab (brown), Ridge (black), Lasso (dark grey) and PLS (pink). A curve of the fitting experimental data is represented by the black curve. See table 4.5 for the statistical measurements of each model.*

Nonlinear Machine learning methods are efficient for flux prediction

We built different ML models and evaluated their performance. Of the thirteen models built, only five predict well the flux for both training and test sets: the random forest (QRF and Rborist), XGBoost Linear, Cubist and ANN (figure 4.16 B-D and figure 4.17 A-E). These models have a cvRMSE range of 0.181-0.109 nmol·min^{-1} and cvR^2 of 0.997-0.999 for the training set, and RMSE range of 0.021-0.098 nmol·min^{-1} and R^2 of 0.999-1 for the test set (table 4.5).

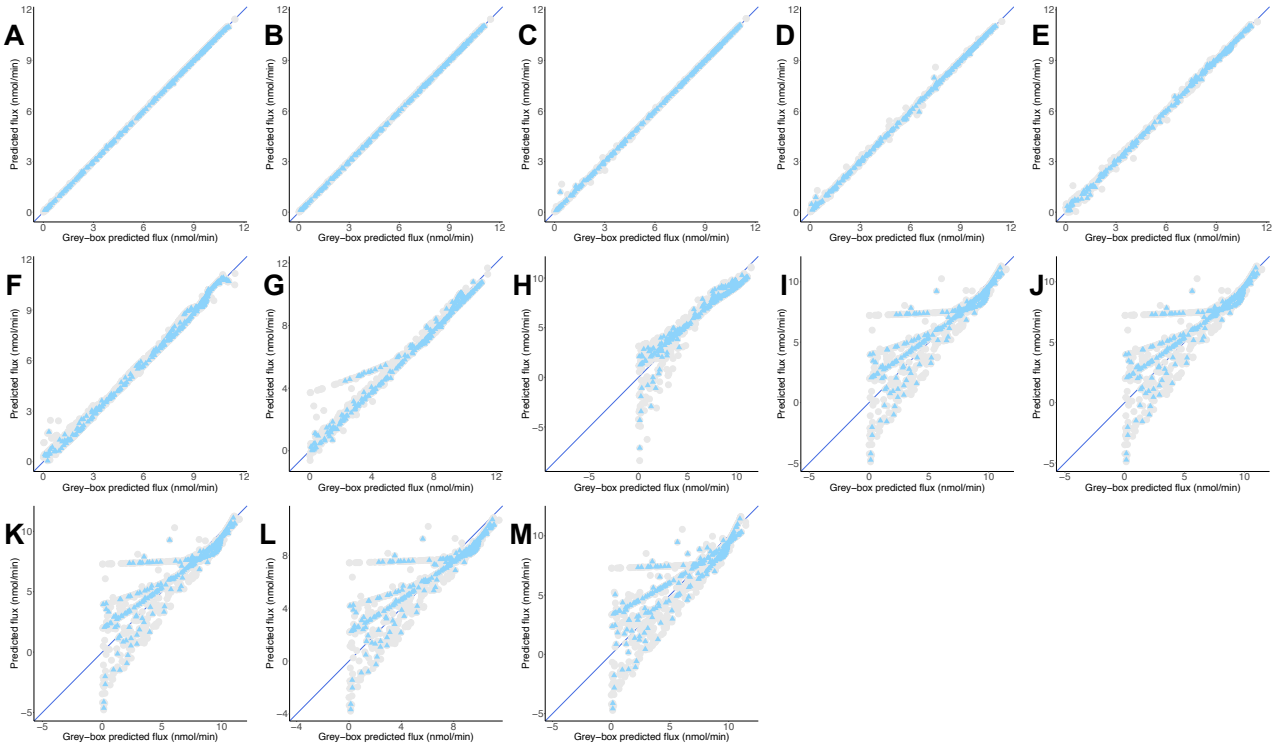

FIGURE 4.17 : *Predictions of grey-box predicted flux by different predictive models for Dataset 2.* ***(A-M)*** *Flux from the second dataset (Table B.1 in Appendix B) predicted by the QRF* ***(A)****, XGBoost Linear* ***(B)****, Cubist* ***(C)****, Rborist* ***(D)****, ANN* ***(E)****, SVM Radial* ***(F)****, SVM Poly* ***(G)****, bagEarth GCV* ***(H)****, Bayesian GLM* ***(I)****, Spike-and-slab* ***(J)****, Ridge* ***(K)****, Lasso* ***(L)*** *and PLS* ***(M)*** *models. Grey circles: training set, and blue triangles: test set. See table 4.5 for the statistical measurements of each model.*

The following three models (SVM Radial, SVM Poly and bagEarth GCV) predict moderately well the flux of peroxide detoxification (figure 4.16 E-G and figure 4.17 F-H), with cvRMSE between 0.329 and 0.949 nmol·min^{-1}, and cvR^2 between 0.917 and 0.99 (table 4.5). With the test set, their performance is slightly lower, with RMSE between 0.243 and 0.967 nmol·min^{-1} and R^2 between 0.913 and 0.995 (table 4.5).

In contrast, the last five models can hardly predict the flux from enzymatic activities for both training and test sets, particularly for flux below 7.5 nmol·min^{-1} which is within the physiological and experimentally determined value (figure 4.16 E-G and figure 4.17 I-M). These models present

higher RSME and lower R^2 values for the training set (cvRMSE range of 1.44-1.584 nmol·min^{-1} and cvR^2 range of 0.763-0.805), test set (RMSE between 1.379 and 1.55 nmol·min^{-1} and R^2 range of 0.777-0.823), confirming their poorer performance not only in terms of learning but also in terms of generalization, in making robust predictions on new data (table 4.5). We also observe that models Bayesian GLM, Spike-and-slab and Ridge give comparable results (table 4.5 and figure 4.17 I-K).

These results, together with those in example 1, allow us to confirm that nonlinear models are more appropriate to predict the flux of a metabolic pathway than linear ones.

4.3.3. Example 3: The industrial-scale penicillin fermentation process of *Penicillium chrysogenum*

In addition, another type of metabolic pathway we can examine are the production pathways; their modeling would allow the development of an optimized overall process. In fact, another study revealed that ML methods can accelerate the optimization of chemical synthesis (Hein, 2021). As stated before, we do not need to enlarge this dataset, which is composed of records of the various parameters of an industrial-scale penicillin fermentation process. It is important to consider that the inputs of our models are no longer the enzymatic activities, but different variables such as: batch time, oil flow, aeration rate, vessel volume and weight, carbon evolution rate and CO_2 percentage in off-gas. A slight variation of CO_2 in off-gas is recorded (Table C.1 in *Appendix C*); this can be explained by the implementation of a system, by the operators, allowing corrective measures to be taken when the CO_2 level is too high, thus avoiding the detrimental effect of an accumulation of CO_2 on the growth of *Penicillium chrysogenum* and the production of penicillin. As the percentage of CO_2 in off-gas is maintained at a certain level, it is not surprising that the carbon evolution rate does not vary much either and presents a low standard deviation (table 4.8). Also, the output we are interested in is not the pathway flux, but the final concentration of penicillin (figure 4.5). As regards the correlation coefficient between the variables, we note that it is generally high between the parameters and the final penicillin concentration (table 4.9); this correlation can be positive (e.g., time) or negative (e.g., oil flow). These correlation coefficients reveal the linear nature of the fermentation process studied in Dataset 3.

	Time (h)	Oil flow (L/h)	Aeration rate (L/h)	Vessel weight (kg)	Carbon evolution rate (g/h)	Vessel volume (L)	CO_2 in off-gas (%)	Penicillin concentra-tion (g/L)

Data count	113,935	113,935	113,935	113,935	113,935	113,935	113,935	113,935
Mean	114.75	26.35	65.25	81,076.73	1.25	73,312.87	1.44	14.33
Standard deviation	66.99	4.95	11.69	10,097.23	0.48	8,599.64	0.5	9.93
Minimum value	0.2	22	20	60,395	0.029	56,549	0.075	0
25%	57	23	60	73,018.5	0.98	65,885.5	1.23	5.53
50%	114	23	65	84,367	1.40	75,770	1.6	14.38
75%	171	30	75	88,608	1.62	79,892	1.76	22.69
Maximum value	290	35	75	107,010	2.05	95,716	7.12	36.18

TABLE 4.8 : Description of the new generated dataset (Dataset 3).

	Observed penicillin concentration
Time	0.92
Oil flow	-0.81
Aeration rate	0.78
Vessel weight	0.79
Carbon evolution rate	0.78
Vessel volume	0.76
CO_2 in off-gas	0.68

TABLE 4.9 : Correlation table between the parameters of the bioreactor and the observed penicillin concentration for Dataset 3.

Nonlinear Machine learning methods predict the fermentation process better than the linear methods

The results of penicillin concentration predictions reveal that Random forest models effectively predict experimental concentrations, with cvRMSE=0.824/1.03 $g \cdot L^{-1}$ and cvR^2=0.993/0.989 (QRF/Rborist) for the training set and RMSE=0.105/0.617 $g \cdot L^{-1}$ and R^2=1/0.996 (QRF/Rborist) for the test set (table 4.5 and figure 4.18 A-B). We can then separate the rest of the models into two groups, based on their performance on the test set. The first one, that predicts the penicillin concentration fairly well, has RMSE between 1.094 and 2.417 $g \cdot L^{-1}$, and R^2 between 0.941 and 0.988 (table 4.5 and

figure 4.18 C-H). By contrast, we found that the predictions of the second group are considerably worse, with many more outliers (figure 4.18 I-M), and with RSME higher than 3.5 $g \cdot L^{-1}$ and R^2 lower than 0.9 for the test set (table 4.5). As noted in the previous dataset, we also found many models that give the same results, namely: Bayesian GLM, Ridge, Spike-and-slab and Lasso (table 4.5 and figure 4.18 I-L). Here also, Lasso and PLS were the worst in terms of predictions. Interestingly, compared to the preceding results, Dataset 3 gives the best results for linear models (lowest RMSE and highest R^2 values for the training and test sets); this could be explained by the largely linear nature of the penicillin concentration used with respect to the parameters used. These results support the previous ones and confirm that nonlinear models surpass linear models for the prediction of penicillin concentration through the fermentation process.

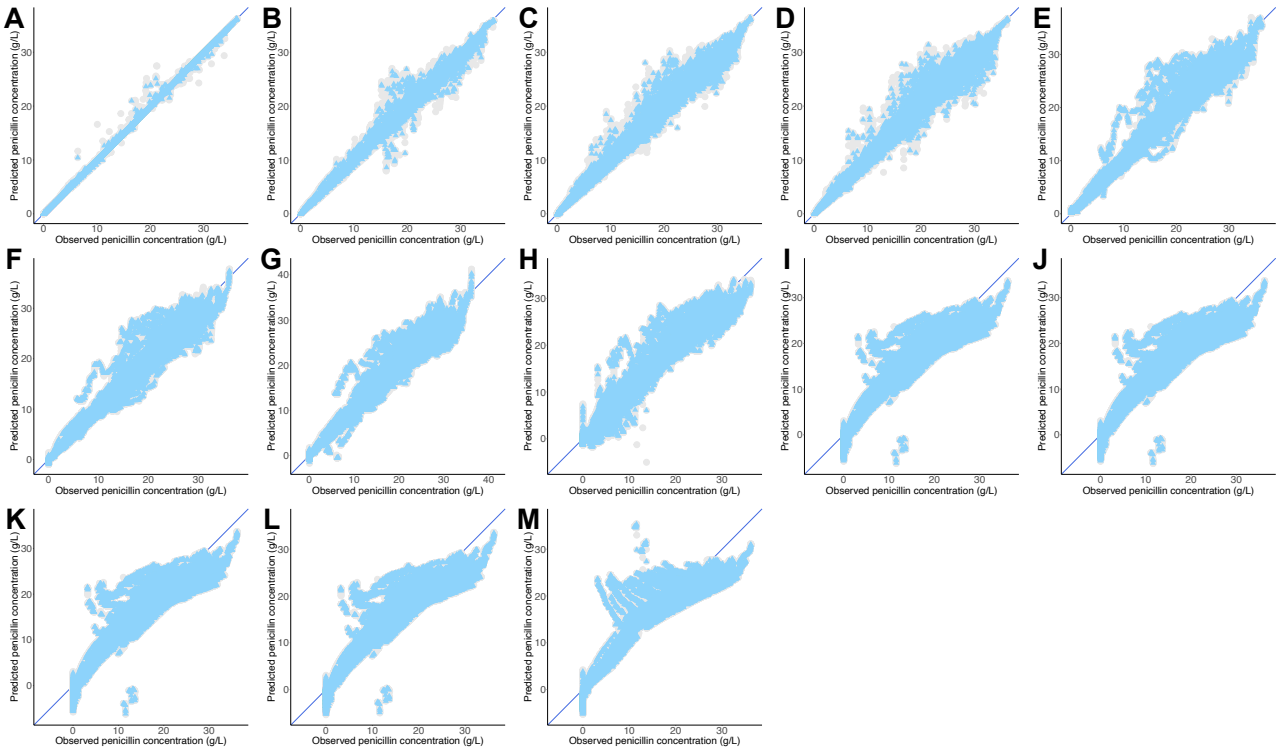

FIGURE 4.18 : *Predictions of observed penicillin concentration by different predictive models.*
(A-M) *Flux from the third dataset (Table C.1 in Appendix C) predicted by the QRF* ***(A)****, Rborist* ***(B)****, XGBoost Linear* ***(C)****, Cubist* ***(D)****, SVM Radial* ***(E)****, ANN* ***(F)****, SVM Poly* ***(G)****, bagEarth GCV* ***(H)****, Bayesian GLM* ***(I)****, Ridge* ***(J)****, Spike-and-slab* ***(K)****, Lasso* ***(L)*** *and PLS* ***(M)*** *models. Grey circles: training set, and blue triangles: test set. See table 4.5 for the statistical measurements of each model.*

Performance comparison of all models

After showing that nonlinear ML methods are more suitable for modeling metabolic pathways, we performed a comparison of the performance of all models. At first glance, the plots further confirm the preceding results and display higher RMSE values and lower R^2 values for the linear

models compared to nonlinear models (figure 4.19). In addition, regardless of the number and/or type of data, we observe that Spike-and-slab, Ridge, Lasso and Bayesian GLM models give almost the same results (figure 4.19 and table 4.5). Also, it appears that some nonlinear models work less well with large datasets; this is the case for ANN, bagEarth GCV, SVM Poly and SVM Radial (figure 4.19). Moreover, it appears that random forest models (QRF and Rborist) are the best suited for metabolic pathway modeling, as they give the best results in term of RMSE and R^2 whatever dataset was used. Furthermore, we can evaluate the impact of the degree of non-linearity of the pathway on the predictions. Indeed, the pathway that has a high nonlinear structure (Dataset 1) gives worse results for linear models than the pathway that presents a less nonlinear structure (Dataset 3), which also gives good results with nonlinear models (figure 4.19A and table 4.5). For example, Dataset 1 performs less well with the Ridge model, with RMSE=30.365 nmol·min^{-1} and R^2=0.694, than Dataset 3, which performs well with the same model, with RMSE=3.582 nmol·min^{-1} and R^2=0.87.

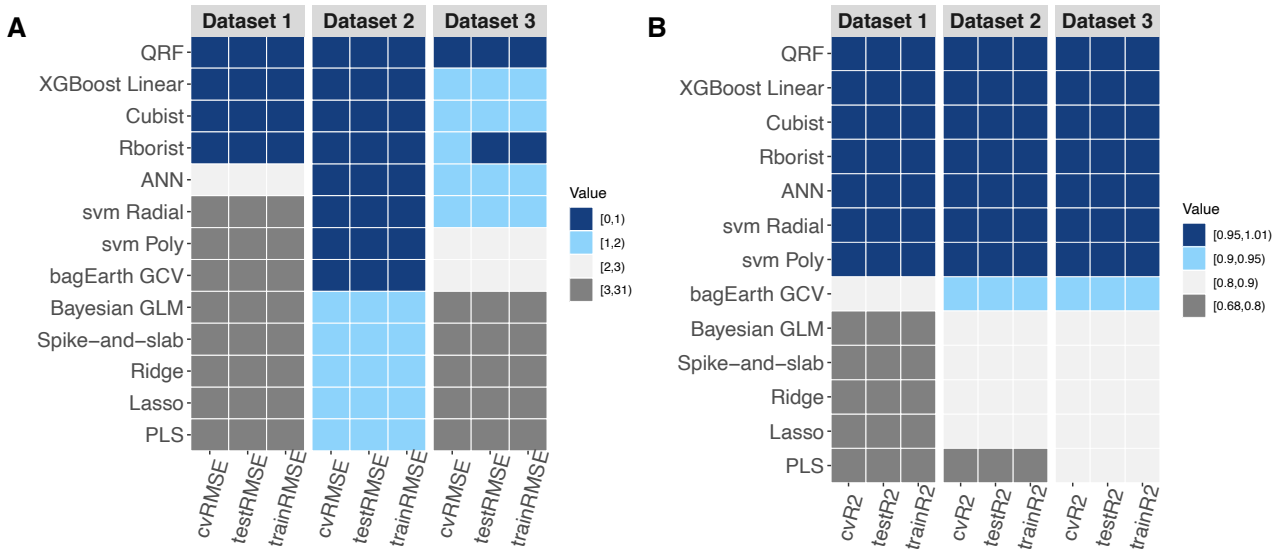

FIGURE 4.19 : *Comparison of the RMSE and R^2 of the three datasets.*
***(A-B)** Variation of RMSE **(A)** and R^2 **(B)** values for the different models and for each dataset.*

4.4. Discussion

4.4.1. Comparison and applicability of knowledge-based and data-driven approaches

The first objective of this study is to determine what sort of data-driven model could better simulate the biological pathways studied. By using different datasets, we build several models with the enzyme balances or parameters collected from a bioreactor and reveal that Random Forests (QRF and Rborist), Cubist and XGBoost Linear are three good methods to predict the final flux or concentration of a final product. This works is part of a larger study about the applicability of either a knowledge-based or a data-driven approach. Indeed, in other fields such as fault detection and diagnosis, a comparison of these two methods demonstrates that they both have comparable performance and can be used (Alzghoul *et al.*, 2014; Yang and Rizzoni, 2016). In biological system modeling, as is the case here, we demonstrated that in instances where little knowledge is available and difficult to obtain on a large scale basis (*e.g.* kinetic parameters k_{cat} and K_m of an enzyme, flux), or when complex feedback regulation mechanisms take place, a data-driven method can be a good alternative for modeling a metabolic pathway, as many authors have shown before (Ramachandran *et al.*, 2011; Hou *et al.*, 2016). By comparison, the knowledge-based method can be laborious and long, due to data mining from the literature or wet laboratory experiments, whereas there is an ease and speed of building models with the data-driven method (Kadarmideen, 2016). Moreover, the suitability of using either method relies on the quantity and quality of the knowledge or the data. Here, to illustrate this point, we simulate two datasets: the first one consisting of an exploration of the experimental data (2,000 data) and the second one composed of enzyme activities from 0 to 1,000 mU (68,950 data). The largest one gives better predictions for the three best models (Random forests, Cubist and XGBoost Linear) than the other dataset, and shows us the importance of having a large dataset before using machine learning methods. In fact, the size of the training set has been shown to be a major driving factor of prediction accuracy (Somarathna *et al.*, 2017). However, we used two datasets made up of a mix of experimental and predicted data to build the models, and even if predicted from a good quality model, they remain mostly predicted data and are not comparable to a fully experimental dataset, which is also difficult to obtain. Thus, it would be worth considering methods using only experimental data, when sufficient data are available to build the models. Interestingly, a data-driven approach is often used to discover biological pathways or unravel pathways that are not well understood. Thus, combined with the knowledge-based approach, this can quickly make clear

the complexity of biological systems modeling. Surprisingly, model performance was weaker for the largest dataset from the bioreactor records than for the smaller datasets. The reason for this result may lie in the choice of input variables. Several studies have highlighted the need for variable selection in order to have better predictions (Camacho *et al.*, 2018; Genuer *et al.*; Awan *et al.*, 2019). Indeed, variable selection allows the use of the most informative variables to predict the output variable(s) and reduce the time of computing. Unlike the knowledge-based model, a diversity of variables for data-based models does not always mean better performance. This is one of the limitations of our study, since only one combination of input variables was tested during the work. It would be interesting for a future study to compare, for the same dataset, models using different sets of input variables, and to analyze their impact on model effectiveness.

4.4.2. Interpretability of machine-learning approaches

Another major issue facing users of machine learning approaches is the interpretability of these models. Even if, at this time, we do not have a common general definition of this term, many researchers, such as Schmidt *et al.* (Schmidt *et al.*, 2019), define a model's interpretability based on two aspects: a) intrinsic interpretability (or transparency): the ability to understand the inner mechanism of the model in the context of the study (*e.g.*, identification of variables most involved in the predictions), and b) *post hoc* interpretability: the ability to extract new information from the model or provide new insights into the relationships discovered during the process (*e.g.*, effect of a variable on another one) (Pintelas *et al.*, 2020; Murdoch *et al.*, 2019; Schmidt *et al.*, 2019). Although some ML methods, such as decision trees or linear regression models, are easily interpretable; this is not the case for most of the models developed here (*e.g.*, XGBoost Linear, bagging MARS, ANN). Nevertheless, using the variables that are directly related to the variable to be predicted, as we do here, allows us to gain some understanding of how the model works and the types of relationships that are revealed, enabling us to rely on the models. Furthermore, while we identified Random forests as one of the best methods for predicting final flux or product concentration, Pintelas *et al.* classifies it as a model that is hard to interpret (Pintelas *et al.*, 2020). Therefore, it would be interesting to compute variable importance or to apply different techniques to explain the model in order to increase its interpretability (Zhou *et al.*, 2019; Azodi *et al.*, 2020). Besides, knowing that models based on decision trees are among the simplest to interpret, we support the idea of Schmidt *et al.* that RF models are more accessible than others from an interpretability point of view (Schmidt *et al.*, 2019). An alternate solution would be to develop simpler models, but this would certainly reduce their overall performance.

Moreover, one of the key factors in the interpretability of the models is linked to the equations used. In fact, compared to knowledge-based models that use well-defined equations with a biological significance, ML models are governed by other equations, which sometimes are "outside our understanding" as Schmidt *et al.* observed in their study of the applications of ML in solid-state materials science (Schmidt *et al.*, 2019). This raises a real problem of confidence in the prediction results obtained with such methods. As these authors point out, the fact that these models were not based on physical principles in their study, or on biological principles in ours, could result in wrong predictions in completely unexpected cases, while providing great results overall. And in the present case where the models are used in the context of biomarker identification or optimization of an industrial bioreactor, we cannot risk obtaining such results from our models in these specific situations. Far from hindering us in the use of ML models, awareness of these problems allows us to formulate several recommendations for future research. These include the combination of interpretable models, *e.g.*, knowledge-based kinetic models with ML models, *e.g.*, random forests models; the prediction of a new set of experimental data with unexpected values. In this latter instance, this would require experimentally testing a range of "extreme" data that would be found in the parasites studied, or recording the bioreactor data even during failures of the penicillin production.

4.4.3. Strengths and weaknesses of the modeling methods

After analyzing the interpretability of the different modeling methods, it is worthwhile to note some advantages and disadvantages of their use in flux and concentration prediction. One of the best methods in our case is the random forest (QRF and Rborist). Many studies report the use of random forest in the biological field for the prediction of: protein interaction (Qi *et al.*, 2006), drug response based on protein markers (Ma *et al.*, 2006) and *in-vitro* drug sensitivity (Riddick *et al.*, 2011). Also, Riddick *et al.* used SVM and random forest to predict the flux of N_2O emissions, and found that random forest achieves the best performances among the built models (Villa-Vialaneix *et al.*, 2010). They highlighted that these models offered the advantage of having a low computational cost, compared to the SVM method. However, in our case, we notice that random forest is the least accurate predictability model compared to SVM methods, with the highest computation time for almost all datasets. Moreover, among the random forest packages developed on R, Rborist is quite a recent implementation, designed for multicore hardware, which minimizes data movement within memory to increase the performance and decrease the processing time (Wright and Ziegler, 2017). Surprisingly, here, Rborist package is the one that has the longest time

of computation and is more efficient on big datasets compared to other methods. It would be of interest to create variant models combining the random forest method and other methods, as in previous studies (Chen *et al.*, 2018; Zampieri *et al.*, 2019). An existing variant of random forests is the quantile regression forest (QRF) method, which has the capability of establishing prediction intervals that cover uncertainties, useful in the prediction of possible new data (Meinshausen, 2006). Francke *et al.* demonstrated in their work that this method had the advantage of calculating uncertainties associated with the predicted sediment yields, through the calculation of confidence intervals (Francke *et al.*, 2008). But they also stated that the model predictions will always be within the range of observations, which prevents implausible values but inhibits prediction outside the range of values learned from the training set. We saw here that, overall, QRF models have a good generalization capability; additional prediction of new experimental data, with data separated by a larger stepsize (>25), would be beneficial to confirm or invalidate this capability. This could be useful for the study of metabolic pathways in extremotolerant organisms.

This leads us to note one of the advantages not only of the QRF method but also of other ensemble learning methods, such as XGBoost Linear: prediction from high-dimensional data. Indeed, these models are among the best we have, with any starting dataset we have, from the simplest to the most complex with several types of variables. Remarkably, compared to other models, XGBoost Linear is better ranked for small datasets. This is confirmed by the work of Yang *et al.*, which propose that ensemble methods have the advantage of reducing the potential for overfitting in small sample size problem (Yang *et al.*, 2010). Another strength of XGBoost Linear compared to its peers is the combination of high accuracy and short time of processing. However, despite the great accuracy of these models, they are often more complex and less interpretable, and present a higher computational intensity.

Moreover, Cubist, a model based on modified regression tree theory, has the advantage of analyzing big data with high speed (Xu *et al.*, 2018). This was confirmed by our results, which show that Cubist is one of our best models (*e.g.*, for Dataset 1, Cubist: 2.49 min and QRF: 1,76 hr). However, we noted that the performance was better for the small datasets than for the bigger one. Another advantage that Zhou *et al.* noticed is the fact that the Cubist model is easy to interpret and is a suitable method for beginners (Zhou *et al.*, 2019).

The PLS method turned out not to be appropriate here to model these pathways and predict the final flux starting from enzyme activities, or the final product concentration starting from parameters of a bioreactor. This may be due to the inherent limitation of the PLS method to capture the non-linearities of the metabolic pathways. However, it performs better when we have a smaller dataset, as it has also been noted in a previous study on gluconeogenic flux prediction

(Antoniewicz *et al.*, 2006). But these results contradict those obtained with the PLS model for the prediction of limonene and isopentenol synthesis. In fact, in this work, results showed that the model performed well when the dataset was larger (lower RMSE, better predictions) (Costello and Martin, 2018). Also, one big advantage of the PLS technique remains that it has the shortest calculation time for modeling.

Our findings generally support the idea that nonlinear models are more suitable than linear ones for modeling metabolic pathways.

4.4.4. Decision-making support for pathway modeling

Given the many different methods that exist and continue to emerge, one can struggle with the choice of a model to build from a dataset. Faced with this decision, we can choose to build simple models or to use models being used in the same field of study and giving good results (Camacho *et al.*, 2018; Cifuentes *et al.*, 2020). In view of this, it would be useful to review and define some basic rules for building a decision-making support for future studies on modeling metabolic pathways. The first feature to consider is the quality of the biological dataset (figure 4.20). Do we have an initial dataset of good quality? Data quality can highly impact the model predictions. If the model is not of good quality, it would be better to build a new dataset and generate good quality experimental data. When the dataset is of good quality but small in size, it is useful to do data augmentation, as we did in this study; if this is not possible, we can use an ensemble model to build the metabolic pathway, since such models can deal with small datasets. Another useful criterion we can investigate is the number of variables. If the dataset presents a high number of variables, we can consider doing variable selection before building the model, or we have the option of building the model by using ensemble modeling that gives good accuracy with several input variables. Also, one key factor is the non-linearity of the studied metabolic pathway; do we have a nonlinear or a linear process? If our pathway is linear, we can design a battery of linear models which will give a high performance. But if our study involves a pathway that is nonlinear, then it is preferable to use a nonlinear model. After building our model, an evaluation of its accuracy is necessary to validate it. In case the performance of the model is not suitable, we can plan to refine it, for example by tuning the hyperparameters (Chicco, 2017), or simply to replace it and build a new one.

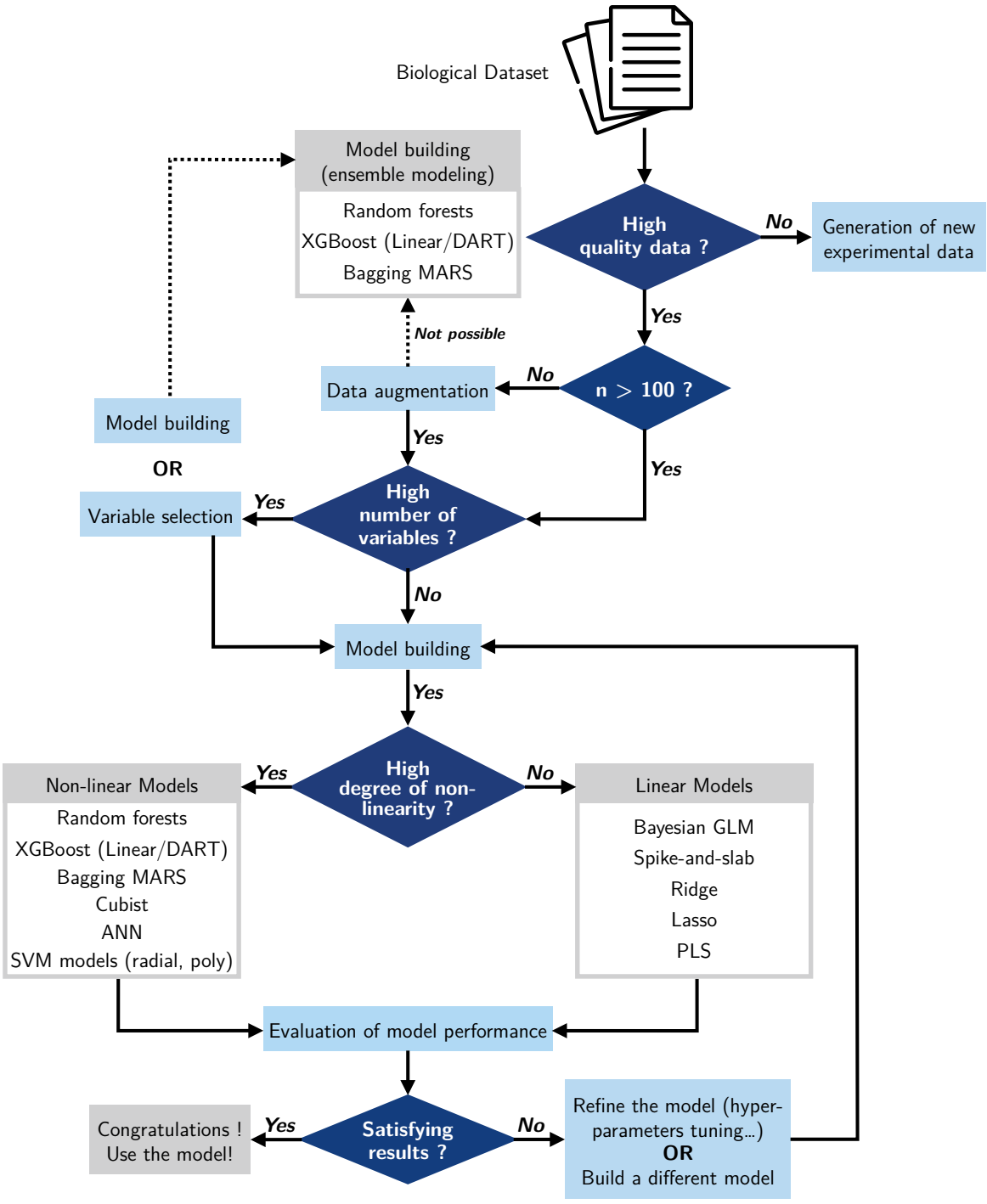

FIGURE 4.20 : *Decision-making support for the construction of metabolic pathway models using machine learning methods.*

4.5. Conclusion

Modeling has become a valuable tool for researchers, both for identifying therapeutic targets in the field of therapeutic research and for optimizing production in the industrial sphere. Among the different techniques of modeling are the machine learning methods, which are either linear or nonlinear. These methods, no matter how robust and effective, have scarcely ever been used to predict the flux of a metabolic pathway or the concentration of the final product.

Therefore, in this study, we aimed to build several linear and nonlinear ML models in order to define which ones are the most appropriate at modeling metabolic pathways and predicting final flux or final product concentration. For this purpose, we decided to model three different metabolic pathways: the lower part of glycolysis of *Entamoeba histolytica*, the peroxide detoxification pathway of *Trypanosoma cruzi* and the industrial-scale penicillin fermentation process of *Penicillium chrysogenum*. We then identified three powerful techniques for modeling these pathways: QRF, Rborist and XGBoost Linear model. The first two models are random forest models, obtained with QRF and Rborist algorithms, and an ensemble model, obtained with XGBoost Linear algorithm. The main characteristic of these models is that they are all nonlinear models. If we compare the other models, we notice that the nonlinear models perform better that the linear ones, and this, whatever the size of the dataset used. We explained this by the non-linear structure of the metabolic pathways studied. It appears that the degree of non-linearity of the pathway would drive the selection of the ML model. Hence, the more a production pathway has a nonlinear structure, the more a nonlinear method will be suitable for its modeling.

In addition, this work allowed the implementation of a first decision support tool for pathway modeling, based on different features of the initial data and more generally of the studied metabolic pathway.

Finally, nonlinear machine learning methods enable us to model metabolic pathways by identifying key-molecules, which are important for the drug-design process, improving disease diagnosis (cancer, viral/parasitic/bacterial infections, neurodegenerative diseases) by highlighting the differences between healthy and pathological situations, or even optimizing industrial production processes.

4.6. Discussion et conclusion du chapitre

Suite aux résultats prometteurs de prédiction de flux obtenus par les réseaux de neurones artificiels concernant la voie basse de la glycolyse chez *E. histolytica*, nous nous sommes consacrés dans ce chapitre à la modélisation de trois voies métaboliques différentes par deux catégories de modèles d'apprentissage automatique : les modèles linéaires (*PLS, Spike-and-Slab, Ridge regression*) et les modèles non-linéaires (*XGBoost, Random forest*).

Parmi les voies étudiées, nous retrouvons :

- La partie basse de la glycolyse chez *E. histolytica* : qui peut être considérée comme un exemple d'application d'une voie de production chez un parasite ;
- La voie de détoxification du peroxyde chez *Trypanosoma cruzi* : qui, comme nous l'avons expliqué au début de cette étude, pourrait être employée dans des systèmes de production afin d'éliminer les sous-produits synthétisés lors de la production et qui seraient potentiellement délétères pour le microorganisme ou pour les enzymes participant à la production ;
- La voie de la synthèse de pénicilline à partir d'un processus de fermentation chez *Penicillium chrysogenum* : qui est notre exemple d'application pour la prédiction, non pas du flux, mais de la concentration en produit final dans un bioréacteur.

Pour les deux premiers exemples, nous avons procédé à ce que nous appelons de l'augmentation de données (« *Data augmentation* »), en faisant appel à des modèles boîte-grise pour générer de nouvelles données et enrichir l'ensemble de données de départ. Ces travaux mettent en valeur l'utilité des modèles hybrides de ce type, présentés en détail deux chapitres plus tôt. Si le premier modèle génère sans difficulté des données sur n'importe quelle gamme d'activité enzymatique, remarquons que ce n'est pas le cas du deuxième modèle. Les modestes connaissances que nous avions au sujet de cette voie nous ont permis de bâtir un modèle qui fonctionne sur une large gamme de données, mais qui fonctionne moins bien dans des gammes de faibles activités pour les enzymes TryR et TXNPx. Loin de nous contrarier quant à l'utilisation générale de ce modèle boîte-grise, cela nous pousse à examiner la construction de notre modèle. En effet, l'une des limites de notre modèle réside dans la représentation de l'oxydation/réduction de la tryparédoxine (TXN). Cette protéine a été décrite en tant qu'enzyme dans notre modèle, alors qu'il s'agit d'une protéine de transport d'électron. Or une étude a montré que, lors de la modélisation d'un tel système, il était plus convenable de décrire une telle protéine en tant que couple oxydo-réducteur plutôt qu'en tant qu'enzyme michaélienne (Pillay *et al.*, 2009). Cette étude

précise également que les paramètres Michaelis-Menten, que nous utilisons par ailleurs ici, n'étaient pas adaptés pour décrire son activité. Une autre étude sur le système thiorédoxine chez *E. coli* nous montre la construction d'un modèle cinétique réaliste de ce système, qui ne fait pas intervenir la thiorédoxine en tant qu'enzyme mais en tant que substrat seul de l'enzyme thiorédoxine-réductase (Pillay *et al.*, 2011). Une amélioration du modèle hybride pourrait alors être envisagée, incluant une modification de la représentation de l'activité de la tryparédoxine.

Concernant la dernière voie étudiée, nous optons pour des données initiales différentes, intégrant cette fois des paramètres du bioréacteur dans lequel se déroule le processus de fermentation. L'ensemble de données utilisées comporte 37 variables différentes pour décrire la production de pénicilline, ce qui est nettement supérieur au nombre de variables retrouvé dans les autres ensembles de données. Suite à la sélection des variables les plus importantes pour décrire la concentration finale de pénicilline produite et la modélisation par les différents algorithmes, nous avons constaté une moins bonne modélisation de cette voie par rapport aux deux autres. Toutefois, de bons résultats sont généralement obtenus par les méthodes d'apprentissage automatique non-linéaires. Comme le nombre de descripteurs de la voie (variables en entrée) et le nombre de données sont plus nombreux pour la production de pénicilline que pour les deux premières voies métaboliques, on s'attendrait plutôt à obtenir de meilleurs résultats de prédiction pour cet exemple. Aussi, nous en déduisons que la sélection des variables descriptives de notre modèle aurait un impact important sur les résultats de la modélisation. D'autres études ont montré que la sélection des variables avait des effets sur la performance de modèle d'apprentissage automatique (Oreski *et al.*, 2017; Al Imran *et al.*, 2018). Dans l'un de ces travaux, il a été démontré la supériorité des modèles incluant une bonne sélection de variables sur les modèles qui n'opéraient pas cette sélection (Al Imran *et al.*, 2018). Il serait alors judicieux de notre part de considérer d'autres combinaisons de variables pour l'obtention de meilleurs modèles plus « représentatifs » du processus industriel de production de la pénicilline.

Par ailleurs, la différence de prédiction constatée entre les deux premières voies et la troisième pourrait être expliquée par la nature même des données utilisées. Comme il a été précisé plus tôt, le dernier ensemble de données contient plus des données sur les paramètres du bioréacteur (vitesse d'aération, vitesse d'agitation, température, pH…) que sur la voie en elle-même (concentrations en substrats, cosubstrats). Mais supprimer de telles variables de notre système pourrait avoir des effets délétères sur la modélisation, car ces paramètres participent à la régulation de la production de notre molécule d'intérêt. Nous pouvons alors faire l'hypothèse qu'un modèle « idéal » serait un modèle comprenant à la fois des données spécifiques à la voie de production et des données sur des paramètres influençant cette production.

De ces travaux de modélisation employant des modèles d'apprentissage automatique, nous avons construit un outil d'aide à la décision quant au choix de la méthode à employer et nous y avons inclus les principaux critères de sélection pour la modélisation de voie métabolique :

- La quantité et la qualité des données expérimentales ;
- La nature des variables d'entrée de notre système ;
- Le degré de non-linéarité de la voie étudiée.

Aussi, ces travaux constituent la première étude de comparaison de la performance de plusieurs algorithmes d'apprentissage automatique pour la modélisation de voies métaboliques. Ils attestent également la reproductibilité de ces méthodes pour les voies métaboliques en général, en développant plusieurs exemples d'application de méthodes non-linéaires. Une comparaison de ces modèles intelligemment conçus avec les modèles obtenus dans le chapitre précédent (**chapitre 3**) serait intéressante à effectuer. Toutefois, cette étude comparative ne pourrait se faire dans l'immédiat puisqu'il nous faudrait un ensemble de données expérimentales considérable.

Pour conclure ce chapitre, notre étude élucide notre interrogation sur l'utilisation de modèles d'apprentissage automatique pour la représentation d'une voie métabolique à partir de données initiales diverses. Nous sommes convaincus que parachever ces divers modèles, par l'optimisation de ces paramètres ou l'ajout de données sur la voie complète de production, pourrait mener à de meilleures prédictions des sorties d'intérêt (flux et concentration). Alors qu'il nous reste un pan de l'étude à développer avec la mise en place d'un système de rétrocontrôle au sein de nos modèles, nous prenons l'initiative de poursuivre notre étude avec le modèle hybride boîte-grise pour l'implémentation du système de contrôle. En effet, ce choix s'explique par le fait que les régulations opérées par notre système de contrôle se font sur les variables mêmes du système de production, or cela n'est pas envisageable pour des modèles d'apprentissage automatique puisqu'ils ne contiennent qu'une partie des variables définissant la voie de production, à savoir les activités des enzymes. Pour rendre possible cette implémentation sur de tels modèles, des expériences supplémentaires sont requises et devraient inclure le suivi de certaines concentrations ou de certains flux de métabolites.

Dans ce chapitre nous avons jaugé la capacité des algorithmes d'apprentissage automatique à modéliser des voies métaboliques pour leur application dans un futur proche. La comparaison des performances de ces modèles ont permis l'identification des meilleurs méthodes pour prédire le flux ou la concentration finale : les modèles de forêts aléatoires (RF) et les modèles d'apprentissage d'ensemble (« *Ensemble Modeling* ») avec la méthode de XGBoost. Nous avons mis également en lumière l'impact du degré de non-linéarité des voies étudiées sur la sélection des algorithmes,

expliquant la meilleure performance des modèles non-linéaires pour nos trois exemples d'application. Malheureusement, la faible quantité des données expérimentales couplée au modeste nombre de variables décrivant la voie font d'eux des modèles non-exploitables pour la suite de notre étude. Au chapitre suivant, nous nous appliquons à mettre au point un système de contrôle pouvant réguler automatiquement certaines variables de la voie basse de la glycolyse de *E. histolytica*, afin de maintenir un flux optimal.

Chapitre 5

Implémentation d'un système de rétrocontrôle sur la modélisation de voies métaboliques

Lors des travaux réalisés dans les chapitres antérieurs, plusieurs modèles du segment de la partie basse de la glycolyse d'*E. histolytica* ont été construits. L'un de ces modèles a montré une bonne performance en matière de prédiction du flux final de la voie modélisée : le modèle hybride boîte-grise. Celui-ci sera utilisé dans les paragraphes qui suivent avec pour objectif l'établissement d'un système de rétrocontrôle participant dans la régulation du flux en sortie de la voie.

Les systèmes de rétrocontrôle sont essentiels autant dans la régulation naturelle des voies métaboliques présentes dans un microorganisme, que dans la régulation d'une voie métabolique servant à la production d'une molécule d'intérêt. Plusieurs de ces procédés de contrôle ont été développés dans le **chapitre 1** « **Introduction** », et nous en avons mentionné un qui semble intéressant à modéliser : la **régulation Proportionnelle-Intégrale-Dérivée** (ou régulateur PID). Ce régulateur PID suscite particulièrement notre intérêt puisqu'il s'agit d'un système couramment utilisé en industrie pour établir un équilibre de production (Persad *et al.*, 2013). De par sa facilité de mise en place sur un système « réel » et sa robustesse, il constitue une méthode de choix pour la régulation de nos voies métaboliques pour la production de molécule. Plusieurs stratégies de contrôle reposant sur un régulateur PID ont été développées pour contrôler des bioréacteurs au niveau des paramètres physico-chimiques (pH, température) ou encore au niveau du taux de croissance. Néanmoins, ce type de régulation n'a pas encore été développé d'une part sur une voie métabolique seule ; les régulateurs PID étant habituellement dédiés au contrôle des actionneurs de bioréacteurs (Pachauri *et al.*, 2017; Lee *et al.*, 1991, 199; Imtiaz *et al.*, 2014). D'autre part, il n'a pas

encore été fait de contrôle sur le flux en sortie de la voie métabolique par la modulation des acteurs de la voie (concentrations en substrats, cosubstrats, enzymes).

Ce nouveau chapitre s'ouvre sur l'ajout d'un système de contrôle de type PID sur un modèle hybride de la voie basse de la glycolyse développé précédemment. Cette implémentation se décline en plusieurs étapes où le régulateur sera adapté au modèle boîte-grise pour assurer un flux stable en sortie de la voie métabolique. Cette régulation se fera par le biais de deux variables de commande : la concentration en AMP et/ou en PP_i.

La construction de ce régulateur PID est faite sur le logiciel COPASI, avec lequel nous avions également bâti les modèles cinétiques dont notre modèle boîte-grise. Ce système de rétrocontrôle est construit de sorte à réguler automatiquement certaines entrées de notre voie de production, afin de permettre le maintien du flux final à une valeur donnée. La première phase, avant la mise en place de ce rétrocontrôle, réside dans l'addition de perturbations dans notre modèle. Il est évident qu'un système à l'état d'équilibre n'a pas besoin d'être régulé, puisqu'il atteint déjà une valeur d'équilibre ; mais ce n'est pas le cas d'une voie métabolique intégrant des perturbations comme : des valeurs de concentrations en substrats ou cosubstrats qui ne sont plus constantes au cours du processus de production, ou encore une variation du pH ou de la température lors de la synthèse de la molécule. Une fois les perturbations ajoutées à la voie métabolique, le contrôle est alors implémenté sous forme d'équations venant modifier les valeurs de certaines entrées de la voie de production.

Lorsque l'ajout du système de contrôle au modèle est terminé, une phase de réglage du régulateur est amorcée. Comme il a été expliqué au **Chapitre 1**, ce régulateur est constitué de trois termes principaux : Proportionnel, Intégral et Dérivé, qui sont caractérisés par des paramètres appelés « gains », de manière respective, K_P, K_I et K_D. Ce sont ces gains qui font l'objet du réglage mentionné plus tôt. Cette étape d'ajustement du contrôle à la voie métabolique s'effectue par le biais du logiciel de départ COPASI, ainsi que du logiciel RStudio. La comparaison des divers ajustements effectués sur le contrôle permet de mettre en valeur le meilleur système de rétrocontrôle pour la voie étudiée.

L'avantage d'utiliser cette méthode de rétrocontrôle sur une voie métabolique, destinée à la production de molécule, est évalué au travers des différents résultats présentés dans ce chapitre. Aussi, ces quelques éléments développés dans cette section de nos travaux nous permettront d'analyser l'impact du choix des variables de commande sur le procédé de rétrocontrôle développé au sein du modèle. Ces modèles hybrides boîte-grise incluant un régulateur PID fonctionnel sont les premiers à être développés pour une telle problématique de recherche. Ainsi,

ils clôturent nos travaux de recherche sur la modélisation des voies métaboliques en vue de la production de molécule, mais ils ouvrent une nouvelle porte sur la représentation détaillée de voies métaboliques régulées par des systèmes évolutifs et adaptables i) à la voie étudiée et ii) à l'objectif de production fixé.

5.1. Introduction

L'utilisation de la voie de la glycolyse chez le parasite *Entamoeba histolytica* est sa principale stratégie de production d'énergie, sous forme d'ATP (Saavedra *et al.*, 2005; Pineda *et al.*, 2015, 201; Saidin *et al.*, 2017). De ce fait, cette voie est soumise à différentes régulations dans le microorganisme qui dispense une production optimale d'énergie, suffisante pour permettre la survie et la multiplication du parasite au sein de son hôte. La compréhension de ces mécanismes de régulation s'avère donc très importante, particulièrement au niveau thérapeutique, lors du développement d'un vaccin ou d'un médicament. Aussi, si nous nous plaçons dans un objectif de production de molécule, il est également important de considérer deux principes :

- La maîtrise des régulations existantes au sein du parasite pour maintenir une synthèse maximale, dans le cas où nous utilisons les parasites directement comme usine de production, en conditions *in-vivo* ;
- La mise au point des mécanismes de contrôle adéquats au sein de la voie métabolique pour établir une synthèse optimale du produit, dans le cas où nous utilisons la machinerie enzymatique du parasite en conditions *in-vitro*.

Dans cette étude, nous nous intéressons plutôt au deuxième principe de production i.e. *in-vitro*. Notre objectif sera alors de mettre en place des mécanismes de contrôle adaptés à la voie basse de la glycolyse pour maintenir une synthèse optimale de pyruvate. Ce segment de la glycolyse a déjà été modélisé auparavant et part du 3-phosphoglycérate pour finir au pyruvate. Il est intéressant de noter que cette voie chez le parasite contient une réaction supplémentaire transformant le pyruvate en acétyl-CoA (figure 5.1).

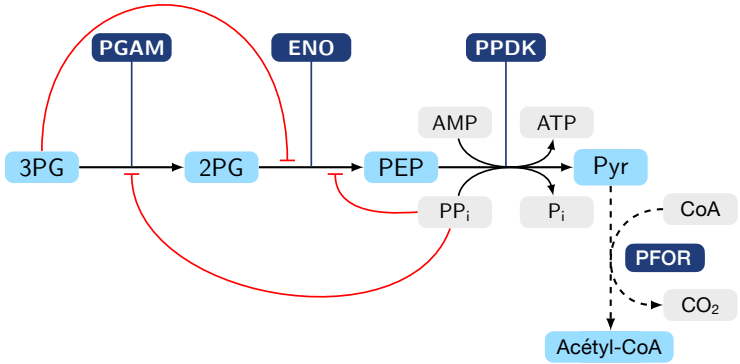

FIGURE 5.1 : *Schéma de la voie basse de la glycolyse d'E. histolytica. Formation de pyruvate (Pyr) à partir de 3- phosphoglycérate (3PG).*
La formation d'acétyl-coenzyme A (Acétyl-CoA), en lignes pointillées, est présente dans la voie naturelle, mais n'a pas été représentée dans nos modèles. Les inhibitions sont représentées en rouge. PGAM, 3-

phosphoglycérate mutase ; 2PG, 2-phosphoglycérate ; ENO, énolase ; PEP, phosphoénolpyruvate ; PPDK, pyruvate phosphate dikinase ; CoA, coenzyme A ; PFOR, pyruvate-ferrédoxine oxidoréductase (Moreno-Sánchez et al., 2008; Saavedra et al., 2007).

Aussi, qu'il s'agisse du pyruvate ou de l'acétyl-CoA, ces deux molécules sont des précurseurs de plusieurs autres molécules d'intérêt, telles que : l'éthanol, l'hydrogène, butanol, participant à la synthèse de biocarburants (Fortman *et al.*, 2008) ; les flavonoïdes, qui sont des antioxydants bénéfiques pour la santé. (Kang *et al.*, 2014) ; ou encore l'acide 3-hydroxypropionique, utilisé comme brique de production pour la formation de biopolymères (Shi *et al.*, 2014).

Le système de rétrocontrôle que nous construisons ici est un régulateur de type PID, qui agit sur les entrées de la voie métabolique pour en contrôler le flux final. Le schéma fonctionnel de notre mécanisme de rétrocontrôle est représenté en figure 5.2. Notre présent travail se divisera, comme nous l'avons annoncé en introduction de ce chapitre, en plusieurs phases :

- L'ajout de perturbations sur notre modèle de départ qui atteint normalement un état quasi-stationnaire ;
- L'identification des variables de commande sur lesquelles agira notre système de contrôle ;
- L'ajout du régulateur de type PID sur le modèle boîte-grise avec différentes variables de commande ;
- L'amélioration du régulateur afin de maintenir au mieux la valeur consigne intégrée au départ.

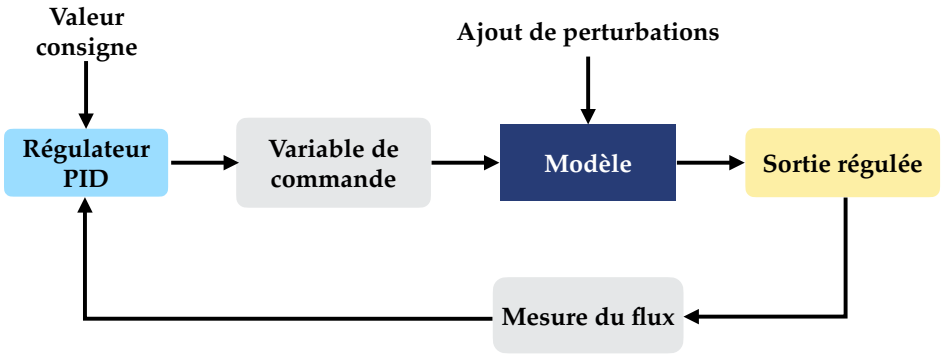

FIGURE 5.2 : *Schéma fonctionnel de la régulation effectuée sur la voie basse de la glycolyse d'E. histolytica. Le modèle de la voie métabolique est soumis à un régulateur PID qui contrôle certaines variables d'entrée du système appelées les variables de commande pour réguler la sortie. La valeur consigne de la sortie est donnée au régulateur qui la compare alors avec les mesures de flux en sortie du système.*

5.2. Méthodes

5.2.1. Modification du modèle hybride boîte-grise

Le modèle hybride boîte-grise qui a été modélisé auparavant, mène à la prédiction du flux final de la voie de la glycolyse lorsqu'elle atteint un état quasi-stationnaire. L'ajout d'un système de contrôle sur un tel modèle serait inutile. De plus, un état d'équilibre n'est pas systématiquement obtenu en production industrielle, du moins pas à la valeur optimale de production souhaitée. Aussi, afin d'évaluer la performance des modèles de rétrocontrôle à maintenir le flux final de la voie à une valeur donnée, nous procédons à quelques changements.

La première modification consiste à remplacer l'équation de la lactate déshydrogénase présente dans le modèle boîte-grise, par la réaction qui a réellement lieu dans le parasite : la transformation du pyruvate en acétyl-CoA catalysée par la pyruvate-ferrédoxine oxydoréductase (ou PFOR). Deux équations sont utilisées pour décrire cette nouvelle réaction :

- L'équation irréversible Bi-Bi :

$$v = \frac{V_{max} * [Pyr] * [CoA]}{K_{m\_Pyr} * K_{m\_CoA} + [Pyr] * K_{m\_CoA} + [CoA] * K_{m\_Pyr} + [Pyr] * [CoA]},$$

avec $\boldsymbol{v}$ la vitesse de la réaction ; $\boldsymbol{V_{max}}$ la vitesse initiale maximale de l'enzyme; [**Pyr**] et [**CoA**] les concentrations respectives en Pyr et CoA ; $\boldsymbol{K_m}$ la constante de Michaelis pour les différentes espèces indiquées.

- L'équation de la loi d'action de masse :

$v = k * [Pyr]$, où $\boldsymbol{k}$ est une constante de vitesse, $k = 10\,000\ min^{-1}$.

Les paramètres cinétiques et les concentrations utilisés pour la première équation figurent dans le tableau 5.1 ci-dessous. Ces données sont issues des travaux effectués en *in-vitro* sur le parasite étudié (Pineda *et al.*, 2010).

Parameters/Concentrations	Value
$\mathbf{V_{max}}$	900 mU
$\mathbf{K_{m\_Pyr}}$	3,500 μM
$\mathbf{K_{m\_CoA}}$	13 μM
CoA	100 μM

TABLEAU 5.1 : Paramètres cinétiques et concentrations des métabolites utilisées dans le modèle boîte-grise incluant la réaction catalysée par l'enzyme PFOR.

Nous utilisons également les concentrations en substrats et cosubstrats retrouvées dans les conditions physiologiques (tableau 5.2).

Metabolite	Physiological concentrations (in μM)
3PG	400
AMP	1,600
PP_i	450
ATP	5,000
P_i	5,400

TABLEAU 5.2 : Concentrations des métabolites utilisés dans le modèle cinétique.
Données extraites des travaux précédents et reprises dans nos travaux (Moreno-Sánchez *et al.*, 2008; Lo-Thong *et al.*, 2020). 3PG, 3-phosphoglycérate; AMP, adénosine monophosphate ; PP_i, pyrophosphate inorganique ; ATP, adénosine triphosphate ; P_i, phosphate inorganique.

Suite à cette modification du modèle, nous ajoutons les perturbations, énoncées précédemment, sur le modèle. Dans le cas d'une voie métabolique, ces perturbations peuvent être :

- L'addition d'inhibiteurs, de substances médicamenteuses ou de molécules oxydantes (par exemple : le peroxyde d'hydrogène ou H_2O_2) ;
- L'ajout de voies alternatives pouvant modifier le flux en sortie ;
- La modification des concentrations de départ.

Ces perturbations sont généralement ajoutées afin d'évaluer la capacité du régulateur à assurer un contrôle du modèle, même en présence d'évènement perturbateur. La perturbation, qui sera ajoutée au modèle et qui fera l'objet de notre étude ici, sera la modification des concentrations de départ en cosubstrats. En effet, dans le modèle boîte-grise, les concentrations en substrats et cosubstrats sont fixées pour permettre l'établissement d'un équilibre. Nous modifions alors le modèle afin de rendre variable les deux cosubstrats présents dans la voie métabolique : AMP et PP_i. Ainsi lorsqu'ils sont consommés lors de la réaction catalysée par PPDK, leur concentration diminue et ne reste pas à la valeur initiale.

Dans un second temps, nous ajoutons une autre perturbation illustrant le phénomène de dégradation des enzymes. Cette perturbation est ajoutée sous la forme du temps de demi-vie ($t_{1/2}$) uniquement pour l'enzyme PGAM. La concentration de l'enzyme PGAM est égale à : $[PGAM] = [PGAM] * \left(-\frac{\ln(2)}{t_{1/2}} \right)$ où $[\boldsymbol{PGAM}]$ est la concentration en PGAM et $\left(-\frac{\ln(2)}{t_{1/2}} \right)$ correspond à la constante de vitesse d'inactivation (El Seoud *et al.*, 2016; Dutta and Saha, 2018).

5.2.2. Modélisation du régulateur PID couplé au modèle hybride boîte-grise

Après une modification du modèle de la voie basse de la glycolyse et l'ajout de perturbations dans celui-ci, nous passons à la modélisation du régulateur PID.

Reprenons au préalable l'architecture de ce système de rétrocontrôle en boucle fermée développé au **chapitre 1**. Il est constitué de trois composantes ou autrement dit trois actions de contrôle : la composante Proportionnelle, la composante Intégrale et la composante Dérivée (figure 5.3). Ces trois composantes seront définies de manière détaillée au cours de cette section. Le système de rétrocontrôle agit sur des variables de commande présentes dans le modèle boîte-grise. Dans cette étude, nous nous intéresserons aux variables que sont les concentrations en cosubstrats : AMP et PP_i. Le modèle boîte-grise est également défini par des variables d'état qui ont pour seul objectif sa description (*e.g.*, la concentration en ATP ou en Pyr). Aussi, la variable de sortie que l'on souhaite réguler ici est le flux de pyruvate ; et la valeur consigne donnée au régulateur pour effectuer le rétrocontrôle est de 11.07 nmol·min^{-1}, correspondant à la valeur prédite par le modèle boîte-grise avant l'ajout des perturbations.

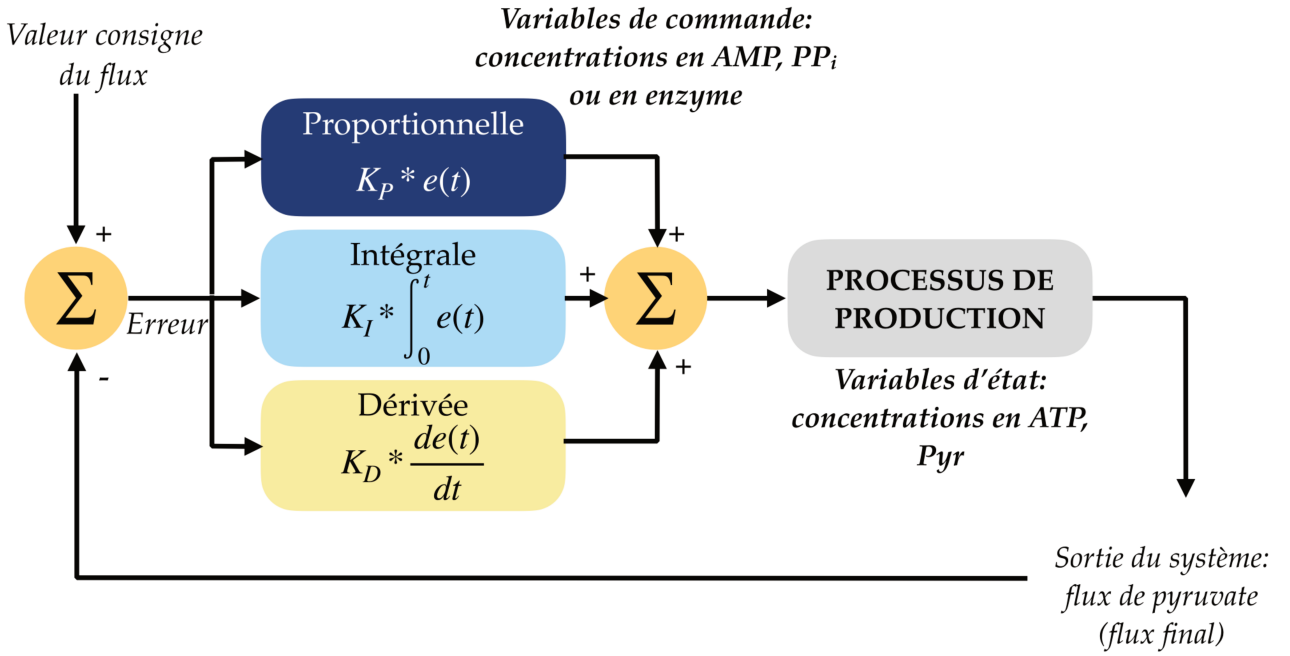

FIGURE 5.3 : *Représentation du rétrocontrôle de type PID implémenté au sein du modèle hybride de la voie basse de la glycolyse.*
La sortie du système est le flux de pyruvate (flux en sortie de la réaction catalysée par PPDK). Cette sortie est mesurée et comparée à la valeur consigne enregistrée. L'erreur $e(t)$ est ensuite intégrée au régulateur PID, qui calcule la commande finale à appliquer sur le processus de production. K_P, K_I et K_D représentent les gains de chaque terme de correction, respectivement celui du terme Proportionnelle, Intégrale et Dérivée. Il s'agit des paramètres du système de contrôle à régler.

Même s'il est constitué de trois termes différents, le régulateur PID peut contenir : i) un seul de ces termes, généralement le terme Proportionnel, il prend donc le nom de régulateur P ; ii) deux termes, soit Proportionnel-Intégral ou soit Proportionnel-Dérivée, il prend respectivement les noms de régulateur PI ou PD et iii) il peut contenir l'ensemble des termes et est dans ce cas appelé régulateur PID. Notre construction de ce rétrocontrôle se fera par étape en intégrant un terme à la fois. Nous pourrons alors évaluer la performance des différents régulateurs construits.

Le rétrocontrôle de type Proportionnel (ou régulateur P)

Le premier système de contrôle testé est le régulateur P. Il est constitué uniquement de la composante Proportionnelle qui correspond à une correction, appliquée aux variables de commande, proportionnelle à l'erreur. Il corrige donc de manière quasi-instantanée l'erreur entre le flux prédit par le modèle et la valeur consigne. Par le biais du gain (K_P) par lequel l'erreur est multipliée (figure 5.3), ce terme vient amplifier l'erreur réelle mesurée afin de corriger rapidement l'erreur au sein du modèle.

L'implémentation de ce régulateur se fait sur COPASI, où il est décrit en tant qu'évènement dans la section « *Events* ». L'évènement, et donc la régulation, est pris en compte dès le début de la simulation.

La consigne entrée est la suivante :

SI $\mathbf{Flux_{PPDK} < 11\ nmol{\cdot}min^{-1}}$ **ET** $\mathbf{[AMP] < 3{,}360\ \mu M}$ **ET/OU** $\mathbf{[PP_i] < 270\ \mu M}$

ALORS

$$[AMP] = [AMP] + K_P{*}E \qquad \text{ET/OU } [PP_i] = [PP_i] + K_P{*}E$$

Où $\mathbf{Flux_{PPDK}}$ désigne le flux final de la voie de la partie basse de la glycolyse ; **[AMP]** et $\mathbf{[PP_i]}$ sont les concentrations en AMP et PP_i ; $\boldsymbol{K_P}$ est le gain caractérisant le terme proportionnel et $\boldsymbol{E}$ est l'erreur calculée entre la valeur consigne et la valeur prédite par la modèle à chaque pas de temps (seconde) de la simulation.

Nous avons ajouté des saturateurs à notre contrôle (« *[AMP] < 3,360 µM et [PPi] < 270 µM* ») afin de diminuer les oscillations de la variable à réguler lors du contrôle, et éviter ainsi à notre système d'atteindre ses limites de prédiction. Lorsque les valeurs de concentration en AMP et en PP_i dépassent celles énoncées dans la consigne, la valeur du flux final dépasse la valeur consigne entrée au départ.

Le paramétrage des gains se fait à l'aide de RStudio avec le package « *Corc* »[5], permettant la lecture et la modification du modèle COPASI. Une gamme allant de 0-20 (sans unité) est testée pour identifier la valeur du gain proportionnel (Marlin, 2000).

Le rétrocontrôle de type Proportionnel-Intégral (ou régulateur PI)

Le deuxième régulateur créé est celui intégrant cette fois deux termes : Proportionnel et Intégral. Le terme Intégral complète la régulation menée par le terme proportionnel. Il vient stabiliser l'action de la régulation proportionnelle dans le temps et vient diminuer l'erreur statique, autrement dit l'erreur finale mesurée lorsque le système est stabilisé. Lorsque la valeur du flux mesuré en sortie du modèle se rapproche de la valeur consigne, la composante proportionnelle n'agit plus mais la composante intégrale, elle, sera toujours présente ; permettant une action continue du régulateur. La composante intégrale est également caractérisée par un gain (K_I). Le régulateur qui en résulte est plus stable.

L'évènement contenu dans le modèle hybride et illustrant le régulateur P est modifié afin d'ajouter la composante intégrale, comme suit :

[5] Le package est disponible sur Github à l'adresse suivante : https://github.com/jpahle/CoRC

SI $\mathbf{Flux_{PPDK} < 11\ nmol{\cdot}min^{-1}}$ **ET [AMP] < 3,360 μM** **ET [PP_i] < 270 μM**

ALORS

$$[AMP] = [AMP] + K_P{*}E_t + K_I{*}(E_t + E_{t-1}){*}\Delta t$$

ET/OU

$$[PP_i] = [PP_i] + K_P{*}E_t + K_I{*}(E_t + E_{t-1}){*}\Delta t$$

Où $\mathbf{Flux_{PPDK}}$ désigne le flux final de la voie de la partie basse de la glycolyse ; **[AMP]** et **[PP_i]** sont les concentrations en AMP et PP_i ; K_P est le gain caractérisant le terme proportionnel ; K_I est le gain caractérisant le terme intégral et est égal à $K_I = \frac{K_P}{y}$, avec **y** un terme à ajuster tout comme les autres gains; E_t est l'erreur calculée entre la valeur consigne et la valeur prédite par la modèle à un instant **t** ; E_{t-1} est l'erreur calculée à l'instant **t-1** ; Δt est la différence de temps entre deux calculs d'erreur au sein du modèle.

Le paramétrage des gains se fait de la même manière que pour le régulateur P. Ici, seul **y** est paramétré ; une gamme plus faible du paramètre **y** est testée (1-3).

Le rétrocontrôle filtré de type Proportionnel-Intégral (ou régulateur PI filtré)

Le dernier système de rétrocontrôle bâti dans cette étude est celui intégrant un filtre au niveau du contrôle et de manière plus précise, au niveau de la consigne entrée dans le modèle boîte-grise. Ce filtre, couramment utilisé, permet de diminuer les oscillations autour de la valeur de consigne. Il porte le nom de filtre des moments.

Il est ajouté également sur COPASI, au niveau de l'équation définissant la valeur de la variable de commande. La consigne prend alors la forme suivante :

SI $\mathbf{Flux_{PPDK} < 11\ nmol{\cdot}min^{-1}}$ **ET [AMP] < 3,360 μM** **ET [PP_i] < 270 μM**

ALORS

$$[AMP] = \alpha{*}[AMP] + (1-\alpha){*}([AMP] + K_P{*}E_t + K_I{*}(E_t + E_{t-1}){*}\Delta t)$$

ET/OU

$$[PP_i] = \alpha{*}[PP_i] + (1-\alpha){*}([PP_i] + K_P{*}E_t + K_I{*}(E_t + E_{t-1}){*}\Delta t)$$

Avec α désignant un paramètre à ajuster et compris entre 0-1.

Le filtre a été ajouté sur les meilleurs modèles de rétrocontrôle que nous avons développés au début de nos travaux.

5.2.3. Évaluation de la performance des systèmes de rétrocontrôle construits

Une fois les systèmes de rétrocontrôle construits, il est utile de comparer leurs performances à maintenir le flux final de voie à la valeur consigne renseignée dans le régulateur. Cette évaluation se fait en fonction du RMSE. L'équation pour le calcul du RMSE est la suivante :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$$

Où $\boldsymbol{Y_i}$ et $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ sont respectivement les valeurs prédites et la valeur consigne de référence, **n** est le nombre total de valeurs et **i** = 1, 2…n.

L'évaluation de la performance des régulateurs est également effectuée au travers de deux critères supplémentaires (Bucz and Kozáková, 2018; Scherlozer *et al.*, 2016) :

- Le temps de montée (T_m) : il correspond au temps qu'il faut pour atteindre la valeur consigne. Plus ce temps est court, plus le régulateur est rapide.
- Le dépassement (D) : il correspond à l'écart entre la consigne et la valeur maximale atteinte. Il est souvent exprimé en pourcentage et est calculé de la manière suivante : $D = 100\frac{y_{max} - y(\infty)}{y(\infty)}$, où $\boldsymbol{y_{max}}$ est la valeur maximale du flux atteinte et $\boldsymbol{y}(\boldsymbol{\infty})$ est celle obtenue quand le système est stable.

5.3. Résultats et discussion

5.3.1. Rectification du modèle hybride boîte-grise

Dans le but de préparer le modèle de la voie basse de la glycolyse à l'implémentation du système de rétrocontrôle, nous opérons quelques modifications au modèle boîte-grise. La première modification consiste à modifier la réaction de conversion du pyruvate en lactate par celle de la conversion du pyruvate en acétyl-CoA, effectuée par la PFOR. La première équation qui est utilisée est la réaction irréversible Bi-Bi ; les résultats de ce nouveau modèle sont présentés sur la figure 5.4 ci-dessous.

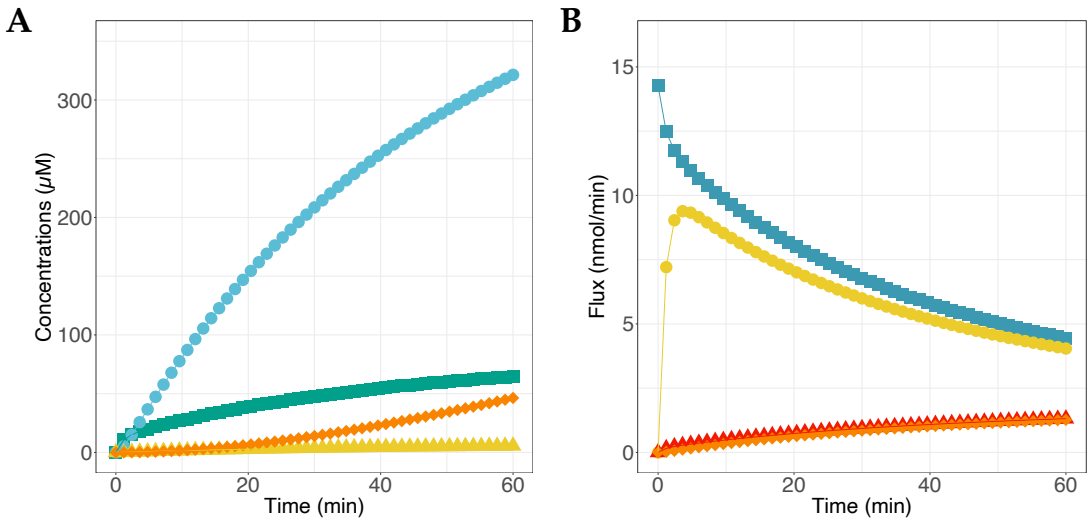

FIGURE 5.4 : *Prédictions des concentrations en métabolites et des flux par le modèle incluant la réaction catalysée par PFOR et utilisant l'équation Bi-Bi.*
*(**A**) Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles), en jaune pour Pyr (triangles) et en orange pour l'Acétyl-CoA (losanges). (**B**) Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles), en rouge pour PPDK (triangles) et en orange pour PFOR (losanges).*

Au bout d'une heure de simulation, nous obtenons une concentration de ~64.59 μM pour 2-PG pour une valeur expérimentale de 58 ± 29 μM et ~321.36 μM pour PEP pour une valeur expérimentale mesurée à 37 ± 16 μM (figure 5.4 A). Ce modèle est le premier à présenter une valeur pour 2-PG qui soit aussi proche de la valeur retrouvée en expérimental (Moreno-Sánchez *et al.*, 2008). Même si la courbe de la concentration en pyruvate semble être nulle visuellement, nous obtenons une concentration finale de Pyr ≈ 5.58 μM ; ce qui est considérablement plus élevé que la valeur que nous obtenions avec le modèle boîte-grise de départ (~1,26.10^{-2} μM). Cela vient appuyer

la proposition qui avait été faite précédemment au chapitre 2 « Modélisation de voies métaboliques par une méthode dite « boîte-grise », pour atteindre la quantité « totale » d'ATP synthétisée par PPDK. Rappelons-le, en raison de son équilibre thermodynamique, la réaction de synthèse d'ATP n'est pas favorisée dans les conditions physiologiques en raison de son équilibre thermodynamique (Varela-Gómez *et al.*, 2004). Ainsi, nous avons émis l'hypothèse qu'une accumulation de substrats ou une diminution des produits dans le système pourrait rendre compte de la synthèse totale « réelle » d'ATP par l'enzyme PPDK. Cette hypothèse est vérifiée ici puisque, l'ajout d'enzymes permet bien la diminution du pyruvate dans le système, ce qui a pour effet de favoriser la réaction catalysée par PPDK vers la synthèse d'ATP et par conséquent d'augmenter la concentration en pyruvate et en ATP.

En ce qui concerne les flux de nos différentes réactions, nous observons qu'ils sont dans l'ensemble moins élevés que ceux obtenus par le modèle boîte-grise, avant les modifications (figure 5.4B). De plus, le modèle n'atteint pas le flux final retrouvé lors de la reconstitution *in-vitro* qui était de 27 nmol·min^{-1} (Moreno-Sánchez *et al.*, 2008), il est ici prédit à ~1.3 nmol·min^{-1}.

Par conséquent, nous décidons d'utiliser une équation plus simple, qui est celle de la loi d'action de masse, pour représenter la réaction catalysée par l'enzyme PFOR (figure 5.5).

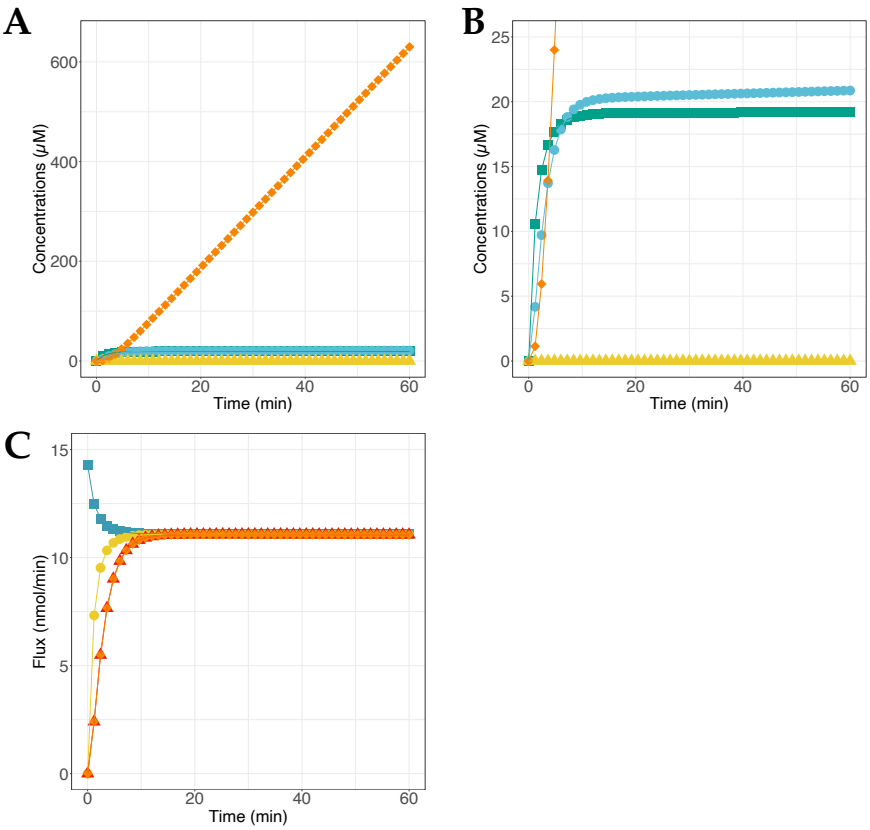

FIGURE 5.5 : *Prédictions des concentrations en métabolites et des flux par le modèle incluant la réaction catalysée par PFOR et utilisant l'équation de la loi d'action de masse.*

(A, B) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles), en jaune pour Pyr (triangles) et en orange pour l'Acétyl-CoA (losanges). La figure B représente un agrandissement de la zone basse des concentrations.* ***(C)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles), en rouge pour PPDK (triangles) et en orange pour PFOR (losanges).*

Ces nouveaux résultats nous montrent des concentrations moins élevées que les précédentes pour 2-PG ≈19.21 μM et PEP ≈20.87 μM (figure 5.5 A et B). La concentration de pyruvate ici est quasiment nulle. Cela s'explique par la production instantanée d'acétyl-CoA à partir du pyruvate, synthèse qui atteint ≈630 μM au bout d'une heure de simulation (figure 5.5 A). Ces concentrations sont inférieures à celles mesurées lors de la reconstitution *in-vitro* (Moreno-Sánchez *et al.*, 2008) qui étaient de 58 μM pour 2-PG et 106.3 μM pour PEP. Quant au flux, il est plus élevé que le précédent et nous obtenons bien un système qui atteint un état stationnaire à 11.07 nmol·min^{-1} (figure 5.5 C). Quand bien même ce flux est très inférieur à celui retrouvé lors de la reconstitution *in-vitro*, nous gardons ce modèle pour la suite de notre étude sur la mise en place du contrôle, puisqu'il représente au mieux la voie de la glycolyse du parasite étudié. La valeur consigne entrée au sein du régulateur PID sera celle correspondante au flux final de pyruvate obtenu ici avant l'ajout de perturbations, à savoir : $Flux_{PPDK} \approx$ 11.07 nmol·min^{-1}. Par ailleurs, la difficulté de modélisation de la réaction catalysée par PFOR est expliquée par la faible connaissance que nous avons sur la cinétique de cette enzyme (Saavedra *et al.*, 2007). Dans d'autres travaux, cette réaction a été regroupée à celle catalysée par l'aldéhyde déshydrogénase (AldDH) avant d'être modélisée par une réaction réversible impliquant deux substrats (ou « *Reversible Bisubstrate reaction* ») (Saavedra *et al.*, 2007). Cette enzyme a été identifiée comme cible des espèces réactives à l'oxygène (ou « *Reactive Oxygen Species* », ROS) dans certaines conditions (Pineda *et al.*, 2010), elle pourrait donc faire l'objet d'ajout de perturbations dans un système tel que le nôtre.

Suite à la modification de notre modèle hybride, nous lui ajoutons quelques perturbations. En effet, notre modèle demeure dans un état quasi-stationnaire car il se trouve que les concentrations initiales en 3-PG (substrat initial) et en AMP et PP_i (cosubstrats) sont fixées. Par conséquent, leur valeur ne varie pas durant la simulation. Nous modifions donc les paramètres des cosubstrats pour les rendre variables et nous laissons la concentration en 3-PG constante pour mimer un ajout continu et constant de substrat dans notre système de départ.

Les nouveaux résultats de ce modèle « perturbé » sont illustrés en figure 5.6. Tout d'abord, nous observons bien une irrégularité au niveau des concentrations prédites et des flux prédits. Pour les concentrations, elles sont plus importantes que celles obtenues avec le modèle atteignant l'état quasi-stationnaire, avec 2-PG ≈ 63.36 μM et PEP ≈ 300.65 μM. Il est intéressant de remarquer que ce

nouveau modèle donne des résultats similaires à ceux obtenus par le premier modèle présenté dans cette section (figure 5.4A), utilisant l'équation Bi-Bi pour modéliser la réaction catalysée par PFOR. Ceci n'est valable que pour les concentrations des deux premiers métabolites, 2-PG et PEP, puisque celle de Pyr est maintenant à $4{,}07.10^{-5}$ μM.

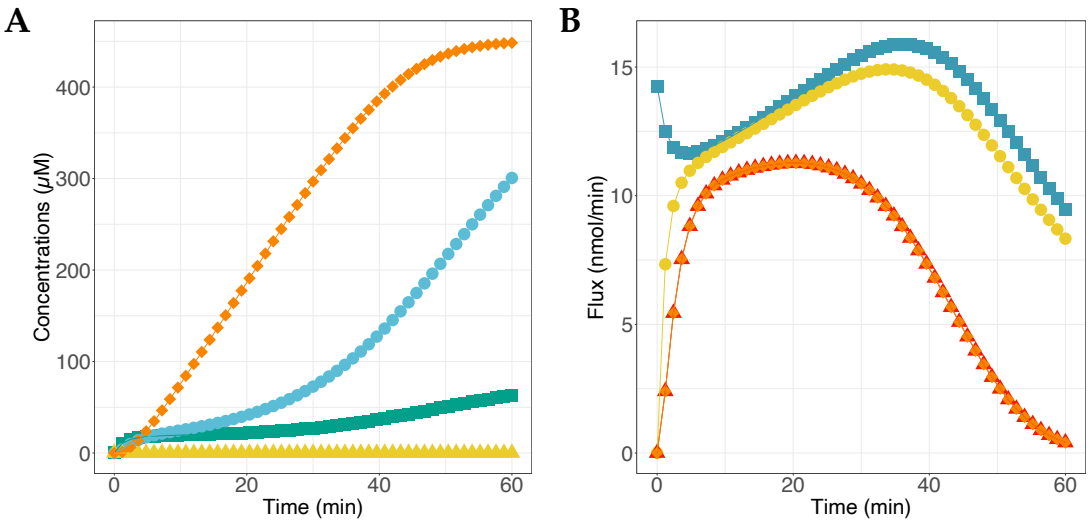

FIGURE 5.6 : *Prédictions des concentrations en métabolites et des flux par le modèle utilisant l'équation Tri-Réactants.*
(A) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles) et en jaune pour Pyr (triangles).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles) et en rouge pour PPDK (triangles).*

Les flux prédits pour chaque tronçon de la voie basse de la glycolyse n'atteignent pas une seule et même valeur (11.07 nmol·min^{-1}) et sont plus faibles que ceux mesurés sans les perturbations après une heure de simulation (figure 5.5). Ces flux augmentent au début de la simulation, puis diminuent, jusqu'à atteindre 0.41 nmol·min^{-1} pour le flux final. Cette variation du flux est « normale » puisque les substrats et cosubstrats seront consommés lors des différentes réactions, ce qui entraînera la diminution de leur concentration, qui n'est pas renouvelée et qui entraîne donc une diminution du flux de la réaction et du flux final.

Ces résultats nous indiquent que le modèle a bien été perturbé suite aux modifications faites et qu'il nécessite, de ce fait, un système de contrôle pour pouvoir réguler le flux en sortie de la voie.

5.3.2. Modélisation des systèmes de rétrocontrôle

Dans les paragraphes qui vont suivre, plusieurs systèmes de rétrocontrôle seront ajoutés au modèle de la voie basse de la glycolyse d'*E. histolytica*. Ces différents régulateurs seront analysés

puis comparés les uns aux autres au niveau de leur contrôle du flux en sortie de ce segment de voie métabolique.

Le rétrocontrôle de type Proportionnel (ou régulateur P)

Le premier rétrocontrôle que nous ajoutons sur notre modèle boîte-grise perturbé est le rétrocontrôle de type Proportionnel. Comme son nom l'indique, il ne contient que la composante proportionnelle d'un régulateur PID ainsi qu'une seule variable de commande : la concentration en AMP. Une gamme allant de 1-10 a été testée pour la valeur du gain, et la simulation est lancée pour 120 min correspondant à deux heures de production dans un bioréacteur (figure 5.7).

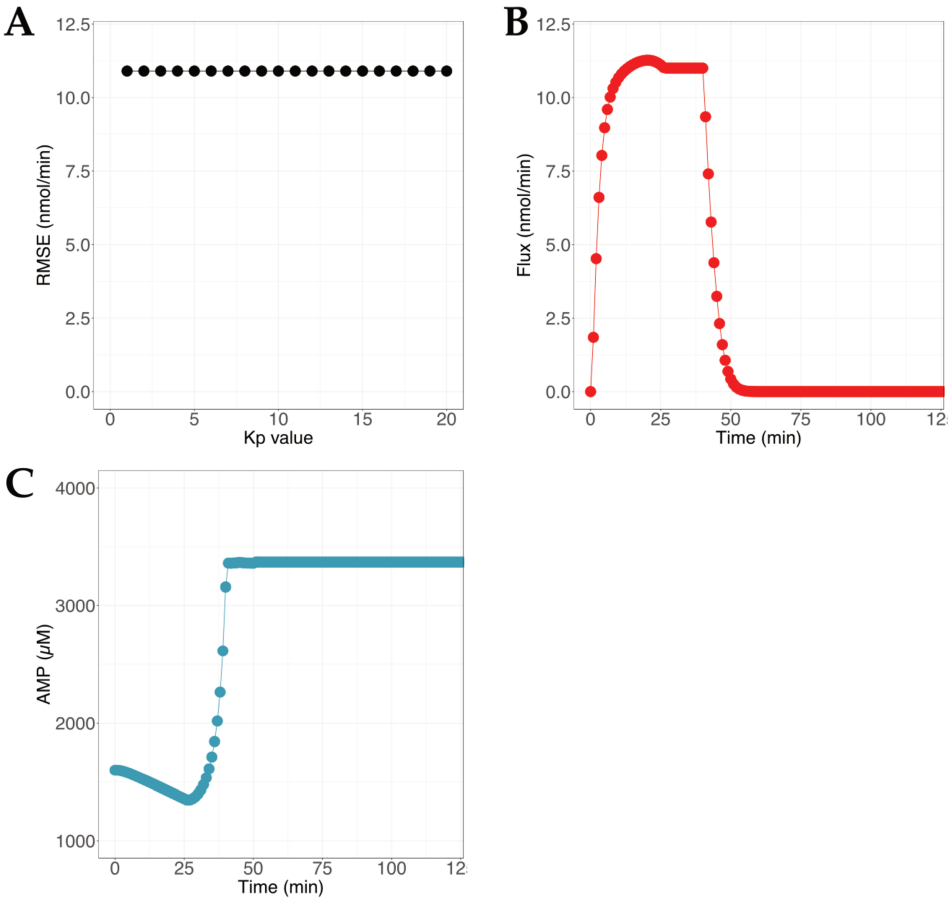

FIGURE 5.7 : *Réglage du régulateur P appliqué à l'unique commande de variable AMP et prédictions du flux final pr le modèle.*
*(**A**) Variation des valeurs de RMSE calculées pour la gamme de K_P testée. (**B, C**) Le modèle contenant le régulateur P, avec K_P=1 a été utilisé comme exemple ici. Le flux final de la voie est représenté en rouge (**B**). La concentration de AMP en fonction du temps de simulation est représentée en bleu (**C**).*

Lors du paramétrage du gain K_P, nous remarquons que le RMSE est le même quelle que soit la valeur du gain et il est égal à ~10.9 nmol·min^{-1} (figure 5.7A). De plus, cette valeur est très élevée, ce qui nous indique que le flux ne semble pas être bien contrôlé par le système de rétrocontrôle.

Quand nous regardons notre flux final pour un régulateur P avec un gain K_P=1, nous observons un premier pic à 11.3 nmol·min^{-1} atteint à ~25 min, puis une nette diminution à partir de 40 min (figure 5.7B). Cette diminution arrive de manière concomitante avec l'augmentation de la concentration en AMP, qui atteint une valeur de 3 370.7 μM (figure 5.7C). Nous en concluons que notre système arrive à saturation à partir de 40 min de simulation.

Nous supposons que la variable de commande n'est pas la meilleure que nous puissions utiliser dans notre modèle de la voie basse de la glycolyse. Nous proposons donc de réitérer l'expérience en changeant notre variable de commande par la concentration en PP_i, qui est l'autre cosubstrat de notre modèle. Les nouveaux résultats, pour un temps de simulation de cinq heures, sont représentés dans la figure 5.8 ci-dessous.

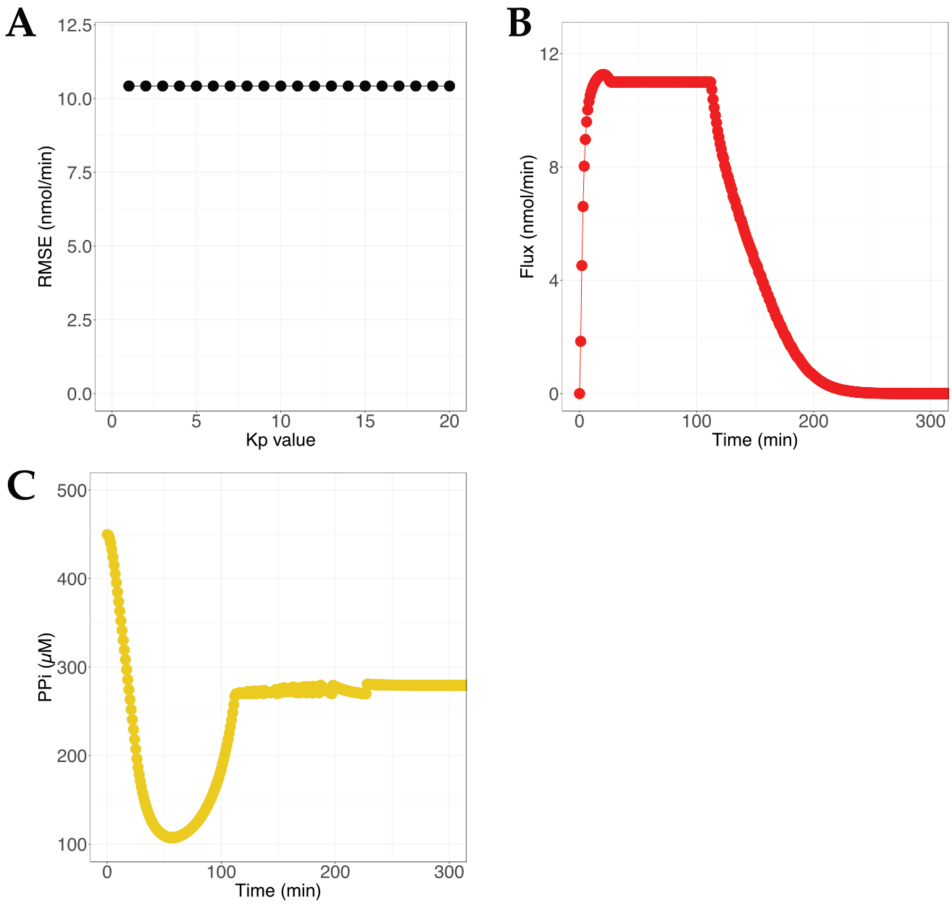

FIGURE 5.8 : *Réglage du régulateur P appliqué à l'unique commande de variable PP_i et prédictions du flux final par le modèle.*
***(A)** Variation des valeurs de RMSE calculées pour la gamme de K_P testée. **(B, C)** Le modèle contenant le régulateur P, avec K_P=1 a été utilisé comme exemple ici. Le flux final de la voie est représenté en rouge **(B)**. La concentration de PP_i en fonction du temps de simulation est représentée en jaune **(C)**.*

La valeur calculée du RMSE est toujours constante pour les différentes valeurs de gain testées et elle est de ~10.43 nmol·min^{-1} (figure 5.8A). Cette valeur est légèrement plus basse que celle du

régulateur P agissant sur la concentration d'AMP uniquement (figure 5.7A). Le modèle n'atteint pas la valeur consigne de 11.07 nmol·min^{-1}. Néanmoins le flux semble être maintenu à cette valeur pendant un certain temps, avant de diminuer à nouveau à partir de 110 min (figure 5.8B). Pour ce qui est de la concentration en PP_i, elle diminue dans un premier temps, ce qui est normal puisque le cosubstrat est consommé ; puis elle augmente et arrive à une concentration de 279.62 μM et ce jusqu'à la fin de la simulation (figure 5.8C). Nous en concluons, une fois encore, que le système de rétrocontrôle arrive rapidement à saturation et que le régulateur P ne fonctionne plus au bout d'un certain temps. Même si ce temps est plus long dans le cas d'une régulation agissant sur la concentration en PP_i ; il n'est pas suffisant pour établir un contrôle stable du flux final de la voie métabolique étudiée. Par ailleurs, nous savons que lorsque le contrôle est effectué sur PP_i, seule cette concentration sera modifiée lors du rétrocontrôle, tandis que celle en AMP va diminuer, sans apport supplémentaire au cours du temps. Cette déplétion en AMP pourrait expliquer la diminution du flux au bout de 100 min (figure 5.8B). Nous faisons également la même hypothèse pour le premier modèle où seule la concentration en AMP était contrôlée par le régulateur. Ainsi, nous supposons qu'un régulateur P agissant sur les deux variables de commande (AMP et PP_i) serait plus approprié pour réguler le flux de la voie basse de la glycolyse.

Nous construisons un nouveau régulateur P en lui ajoutant deux variables de commandes : les concentrations en AMP et PP_i. Le paramétrage des gains pour les deux variables de commande se fait de la même manière que pour les premiers régulateurs P. Le temps de simulation est de 24 h (1 440 min), le RMSE calculé entre la valeur consigne et les valeurs du flux final durant la simulation sont répertoriés dans la figure 5.9.

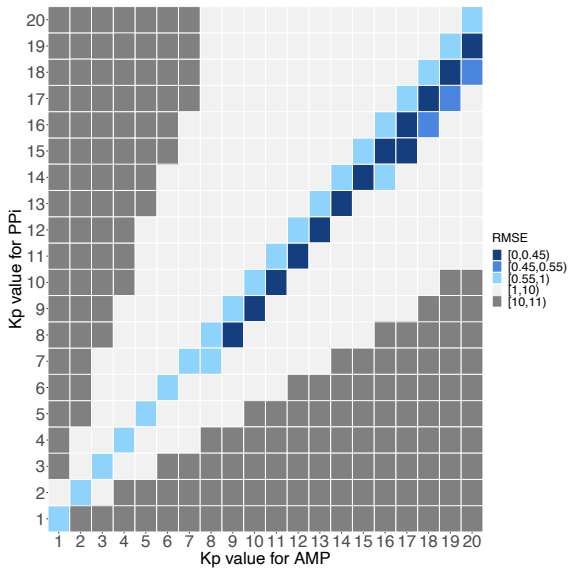

FIGURE 5.9 : *Variation du RMSE des systèmes de rétrocontrôle (régulateur P) agissant sur les concentrations en AMP et PP_i, selon les valeurs des gains K_P.*

Nous remarquons que lorsque les gains sont identiques pour AMP et PP_i, nous obtenons une faible erreur entre la valeur consigne et le flux prédit par notre modèle boîte-grise, avec un RMSE compris entre 0.55 et 1 nmol·min^{-1} (figure 5.9). Le système avec lequel nous obtenons la plus faible erreur est celui avec K_{P_AMP}= 13 et K_{P_PPi}= 12 (RMSE≈ 0.449 nmol·min^{-1}). Regardons de plus près les résultats obtenus avec quelques-uns de ces régulateurs P (figure 5.10).

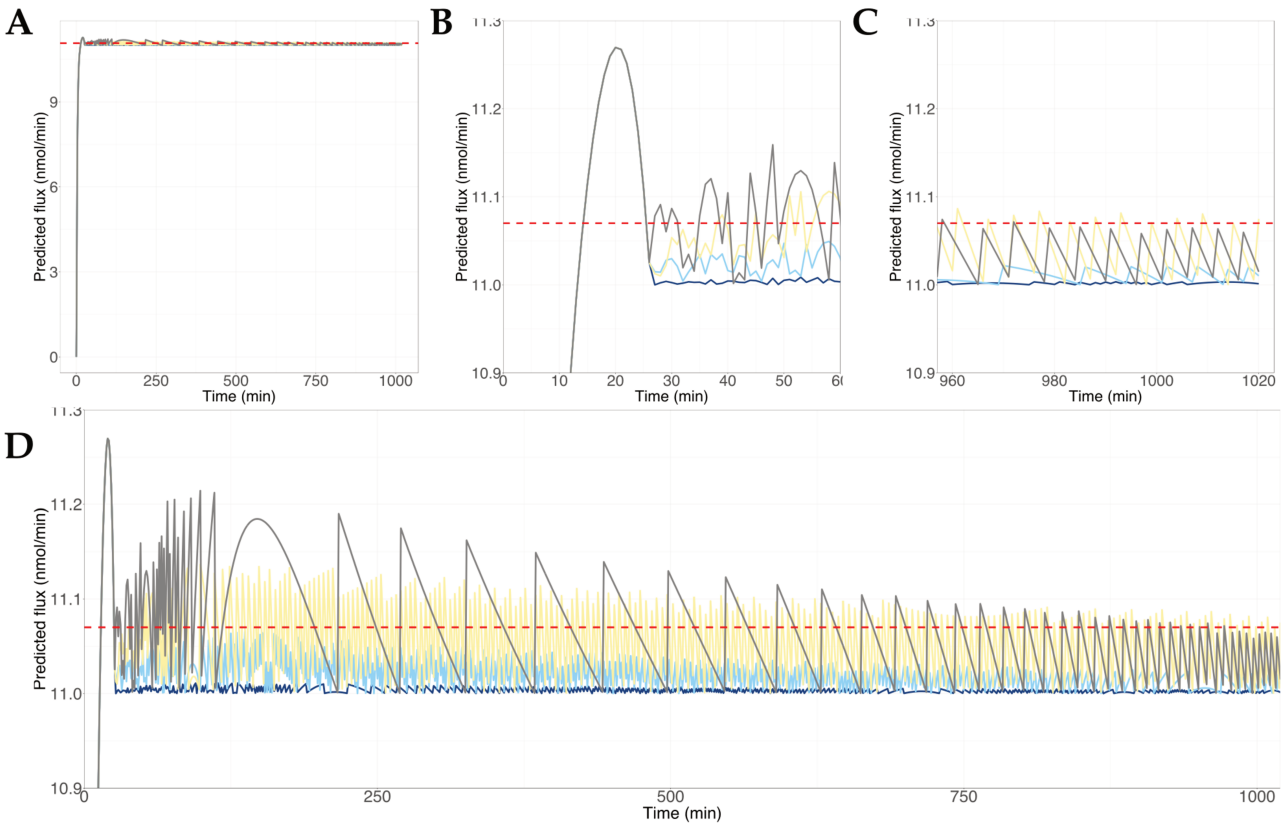

FIGURE 5.10 : *Prédictions du flux final de la voie pour 17h de simulation.*
***(A)** Représentation de l'ensemble de la simulation pour 4 régulateurs différents. **(B-D)** Représentation du flux prédit pour la zone entre 10.9-11.3 nmol·min^{-1} pour la première heure **(B)**, la dernière heure **(C)** ou l'ensemble de la simulation **(D)**. La ligne pointillée en rouge désigne la valeur consigne enregistré pour chaque régulateur P. Les couleurs représentent des systèmes de rétrocontrôle différents, à savoir : K_{P_AMP}=1 et K_{P_PPi}=1 (bleu foncé), K_{P_AMP}=6 et K_{P_PPi}=6 (bleu clair), K_{P_AMP}=13 et K_{P_PPi}=12 (jaune) et K_{P_AMP}=20 et K_{P_PPi}=20 (gris).*

Dans l'ensemble, les régulateurs semblent fonctionner correctement (figure 5.10 A). Lorsque nous faisons un agrandissement sur la zone d'intérêt, nous constatons un dépassement de la valeur consigne de ~0.2 nmol·min^{-1}. Cela peut être considéré comme un avantage lorsque les systèmes de production sont sensibles aux changements, tels qu'une accumulation de molécules toxiques pour l'organisme de production, des changements de pH ou de température (Frahm *et al.*, 2009; Ozturk and Hu, 2006). Il est intéressant de remarquer que le dépassement et le temps de montée sont les mêmes pour l'ensemble des régulateurs, quelle que soit la valeur du gain

proportionnel utilisée. Ces résultats sont étonnants puisque l'augmentation du gain K_P a pour effet d'augmenter le dépassement du système et d'engendrer la diminution du temps de montée (Alargt and Ashur, 2013; Hussien *et al.*, 2015). Aussi, les oscillations au début de la simulation sont plus importantes que celles à la fin de la simulation (figure 5.10 A and B). Cela nous montre qu'il y a une stabilisation du système de rétrocontrôle au fil du temps. Cette stabilisation n'est pas tout à fait complète, effectivement, nous observons encore la présence d'oscillations plus ou moins importantes selon les régulateurs testés, ce qui ne devrait pas être le cas d'un système de rétrocontrôle idéal (Scherlozer *et al.*, 2016).

Nous notons une plus importante oscillation lorsque nous augmentons le gain K_P pour AMP et PP_i (figure 5.10 D). Le régulateur P le plus stable est celui présentant des gains de K_{P_AMP}=1 et K_{P_PPi}=1, avec de très faibles oscillations. Néanmoins, la valeur du flux est inférieure à celle de la valeur consigne. À l'inverse, le régulateur le moins stable est celui qui possède les gains les plus élevés (K_{P_AMP}=20 et K_{P_PPi}=20). En effet, le temps de stabilisation du régulateur est de 250 min et les oscillations sont les plus élevées comparées à celles des autres systèmes de contrôle.

Nous analysons les résultats du meilleur régulateur P obtenu en termes de RMSE, avec : K_{P_AMP}=13 et K_{P_PPi}=12 (figure 5.11). Comme le système de contrôle présente toujours des oscillations à la fin de la simulation et n'atteint pas une valeur constante de flux, nous utilisons donc la valeur de flux obtenue à la fin de la simulation lors du calcul du dépassement, à la place du terme $\boldsymbol{y(\infty)}$, qui est la valeur obtenue quand le système est stable. Le dépassement du régulateur est de 1.71 % par rapport à la valeur consigne ; et en ce qui concerne le temps de montée, il est de 14.25 min (figure 5.10 B). Le contrôle devient quasiment stable à partir de 199 min, avec des oscillations régulières de $9{,}41.10^{-2}$ nmol·min^{-1}.

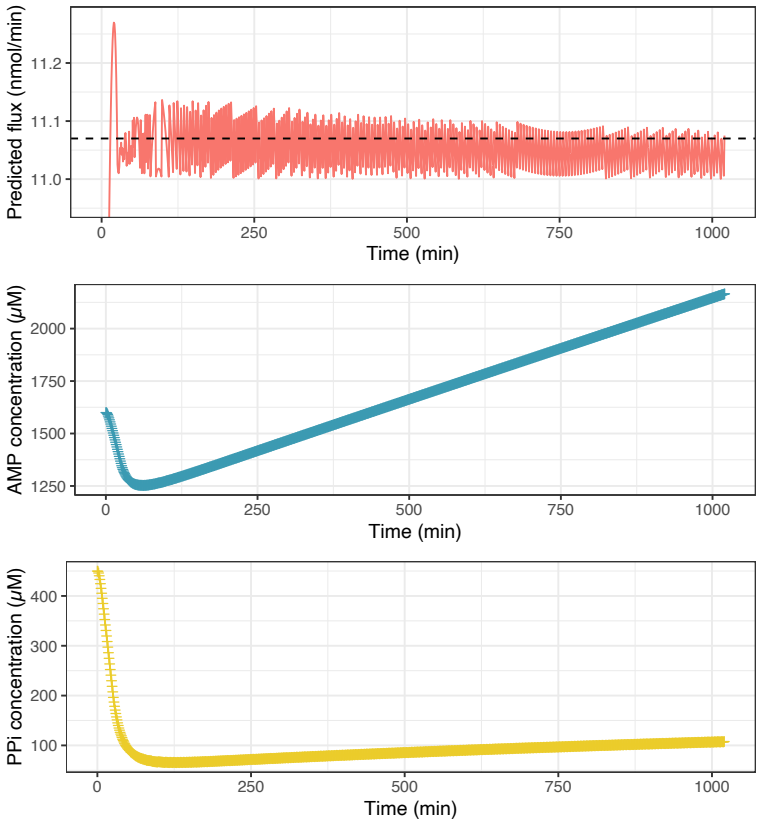

FIGURE 5.11 : *Prédictions du modèle contrôlé par le régulateur P (K_{P_AMP}=13 et K_{P_PPi}=12) pour 17h de simulation.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. La ligne en pointillée noire représente la valeur consigne.

Quand nous nous intéressons à la concentration en PP_i, nous observons que celle en PP_i atteint un plateau à partir de 100 min, après cela, nous notons une légère augmentation jusqu'à la fin de la simulation (figure 5.11). Pour la concentration en AMP, celle-ci diminue jusqu'aux alentours de 50 min, puis augmente jusqu'à la fin de la simulation. Une telle augmentation pourrait avoir des conséquences néfastes sur un système de production utilisant le parasite. Mais, comme nous nous plaçons dans un système acellulaire, ce problème ne se pose pas. En revanche, il est primordial de ne pas gaspiller les ressources primaires utilisées dans un système de production. Par ailleurs, nous savons qu'une enzyme est considérée comme étant saturée à une concentration en substrat 10 voire 20 fois supérieure à la concentration en substrat. Pour l'AMP, cette concentration saturante serait de 320 000 μM. De ce fait, après 17h de simulation, nous nous trouvons à une concentration égale à 2 165.97 μM. Nous en concluons que nous sommes en-dessous de la concentration saturante en AMP pour un système de production acellulaire. Néanmoins, afin de ne pas nous risquer à utiliser plus de substrat qu'il n'en faudrait pour maintenir le flux à un niveau optimal, une simulation plus longue serait nécessaire pour déterminer si la concentration en AMP ne

dépasse pas la concentration saturante citée plus tôt ou même déterminer si cette concentration atteint une valeur quasiment constante, comme c'est le cas pour la concentration en PP_i.

Bien que le précédent régulateur donne de meilleurs résultats pour le RMSE, il présente tout de même de larges oscillations. Nous nous intéressons alors au régulateur P avec les gains suivants : K_{P_AMP}=1 et K_{P_PPi}=1 (figure 5.12). Ce régulateur P a un avantage considérable : ses oscillations sont infimes et de l'ordre de $3{,}11.10^{-3}$ nmol·min^{-1}. Même si la valeur de flux n'atteint pas la valeur consigne lorsque le système est stable, l'erreur statique est faible ($6{,}42.10^{-2}$ nmol·min^{-1}). Le système de contrôle devient quasiment stable à partir de 100 min, avec les mêmes motifs d'oscillation qui surviennent. Par ailleurs, ce système de rétrocontrôle présente la même performance que le régulateur précédent en termes de temps de montée (14.25 min), mais le dépassement est plus élevé et est égale à ~2.45 %.

Il est intéressant de constater que la variation de la concentration en PP_i et celle en AMP est la même, avec une augmentation d'environ 200 μM au bout de la simulation (figure 5.12). La concentration en AMP au bout de 17 h est plus basse que celle obtenue par le précédent régulateur P ; tandis que celle en PP_i est plus élevée que la précédente valeur (~100 μM). Si nous considérons à nouveau la concentration saturante en PP_i, celle-ci est de 9 000 μM ; nous ne dépassons donc pas la valeur au sein de notre système de production. Il serait intéressant là-aussi de faire fonctionner le système pour un temps plus long, afin de voir s'il atteint un équilibre, notamment au niveau des concentrations en cosubstrat.

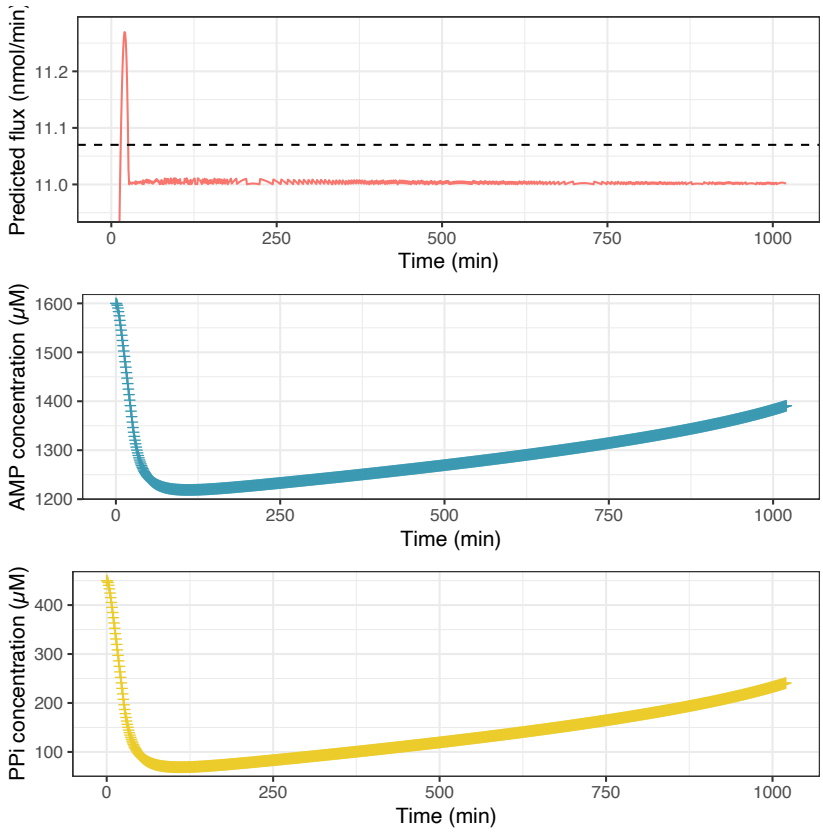

FIGURE 5.12 : *Prédictions du modèle contrôlé par le régulateur P (K_{P_AMP}=1 et K_{P_PPi}=1) pour 17h de simulation.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. La ligne en pointillée noire représente la valeur consigne.

Ces résultats suggèrent que les régulateurs de type P parviennent à contrôler le flux en sortie d'une voie métabolique en agissant sur les concentrations en cosubstrats. Notre meilleur système de contrôle étant celui avec les meilleurs critères de performance, à savoir le régulateur P avec K_{P_AMP}=13 et K_{P_PPi}=12. Ce type de contrôle (régulateur PID) agissant sur plusieurs variables de commande, a été développé pour le contrôle de la température et de l'humidité pour un système d'enceinte destiné à la bio-impression de molécules biologiques. Le système de rétrocontrôle utilisé est un régulateur PID, comprenant donc l'ensemble des composantes du régulateur PID, qui est capable de stabiliser le système en 311 s (Matamoros *et al.*, 2020).

Néanmoins, les régulateurs de ce type sont très peu utilisés en raison de leur incapacité à contrôler les processus étudiés, tels que la régulation de production chimique au sein de réacteurs chimiques (Aslam and Kaur, 2011). De plus, il peut y avoir au sein de ce système une erreur statique persistante qu'ils ne peuvent corriger, et qui peut poser dans certain cas ; sans parler du déclin du système de régulation en présence de perturbations (Ellis, 2012; Lavric *et al.*, 2005). C'est

pourquoi, nous décidons d'implémenter un régulateur Proportionnel-Intégral ou régulateur PI dans le prochain paragraphe.

Le rétrocontrôle de type Proportionnel-Intégral (ou régulateur PI)

Nos systèmes de régulation abordés au paragraphe précédent intégraient uniquement le terme Proportionnel. Dans cette partie, nous nous intéressons à l'implémentation d'un régulateur Proportionnel-Intégral sur notre modèle boîte-grise « perturbé ». Pour ce faire, nous partons des deux meilleurs régulateurs obtenus précédemment : **régulateur PI 1** (K_{P_AMP}=13 et K_{P_PPi}=12) et **régulateur PI 2** (K_{P_AMP}=1 et K_{P_PPi}=1) avec un gain K_P déjà paramétré pour notre modèle de production. À ces modèles a été rajouté le gain de la composante Intégrale, $K_I = \frac{K_P}{y}$; nous paramétrons alors la variable $\boldsymbol{y}$ de notre gain K_I, le terme K_P étant déjà paramétré (figure 5.13). La valeur du RMSE est inférieure à 0.5 nmol·min^{-1} pour des valeurs de $\boldsymbol{y}$ supérieures à 1. Au vu des résultats, nous décidons de fixer la valeur de $\boldsymbol{y}$ à 10, pour laquelle RMSE≈ 0.405 nmol·min^{-1} ; les changements étant peu importants à partir de cette valeur.

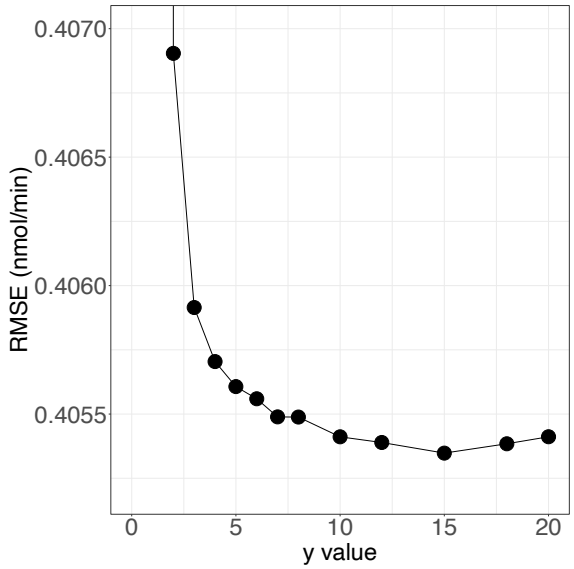

FIGURE 5.13 : *Variation du RMSE calculé pour les différents régulateurs PI, constitué d'un gain K_I dont y varie entre 0-20.*
La valeur du RMSE pour y=1 n'est pas représentée sur le graphique car il est de 5.75 nmol·min^{-1}.

Nous évaluons alors la performance de ce nouveau système de régulation (figure 5.14). La prédiction du flux pendant une simulation de 17 h nous montre que le régulateur PI 1 présente des oscillations plus ou moins régulières à partir de 125 min. Par ailleurs, nous remarquons que l'amplitude des oscillations ne varie pas beaucoup au cours de la production.

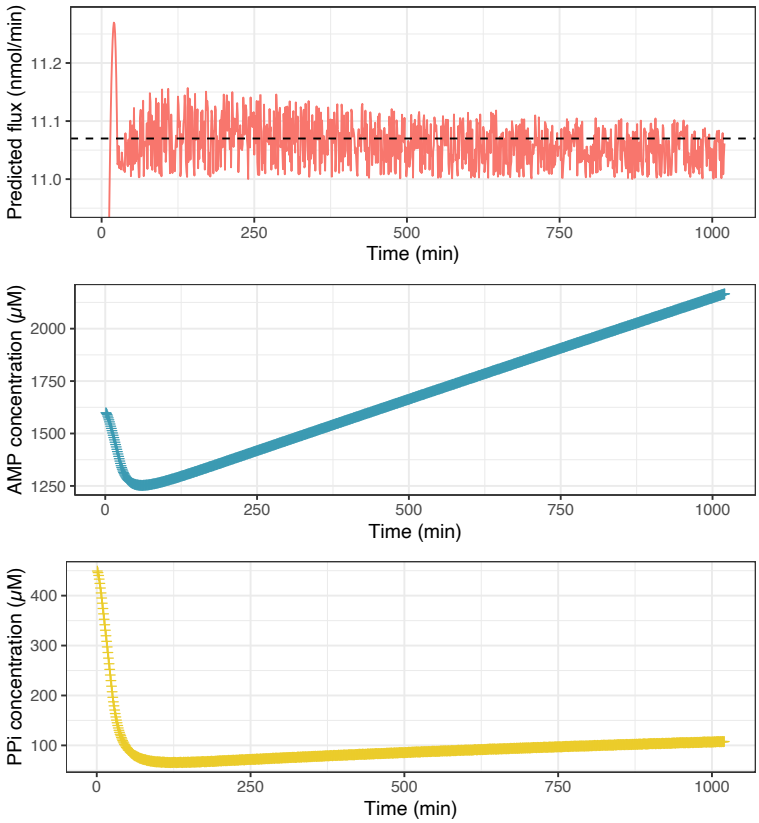

FIGURE 5.14 : *Prédictions du modèle contrôlé par le régulateur PI 1 (K_{P_AMP}=13, K_{P_PPi}=12 et K_{I_AMP}=1.3 et K_{I_PPi}=1.2) pour 17h de simulation.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. La ligne en pointillée noire représente la valeur consigne.

L'ajout du terme Intégral ne semble pas avoir d'effet sur le temps de montée qui est toujours de 14.25 min. F. S. Alargt *et al.* considèrent que lorsque le gain K_I augmente, le temps de montée diminue légèrement, tandis que le dépassement augmente (Alargt and Ashur, 2013). Dans notre cas, lorsque le gain K_I augmente, et que le modèle présente un y ($K_I = \frac{K_P}{y}$) plus petit, cela n'a aucun effet sur les deux critères cités plus tôt (figure 5.15). En effet, les deux courbes se superposent au début de la simulation. Nous supposons que le régulateur PI 1 n'a pas d'impact important sur la rapidité du système de rétrocontrôle implémenté sur le modèle boîte-grise de la voie basse de la glycolyse. Le dépassement calculé pour ce système de rétrocontrôle est de 1.89%. Par contre, il semble que les oscillations soient plus grandes lorsque le gain K_I augmente ; ce qui signifie que le système est généralement moins stable pour ce paramétrage du régulateur PI 1. Cependant, nous gardons à l'esprit que les oscillations notées en figure 5.11 et figure 5.14 proviennent d'un agrandissement sur une zone précise du flux et elles correspondent donc à de petites variations dans notre procédé de production.

En ce qui concerne les concentrations de nos deux variables de commandes, il ne semble pas y avoir de différences majeures entre celles issues du rétrocontrôle par le régulateur P développé dans le paragraphe précédent et celles issues du régulateur PI 1 présenté dans cette partie (figure 5.14). A l'issue de la simulation, nous avons des concentrations légèrement plus élevées en AMP (2 166.3 μM) et PP_i (107.7 μM).

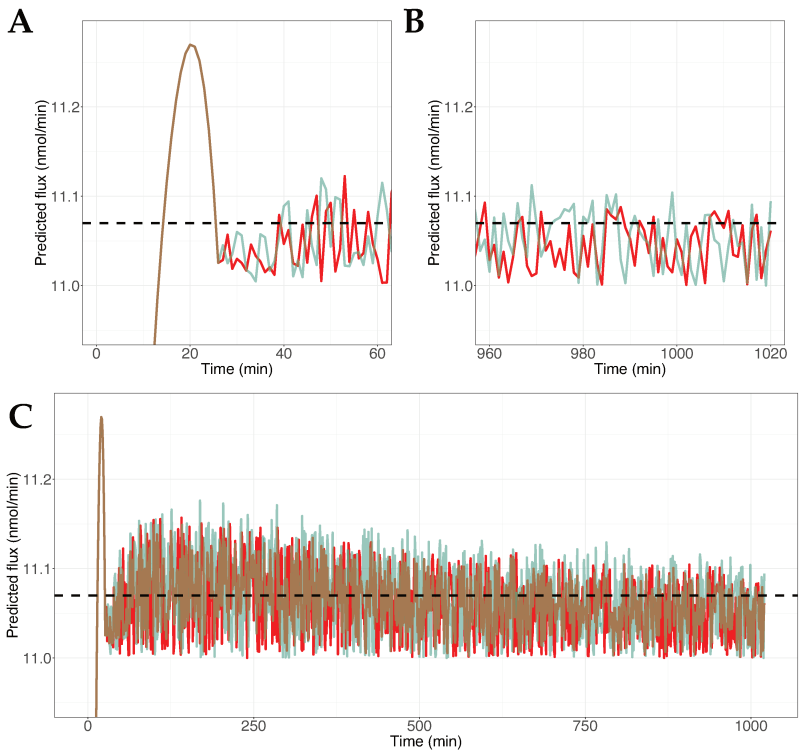

FIGURE 5.15 : *Prédictions du flux par le modèle contrôlé par les régulateurs PI pendant 17 h. Représentation du flux prédit pour la zone entre 10.9-11.3 nmol·min^{-1} pour la première heure* ***(A)****, la dernière heure* ***(C)*** *ou l'ensemble de la simulation* ***(C)****. Les régulateurs testés incluent les paramètres suivants :* K_{P_AMP}*=13/*K_{P_PPi}*=12 et* K_{I_AMP}*=1.3/*K_{I_PPi}*=1.2 pour le premier (courbe en rouge) et* K_{P_AMP}*=13/*K_{P_PPi}*=12 et* K_{I_AMP}*=2.6/*K_{I_PPi}*=2.4 pour le second (courbe en bleu). La ligne en pointillée noire représente la valeur consigne.*

Ces différents résultats nous indiquent que la composante Intégrale n'a pas d'impact majeur sur la régulation du flux en sortie de la voie de la glycolyse. Le régulateur P (K_{P_AMP}=13 et K_{P_PPi}=12) et le régulateur PI 1 sont globalement équivalents au niveau de leur performance sur 17 h de production. Nous notons une légère amélioration du contrôle du flux final avec le régulateur PI 1 (RMSE et dépassement plus faibles). Une simulation d'une plus longue durée serait nécessaire pour discriminer le meilleur modèle entre les deux.

Analysons toutefois les résultats obtenus par le régulateur PI 2, ayant pour paramétrage de départ, pour le terme Proportionnel, K_{P_AMP}=1 et K_{P_PPi}=1. La gamme de paramétrage du gain K_I

est plus petite (1-3) pour ce système de contrôle, en raison du temps de de calcul des sorties de notre modèle boîte-grise. Le meilleur système de régulation est obtenu avec K_{I_AMP}=K_{I_PPi}=1, avec un RMSE de ~0.408 nmol·min^{-1} (figure 5.16). L'erreur calculée pour ce nouveau régulateur est plus faible que celle calculée pour le régulateur intégrant uniquement le terme Proportionnel (RMSE≈0,536 nmol·min^{-1}).

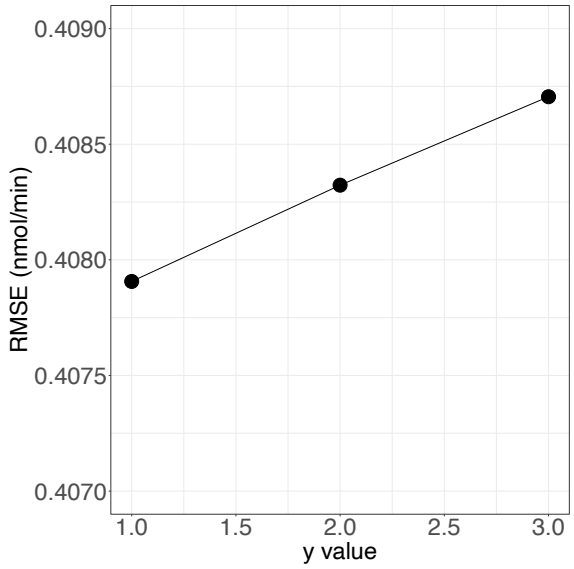

FIGURE 5.16 : *Variation du RMSE calculé pour les différents régulateurs PI, constitué d'un gain K_I dont y varie entre 1 et 3.*

Nous ne constatons aucun changement notable pour les concentrations en cosubstrats (figure 5.17), pour notre modèle intégrant le nouveau régulateur PI 2.

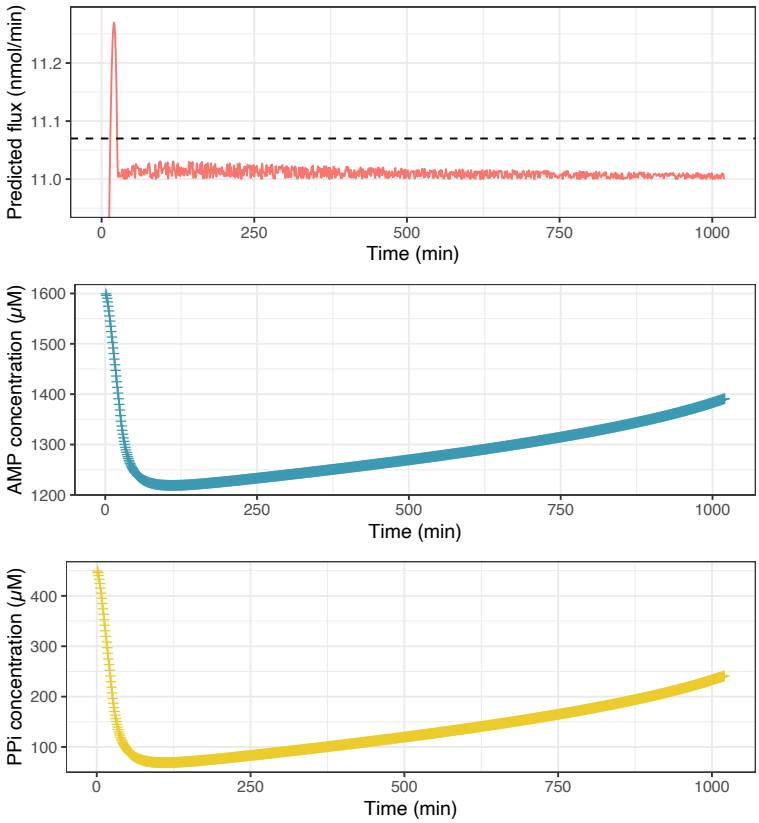

FIGURE 5.17 : *Prédictions du modèle contrôlé par le régulateur PI 2 (K_{P_AMP}=K_{P_PPi}=1 et K_{I_AMP}=K_{I_PPi}=1) pour 17h de simulation.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. La ligne en pointillée noire représente la valeur consigne.

En ce qui concerne le flux final de la voie de la glycolyse, nous observons des oscillations plus importantes lors de l'ajout de la composante Intégrale dans notre système de rétrocontrôle ; mais le système reste stable dans l'ensemble (figure 5.17). Quant à l'erreur statique, elle reste faible ($6{,}78.10^{-2}$ nmol·min^{-1}). Les autres critères de performance restent inchangés comparés à ceux obtenus avec le régulateur P (T_m=14.25 min et D=2.45%).

Nous pouvons conclure que le régulateur PI 1 améliore légèrement le processus de rétrocontrôle du flux final de la voie étudiée par rapport au régulateur P ; tandis que le régulateur PI 2 n'induit pas d'amélioration du contrôle du flux. Le paramétrage des gains effectué au sein des deux régulateurs présentés (régulateur P et PI) a été fait de manière manuelle. Cette méthode, basée sur un principe par essai-erreur, requiert beaucoup de temps, puisqu'elle doit faire plusieurs itérations pour chaque valeur de gain testée. Il existe d'autres méthodes de paramétrage, qu'il serait intéressant de tester (Bansal *et al.*, 2012). Parmi ces méthodes se trouve l'une des plus populaires : la méthode Ziegler Nichols (Ziegler and Nichols, 1993). Elle a été la première à être développée,

permettant l'ajustement rapide des paramètres des régulateurs de type PID. Elle se décline sommairement en deux approches dont l'une se base sur la réponse du modèle sans le régulateur et l'autre requiert d'amener le système de contrôle à la limite de sa stabilité. Ce type de méthode comporte des risques, notamment si notre système de production est sensible aux variations. D'autres méthodes sont proposées dans la revue de H. Bansal *et al.* (Bansal *et al.*, 2012) et sont les mêmes proposés par COPASI pour effectuer une estimation de paramètres (*e.g.*, évolution différentielle, algorithme génétique). De ce fait, nous pourrions utiliser COPASI pour estimer les paramètres de notre régulateur PID.

Le rétrocontrôle filtré de type Proportionnel-Intégral (ou régulateur PI filtré)

Le dernier type de régulateur testé sur notre modèle boîte-grise perturbé est un régulateur PI dont les commandes sont filtrées. Comme annoncé dans la partie Méthodes, le filtre sert à éviter une saturation précoce de la commande et à atténuer les oscillations autour de la valeur de consigne, autrement dit le bruit présent dans le système. Ce filtre est donc ajouté sur le régulateur PI 1 et sur le régulateur PI 2. La simulation est cette fois de 6 h. Le calcul du RMSE des différents régulateurs ajoutés au modèle donne des résultats similaires pour les deux systèmes de rétrocontrôle filtré (figure 5.18 et figure 5.19). Nous procédons quand même à l'identification du système de rétrocontrôle filtré ayant le meilleur RMSE pour le régulateur PI 1 et PI 2 (figure 5.18 B et figure 5.19 B). Les résultats les moins éloignés de la valeur consigne (11.07 nmol·min^{-1}) sont ceux obtenus par le régulateur PI 1 avec $\alpha = 0.3$ et ceux obtenus par le régulateur PI 2 avec $\alpha = 0.1$. Nous faisons l'hypothèse qu'une plus grande durée de simulation serait plus adaptée pour la discrimination du meilleur système de régulation pour cette voie métabolique.

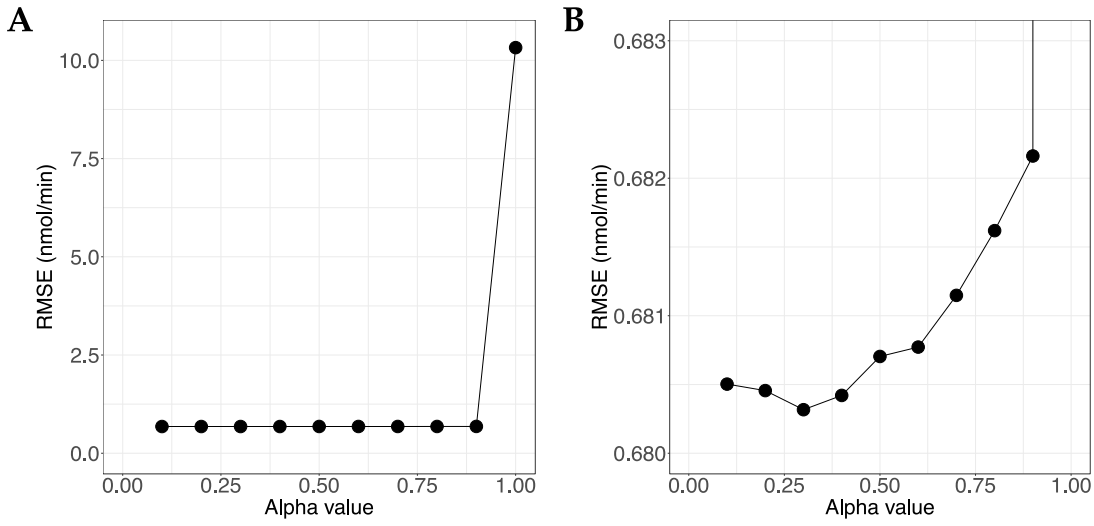

FIGURE 5.18 : *Variation du RMSE calculé pour le régulateur PI 1 (K_{P_AMP}=13, K_{P_PPi}=12 et K_{I_AMP}=1.3 et K_{I_PPi}=1.2) en fonction de la valeur du α (compris entre 0.1 et 1).*

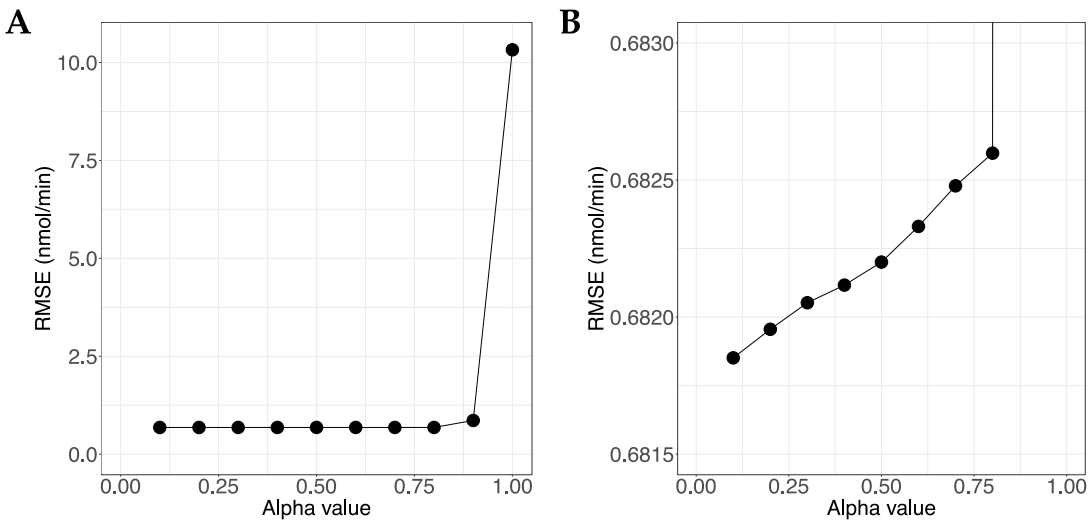

FIGURE 5.19 : *Variation du RMSE calculé pour le régulateur PI 2 (K_{P_AMP}=1, K_{P_PPi}=1 et K_{I_AMP}=K_{I_PPi}=1) en fonction de la valeur du α (compris entre 0.1 et 1).*

Nous analysons les résultats obtenus pour ces deux régulateurs filtrés et nous comparons leurs résultats aux mêmes régulateurs non filtrés. Le premier régulateur PI 1, avec $\alpha = 0.3$ résulte en un bon contrôle du flux en sortie de notre modèle (figure 5.20). Les concentrations de nos deux variables de commande sont similaires à celles obtenues lors du rétrocontrôle sans le filtre des moments ajouté (figure 5.20 A). Nous apercevons que l'ajout du filtre permet de diminuer les oscillations du flux final de la voie métabolique (figure 5.20 B). Dans notre étude, cette oscillation est faible, néanmoins dans le cas où le système de production serait sensible aux moindres variations, cela s'avérerait très important. Nous retrouvons ce cas de figure lors de l'utilisation de cellules de mammifère qui sont très sensibles aux changements (température, pH, osmolarité et agitation), pour la synthèse de produits issus de cultures de cellules de gènes de mammifère, à l'instar des cellules humaines (Werner and Noé, 1993).

En ce qui concerne les deux autres critères de performance, le temps de montée ne change pas (14.25 min) et le dépassement est de 1.78%.

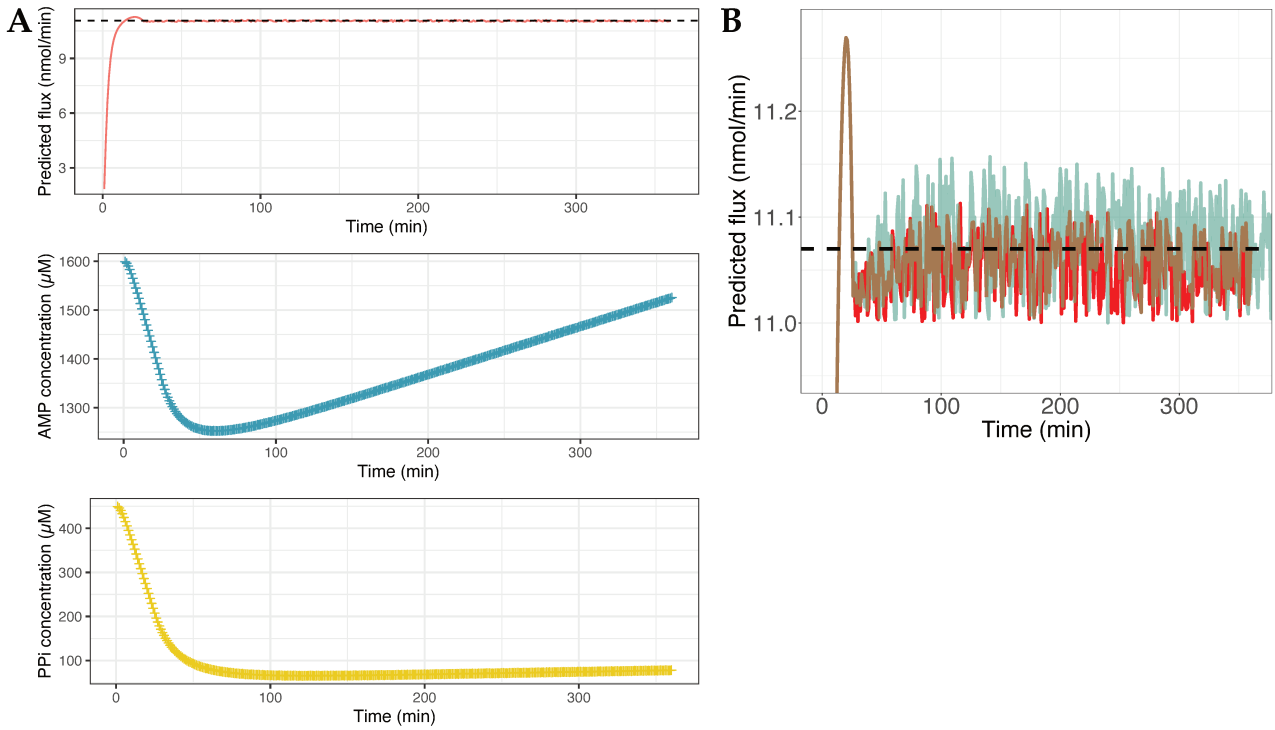

FIGURE 5.20 : *Rétrocontrôle du flux final de la voie de la glycolyse par le régulateur PI 1 filtré.*
***(A)** Prédictions du modèle contrôlé par le régulateur PI 1 avec $\alpha = 0.3$ pour 6 h de simulation. Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. **(B)** Agrandissement de la zone entre 11.0-11.2 nmol·min^{-1}, le contrôle du flux par le régulateur filtré est en rouge et celui avec le même régulateur non filtré est en bleu. La ligne en pointillée noire représente la valeur consigne.*

Le filtre des moments a également été ajouté sur le régulateur PI 2, avec le meilleur paramétrage obtenu avec $\alpha = 0.1$ (RMSE≈ 0.682 nmol·min^{-1}). Le rétrocontrôle obtenu est stable et assure le maintien du flux quasiment à la valeur consigne (~11 nmol·min^{-1}). Les oscillations présentes en présence du filtre sont similaires à celles obtenues sans filtre (figure 5.21 B). L'ajout du filtre ne semble pas avoir d'impact sur le contrôle du flux déjà effectué par le régulateur PI 2. Pour ce nouveau régulateur : le temps de montée reste à 14.25 min et le dépassement est de 2.38%.

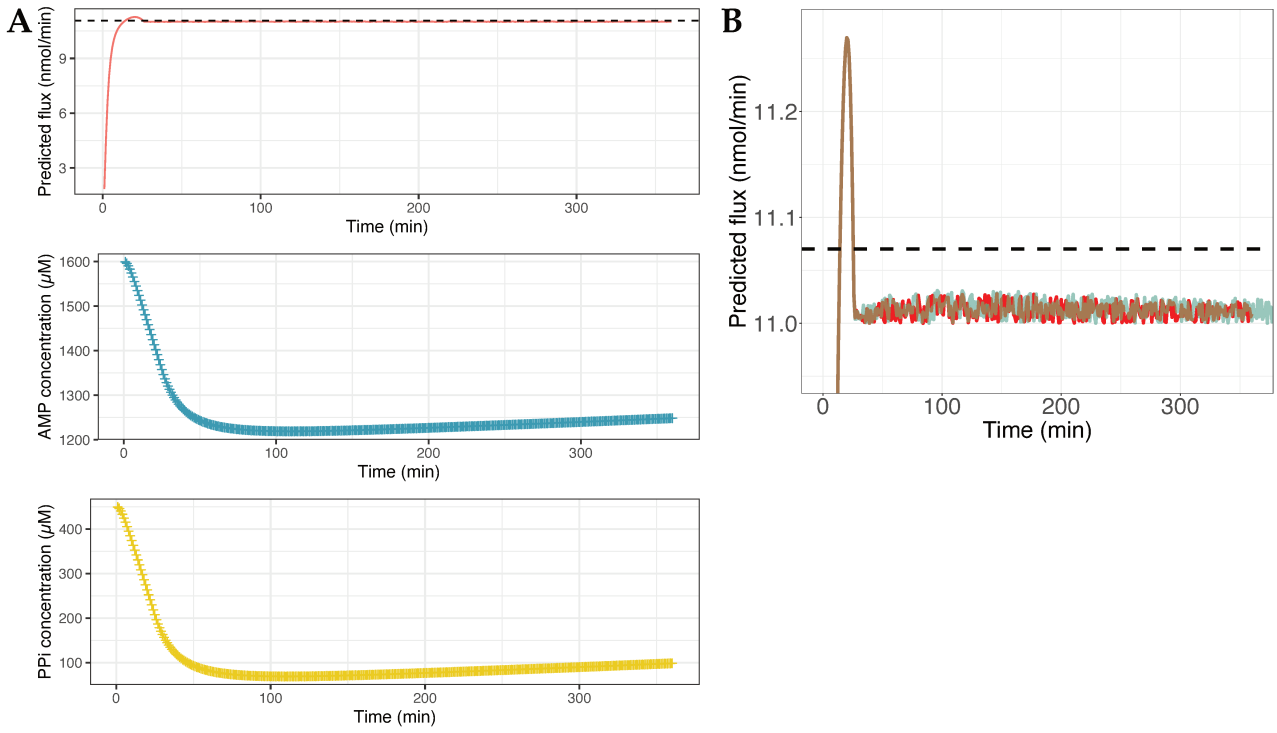

FIGURE 5.21 : *Rétrocontrôle du flux final de la voie de la glycolyse par le régulateur PI 2 filtré.*
(A) *Prédictions du modèle contrôlé par le régulateur PI 2 avec* $\alpha = 0.1$ *pour 6 h de simulation. Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en* PP_i. ***(B)*** *Agrandissement de la zone entre 11.0-11.2 nmol·min*$^{-1}$*, le contrôle du flux par le régulateur filtré est en rouge et celui avec le même régulateur non filtré est en bleu. La ligne en pointillée noire représente la valeur consigne.*

Nos deux systèmes de rétrocontrôle filtrés présentent de faibles oscillations du flux, lorsque nous regardons les variables de commandes, leurs concentrations augmentent continuellement de manière plus ou moins importante (figure 5.20 A et figure 5.21 A). Pour les concentrations qui semblent atteindre une valeur donnée (*e.g.*, PP_i pour le régulateur PI 1 filtré et AMP et PP_i pour le régulateur PI 2 filtré), le régulateur rajoute à chaque pas de temps une petite quantité de cosubstrat (~0.3 μM). Cela peut être considéré comme un inconvénient de notre système de rétrocontrôle. En effet, des changements minimes effectués tout au long d'une production peuvent mener à une usure prématurée des éléments du système de production, tels que les valves ou les sondes (McMillan, 2012).

5.3.3. Impact de l'ajout de perturbations supplémentaires sur le rétrocontrôle PI filtré

Afin d'évaluer la robustesse des systèmes de rétrocontrôle, nous ajoutons des perturbations supplémentaires au modèle boîte-grise. Ces perturbations consistent à ajouter le phénomène de dégradation des enzymes par le biais de l'ajout du temps de demi-vie des enzymes. Comme son nom l'indique, le temps de demi-vie d'une enzyme correspond au temps qui lui est nécessaire pour que sa concentration soit diminuée de moitié. Comme nous l'avons précisé plus tôt, cette perturbation est ajoutée uniquement sur la concentration en PGAM, car il s'agit de l'enzyme qui contrôle principalement la partie basse de la voie de la glycolyse. Le temps de demi-vie de cette enzyme chez *Entamoeba histolytica* n'a pas été mesuré, nous avons donc pris celle de *Archaeoglobus fulgidus* ($t_{1/2}$= 150 min) qui est également une enzyme indépendante du cofacteur 2,3-biphosphoglycérate (Johnsen and Schönheit, 2007). L'ajout de cette perturbation sur le modèle boîte-grise « perturbé » a très peu d'effet important sur le flux final de la voie et sur les concentrations en produit et métabolite (figure 5.22 et figure 5.6). Aussi l'effet de l'ajout de la perturbation est bien présent sur le flux en sortie de la réaction catalysée par PGAM qui passe d'un flux de ~9.1 nmol·min^{-1} à ~8.99 nmol·min^{-1} après 60 min de simulation (figure 5.6 B et figure 5.22 B). L'effet minime de cette perturbation peut être expliqué par le temps de demi-vie élevé de l'enzyme ($t_{1/2}$= 150 min).

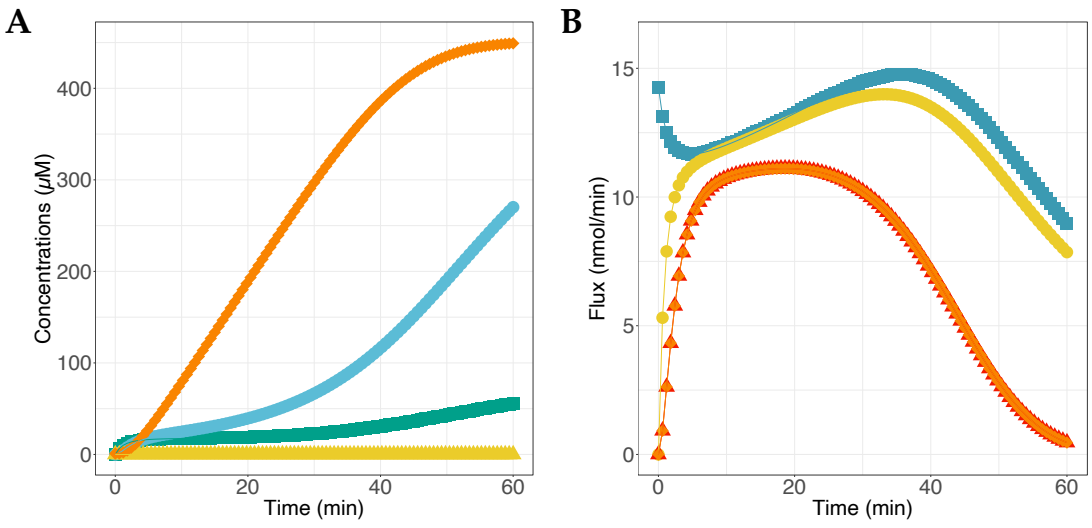

FIGURE 5.22 : *Prédictions des concentrations en métabolites et des flux par le modèle incluant la perturbation sur l'enzyme PGAM.*

(A) *Les concentrations des métabolites sont représentées en vert pour 2-PG (carrés), en bleu pour PEP (cercles), en jaune pour Pyr (triangles) et en orange pour l'Acétyl-CoA (losanges).* ***(B)*** *Les flux de chaque réaction de la voie sont représentés en bleu pour PGAM (carrés), en jaune pour ENO (cercles), en rouge pour PPDK (triangles) et en orange pour PFOR (losanges).*

Le premier régulateur testé est le régulateur PI 1 filtré que nous avons développé dans le paragraphe précédent. Nous menons une simulation de 17h et obtenons un RMSE plus élevé de ~8.58 nmol·min^{-1}. Les nouveaux résultats du système de rétrocontrôle sont représentés en figure 5.23. Nous constatons que le régulateur est bien fonctionnel et tente de maintenir au mieux le flux final à la valeur consigne. Le temps de montée est de 25 min, cette valeur est plus élevée que celle déterminée pour nos autres systèmes de rétrocontrôle. Le régulateur PI 1 fonctionne moins bien en présence de cette nouvelle perturbation ; ce qui remet en question sa robustesse. Le système de production se déroule bien jusqu'à 150 min, puis se décline pour atteindre une valeur de flux final ≈ 0 nmol·min^{-1} (figure 5.23). Le contrôle agit bien sur les deux variables de commande et nous voyons des oscillations de leur concentration au cours des 17 h de simulation (figure 5.23). Mais l'action sur ces variables de commande ne suffit plus après 150 min pour établir un flux optimal.

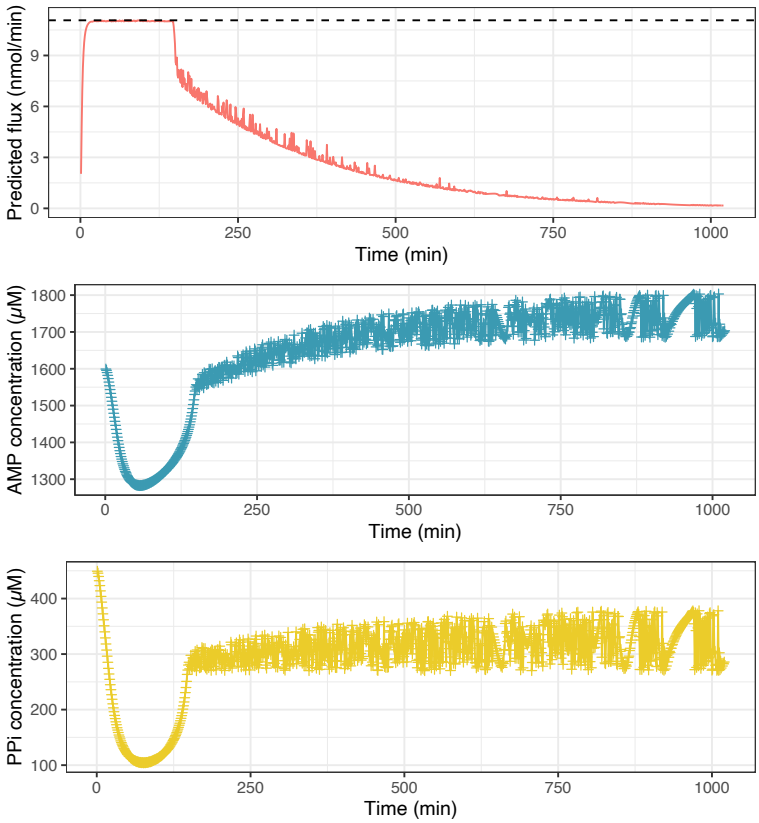

FIGURE 5.23 : *Prédictions du modèle perturbé contrôlé par le régulateur PI 1 (K_{P_AMP}= 13 et K_{P_PPi}=12 ; K_{I_AMP}=1.3 et K_{I_PPi}=1.2 et $\alpha = 0.3$) pour 17h de simulation.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. La ligne en pointillée noire représente la valeur consigne.

Nous obtenons des résultats similaires avec le régulateur PI 2 (figure 5.24). La variation des concentrations en AMP et PP_i nous indique que le système de rétrocontrôle est fonctionnel et qu'il

essaie de maintenir le flux final de la voie en agissant sur les deux variables de commande. Malgré l'action du régulateur PI 2, le flux final de la voie basse de la glycolyse chute après 150 min (figure 5.24). Il est à noter que les oscillations des concentrations en cosubstrat sont plus petites que celles obtenues lors du contrôle par le régulateur PI 1.

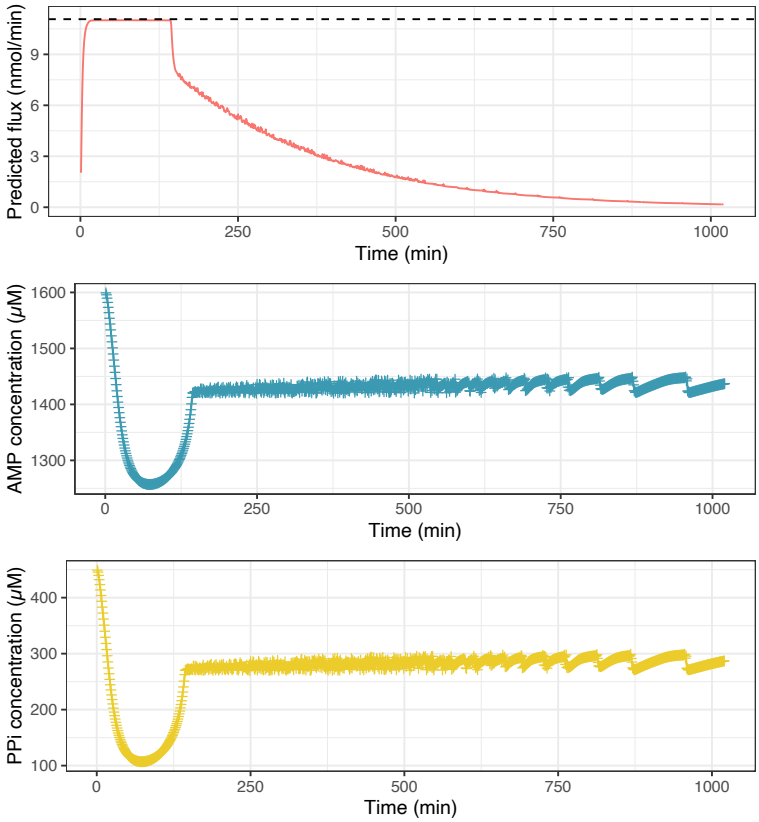

FIGURE 5.24 : *Prédictions du modèle perturbé contrôlé par le régulateur PI 2 (K_{P_AMP}= K_{P_PPi}=1 ; K_{I_AMP}= K_{I_PPi}=1 et $\alpha = 0.1$) pour 17h de simulation.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i. La ligne en pointillée noire représente la valeur consigne.

Nous savons qu'à partir de 150 min l'activité de PGAM va diminuer de moitié entraînant la diminution de la production totale ; l'enzyme PGAM catalysant la première réaction de la voie basse de la glycolyse. Cela explique donc la diminution du flux après 150 min. De plus, nous agissons uniquement sur les cosubstrats de la dernière réaction, catalysée par PPDK. Par conséquent, le régulateur PI (1 et 2) n'agit pas sur la réaction catalysée par l'enzyme PGAM, il n'y a pas de renouvellement ou d'apport en enzyme et la production décline complètement. Une amélioration ultérieure de ce système de rétrocontrôle consisterait à rajouter la concentration en enzyme en tant que variable de commande du régulateur. De plus, nous avons utilisé le temps de

demi-vie d'une enzyme provenant d'un autre organisme que celui étudié, nous ne pouvons donc conclure sur la robustesse réelle de notre régulateur PI.

Nous pouvons conclure de cette partie que le choix des variables de commandes est très important lors de la construction d'un système de rétrocontrôle. Aussi, la connaissance précise de paramètres des enzymes est essentielle pour l'obtention d'une fine modélisation d'un système de production régulé.

5.3.4. Comparaison des performances des différents modèles de rétrocontrôle

Après s'être attachés à la modélisation de plusieurs régulateurs implémentés sur notre modèle boîte-grise de la voie de la glycolyse, prenons le temps de comparer la performance de ces différents systèmes de rétrocontrôle. Pour ce faire, nous numérotons les différents systèmes de rétrocontrôle établis dans ce chapitre (tableau 5.3).

Controller	Control variable(s)	System
Proportional (P)	AMP: K_P=1	System 1
Proportional (P)	PP_i: K_P=1	System 2
Proportionnal (P)	AMP: K_P=13 PP_i: K_P=12	System 3
Proportionnal (P)	AMP: K_P=1 PP_i: K_P=1	System 4
Proportional-Integrative (PI)	AMP: K_P=13 and K_I=1.3 PP_i: K_P=12 and K_I=1.2	System 5
Proportional-Integrative (PI)	AMP: K_P=1 and K_I=1 PP_i: K_P=1 and K_I=1	System 6
Proportional-Integrative (PI) + filter	AMP: K_P=13 and K_I=1.3 PP_i: K_P=12 and K_I=1.2 $\alpha = 0.3$	System 7
Proportional-Integrative (PI) + filter	AMP: K_P=1 and K_I=1 PP_i: K_P=1 and K_I=1 $\alpha = 0.1$	System 8

TABLEAU 5.3 : Systèmes de rétrocontrôle élaborés dans ces travaux.

Ensuite les résultats des différents critères énoncés dans la partie Méthodes sont référencés dans le tableau 5.4 ci-dessous.

Controller	RMSE ($nmol \cdot min^{-1}$)	Rise time (min)	Overshoot (%)
System 1	10.83	14.25	2.45
System 2	10.15	14.25	2.45
System 3	0.533	14.25	1.71
System 4	0.536	14.25	2.45
System 5	0.405	14.25	1.89
System 6	0.408	14.25	2.45
System 7	0.405	14.25	1.99
System 8	0.682	14.25	2.36

TABLEAU 5.4 : Comparaison des critères de performance pour chaque système de rétrocontrôle bâti dans ce chapitre.
Les critères sont calculés pour une simulation de 17 h, excepté pour le système 8, où cela a été fait pour 6 h de simulation. Le dépassement a été calculé sur la partie stable du système de rétrocontrôle.

Les régulateurs ayant l'écart le plus faible entre le flux prédit et la valeur consigne sont ceux qui possèdent les composantes Proportionnelle et Intégrale, avec une préférence pour le système 5 qui possède la valeur la plus basse pour le RMSE et pour le dépassement de la consigne. Le temps de montée ne varie pas selon les régulateurs utilisés ; l'ajout de la composante Intégrale ne semble pas avoir l'effet escompté sur ce critère (temps de montée) qui doit normalement diminuer (Alargt and Ashur, 2013). Enfin, le dépassement le plus bas est obtenu avec le système 3 (régulateur P), puis le système 7 (régulateur PI 1 filtré).

À première vue, il semble que le régulateur P suffit pour maintenir le flux final de notre voie métabolique à la valeur consigne, puisqu'il présente un RMSE et un dépassement qui sont faibles. Nous observons que l'ajout de la composante principale a pour effet de réduire le RMSE. L'ajout du filtre des moments augmente le dépassement pour le système 7 et permet de réduire le dépassement pour le système 8. Nous pouvons considérer notre régulateur PI avec un filtre des moments comme un modèle prometteur qui présente un bon compromis entre l'erreur et le dépassement. De plus, nous avons vu dans la partie précédente que les oscillations étaient réduites avec l'utilisation de ce même filtre.

Par ailleurs, les bioréacteurs sont généralement conçus pour réaliser bien plus que 17 h de production de molécules. Ainsi, ils sont programmés pour une durée d'au moins une semaine voire des mois (Ozturk and Hu, 2006). Nous avons de ce fait réalisé une simulation d'une semaine (figure 5.25). L'analyse de ces résultats nous montre que le régulateur implémenté n'est pas stable, il s'arrête au bout de ~42 h. Suite à cela, il ne semble plus y avoir d'apport en PP_i au sein du système de production. Ce décrochage peut survenir à cause d'un problème numérique au sein du régulateur, de l'atteinte des limites de validation ou encore un problème de mémoire lors de la simulation. Ces quelques pistes de réflexion nous serviront de guide pour le développement d'un modèle optimal pour la régulation et le maintien du flux final de la voie métabolique à une valeur donnée.

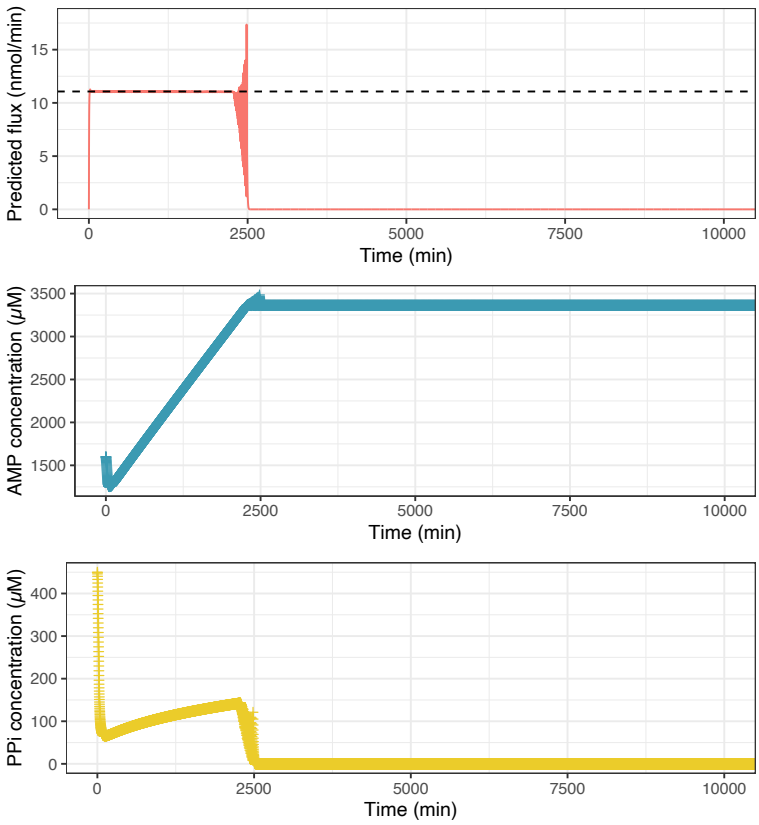

FIGURE 5.25 : *Rétrocontrôle du flux final de la voie de la glycolyse par le régulateur P (K_{P_AMP}=13 et K_{P_PPi}=12) pour une semaine.*
Panel du haut : prédiction du flux final de la voie basse de la glycolyse ; panel du milieu : prédiction de la concentration en AMP et panel du bas : prédiction de la concentration en PP_i.

5.4. Résultats et discussion

1. Conclusion

La production de molécules par le biais des voies métaboliques nécessite souvent l'ajout de système de régulation pour assurer une production optimale de la molécule d'intérêt. Ces systèmes de contrôle agissent sur différentes variables de la voie métabolique (paramètres physico-chimiques du milieu de production, substrat, cosubstrat, enzymes, produits) et sont déjà mis en place dans certaines productions (Cheng *et al.*, 2017; Tan *et al.*, 2016; Burgard *et al.*, 2016). Afin d'accélérer la mise en place de tel processus sur de réels systèmes de production, nous envisageons la modélisation d'un régulateur sur notre modèle de la voie basse de la glycolyse.

Dans notre étude, ce régulateur permettrait le rétrocontrôle de la voie étudiée, pour permettre le maintien du flux à une valeur optimale, en agissant sur les concentrations en cosubstrats. Le régulateur que nous avons choisi d'implémenter sur notre modèle de la voie de la glycolyse est le régulateur PID. Il s'agit d'un régulateur classique, robuste et facile à mettre en place sur des systèmes (Aguilar *et al.*, 2002; Sergei Mikhalevich *et al.*, 2015; Huyett *et al.*, 2015). Le modèle sur lequel sera implémenté ce système de rétrocontrôle est le modèle boîte-grise, lequel a démontré sa performance à prédire le flux en sortie de cette voie.

Plusieurs systèmes de rétrocontrôle ont été établis lors de ces travaux :

- Régulateurs P : agissant sur une ou deux variables de contrôle, à savoir la concentration en AMP et/ou PP_i;
- Régulateurs PI : agissant sur les deux variables de commande citées plus tôt;
- Régulateurs PI incluant un filtre des moments.

Les simulations ont été faites pour une période de 17 h de production, ce qui est relativement peu par rapport au temps de simulation usuels des bioréacteurs. Les résultats nous ont montré une bonne performance générale des régulateurs de type PID à maintenir le flux en sortie de la voie étudiée. Avec une distinction notable pour le régulateur P agissant sur les deux cosubstrats (AMP : K_P=13 et PP_i: K_P=12) qui offre la meilleure performance en termes de dépassement et présente un faible RMSE. Le régulateur PI agissant sur les deux cosubstrats et contenant le filtre des moments (AMP : K_P=13 et K_I=1.3 ; PP_i: K_P=12 et K_I=1.2 ; $\alpha = 0.3$) est un modèle prometteur, puisqu'il permet l'obtention d'une erreur faible entre les flux prédits et la valeur consigne et réduit les oscillations par rapport au régulateur PI sans filtre des moments.

L'addition de perturbations supplémentaires au modèle a révélé un point faible de notre système de rétrocontrôle : l'action sur les cosubstrats uniquement. Toutefois ces résultats peuvent être remis en question puisque le temps de demi-vie utilisé pour ces travaux était celui d'une enzyme provenant d'un organisme différent de notre parasite étudié. De plus, ce temps de demi-vie a été déterminé à une température très élevée (70°C) (Johnsen and Schönheit, 2007). Or, cette température qui n'est généralement pas atteinte dans un système de production, où la température est généralement à 30-37°C (Doran, 2013).

Ces travaux nous révèlent l'importance du choix des paramètres lors de la construction d'un système de rétrocontrôle comme celui-ci. De plus, plusieurs perspectives d'améliorations ont été identifiées, telles que : le test d'un nouveau set de paramètres par d'autres méthodes, l'ajout d'une variable de commande supplémentaire pour pallier d'autres perturbations pouvant survenir lors de la production ou encore l'étude plus approfondie de la synthèse à partir de 2500 min.

5. Discussion et conclusion du chapitre

Dans ce dernier chapitre, nous avons décrit et analysé l'implémentation de plusieurs systèmes de rétrocontrôle sur notre modèle boîte-grise de la partie basse de la glycolyse du parasite *E. histolytica*. Ces régulateurs ont été évalués quant à leur performance à maintenir le flux en sortie de cette voie à une valeur consigne donnée.

Le régulateur que nous avons implémenté au sein du modèle boîte-grise est un régulateur de type Proportionnel-Intégral-Dérivée. Il a été construit de bout en bout, avec l'ajout d'une seule composante à la fois, pour permettre le réglage manuel successif des gains de chaque composante. Une fois le système mis en place, une évaluation de sa performance a été effectuée à l'aide de trois critères :

- Le RMSE qui évalue l'erreur entre la valeur de flux prédite et la valeur consigne renseignée dans le régulateur. Plus cette valeur est basse, plus l'erreur calculée sera faible.
- Le temps de montée qui illustre la rapidité du système de rétrocontrôle à atteindre la valeur consigne. Plus le temps de montée sera faible, plus le régulateur sera réactif.
- Le dépassement, comme son nom l'indique, quantifie le dépassement de la valeur consigne par le système de régulateur. Lors de l'amélioration du système de régulation, nous cherchons à diminuer ce dépassement.

Par ailleurs, une perturbation supplémentaire a été ajoutée au modèle afin d'évaluer la robustesse des systèmes de rétrocontrôle développés : l'ajout de l'impact du temps de demi-vie de l'enzyme PGAM.

L'implémentation de ces systèmes de rétrocontrôle a permis, pour la toute première fois, la modélisation d'un modèle cinétique d'une voie métabolique dont le flux en sortie est contrôlé. En effet, le régulateur P, qui est notre meilleur régulateur permet bien de maintenir le flux final de la voie à ~11 $nmol \cdot min^{-1}$, au bout de 14 min. Ce temps est le meilleur temps de montée que nous ayons obtenu. Ceci est dû aux caractéristiques naturelles des enzymes présentes au sein de cette voie métabolique. En effet, si les enzymes dans des conditions optimales parviennent à obtenir un flux de 11.07 au bout de 14 min, l'ajout d'un système de régulation au sein de cette voie ne peut diminuer ce temps ; c'est ce qui se passe dans notre voie métabolique. Afin d'améliorer le temps de montée, il serait bon d'utiliser d'autres enzymes plus « rapides » ou de réaliser une modification génétique des enzymes importantes dans le contrôle du flux (PGAM et ENO). Le dépassement le

plus faible que nous ayons obtenu est celui obtenu avec le même régulateur P. Le deuxième meilleur modèle est le modèle PI intégrant un filtre des moments qui induit la réduction des oscillations présentes au niveau du flux et présente la plus faible valeur de RMSE.

Plusieurs améliorations peuvent être envisagées, dont l'utilisation de techniques de paramétrage des gains du régulateur PID. Malgré la présence d'oscillations, les régulateurs sont quasiment stables durant 17h de production. Comme nous l'avons énoncé plus tôt, les bioréacteurs fonctionnent pendant plus de 17h : ils sont paramétrés pour tenir des jours voire des mois. La simulation d'une production de 6 jours a été faite et a révélé un problème interne du modèle à maintenir le contrôle au-delà de 42 h. Plusieurs hypothèses ont été faites pour expliquer cet « arrêt » du système de rétrocontrôle (atteinte des limites du modèle, problème numérique…). L'amélioration du système de rétrocontrôle sera par la suite ciblée sur ces points.

L'addition de perturbations sur la concentration en enzyme (PGAM) a mis en avant les limites de notre système de rétrocontrôle. En effet, notre régulateur agit uniquement sur les concentrations en cosubstrats et ne peut donc réagir convenablement lorsque la concentration des enzymes diminuent. Toutefois, comme nous l'avons indiqué et souligné dans cette section, le temps de demi-vie utilisé était celui d'un autre organisme. Nous pouvons supposer que le temps de demi-vie de PPDK chez le parasite que nous étudions est plus élevé. Deux améliorations du système de production sont possibles pour le rendre plus robuste (Gerson *et al.*, 1988) :

- La mise en place d'une **culture en lot** (« Batch »), en fixant la durée de la culture à une valeur où les enzymes sont en concentration suffisante pour maintenir le flux au niveau optimal souhaité ou ;
- La mise en place d'une **culture en mode discontinu alimenté** (« Fed-batch »), en alimentant le système en cosubstrats et en enzymes.

Nous pouvons identifier de ces résultats quelques points clés dans la mise en place d'un modèle performant d'une voie métabolique dont le flux est contrôlé :

- Identification des perturbations retrouvées au sein d'un système de production à grande échelle (augmentation ou diminution de la température, épuisement des catalyseurs, substrats ou cosubstrats…) ;
- Sélection des variables de commande adéquates et palliant au mieux les perturbations ;
- Construction des systèmes de rétrocontrôle en implémentant une composante à la fois ;
- Paramétrage des gains (manuel ou à l'aide de techniques d'estimation de paramètres).

Nous avons vu dans ce chapitre qu'il était possible d'implémenter un système de rétrocontrôle au sein d'un modèle cinétique pour le maintien du flux en sortie d'une voie métabolique à une valeur optimale. Les régulateurs ont permis un contrôle précis du flux final en régulant les concentrations en cosubstrat (AMP et PP_i). Cette étude propose une nouvelle méthode de modélisation de régulateurs impliqués dans la régulation de voie métabolique utilisée dans des systèmes de production, basée sur l'utilisation d'un modèle hybride boîte-grise sur lequel a été implémenté un régulateur de type P ou PI. Suite à cette étude, un régulateur pourrait être mis en place sur la voie complète de la glycolyse, afin de mimer au mieux le système de production. Il serait également intéressant d'améliorer le régulateur et de lui ajouter la composante Dérivée pour déterminer son impact sur le rétrocontrôle ou encore de tester un autre système de rétrocontrôle basé, par exemple, sur l'utilisation des réseaux de neurones comme suggéré au **chapitre 1**.

Chapitre 6

Conclusion et perspectives

De nos jours, deux ressources sont devenues essentielles à chacun : le temps et l'argent. Cela est d'autant plus vrai pour certains acteurs économiques de notre société qui participent à la production de molécules utiles à la communauté. Production, qui dans beaucoup de cas, serait meilleure avec l'ajout de quelques améliorations des systèmes de synthèse.

C'est dans ce cadre-là que s'implantent nos travaux de recherche. Ainsi, les objectifs de cette thèse se résument en la modélisation et le rétrocontrôle des voies métaboliques utilisées à des fins de production de molécules. Cela consistait d'une part à établir plusieurs modèles de voies métaboliques en partant de diverses données de la voie étudiée et à identifier le meilleur modèle de prédiction du flux en sortie de la voie. Puis, d'autre part, il s'agissait de développer un système de rétrocontrôle au sein de la voie métabolique étudiée pour permettre le maintien du flux final à un niveau optimal. Plusieurs exemples d'application ont été développés, dont chacun revêt une importance particulière dans un système de production de molécule (cellulaire ou acellulaire) : i) la voie de la glycolyse du parasite *Entamoeba histolytica* (Moreno-Sánchez *et al.*, 2008), permettant la synthèse de molécule précurseur importante pour la production de diverses molécules d'intérêt ; ii) la voie de détoxification du peroxyde du parasite *Trypanosoma cruzi* (González-Chávez *et al.*, 2015), dont la mise en place dans des procédés de production pourrait aider à l'amélioration de la synthèse en luttant contre l'apparition d'espèces oxydantes et iii) la voie de la fermentation de la pénicilline chez le champignon *Penicillium chrysogenum*, dont les données sont issues d'un processus à l'échelle industrielle (Goldrick *et al.*, 2015). De plus, nos travaux revêtent un aspect biologique supplémentaire avec notamment la modélisation de deux voies métaboliques importantes dans la survie de parasites. En effet, pour la première voie, il s'agit de la voie principale de synthèse d'énergie du parasite, tandis que la seconde correspond à une voie essentielle dans le système de défense du parasite.

Ainsi, cette thèse s'est focalisée sur la modélisation de voies métaboliques tant par des méthodes classiques (Rapoport *et al.*, 1974; Wiechert and Noack, 2011), que par des méthodes peu

utilisées d'apprentissage automatique pour la prédiction du flux final d'une voie métabolique ou par des méthodes innovantes hybrides.

Aussi, afin de contribuer au développement de modèle de voie métabolique intégrant un système de rétrocontrôle permettant de maintenir le flux de sortie à un niveau optimal, nous nous sommes intéressés, au cours des différents chapitres, à :

- La modélisation des voies métaboliques par une méthode que nous avons surnommées : « boîte-blanche » pour les modèles cinétiques qui tiennent compte de la cinétique de chaque enzyme de la voie étudiée.
- La modélisation des voies métaboliques par des modèles de type « boîte-noire » pour les modèles basés sur l'utilisation de données (modèles d'apprentissage automatique), dont les mécanismes internes de fonctionnement ne sont pas toujours distincts et interprétables d'un point de vue biologique.
- La modélisation des voies métaboliques par une méthode innovante « boîte-grise » qui allie les deux avantages des modèles boîte-blanche et boîte-noire en utilisant un modèle cinétique de base sur lequel un terme d'ajustement, paramétré à l'aide de données expérimentales, a été ajouté.
- L'implémentation d'un système de rétrocontrôle, de type régulateur PID, sur le modèle « boîte-grise » afin de maintenir le flux en sortie de la voie métabolique à un niveau optimal souhaité par l'opérateur.

6.1. Modélisation de voie métabolique par des modèles « boîte-blanche »

Plusieurs approches de modélisation de voie métabolique existent et ont été recensées dans la première partie de cette étude (figure 1.4). Parmi ces modèles nous en avons relevé un en particulier qui est couramment utilisé lors de la modélisation de voie biochimiques : les modèles cinétiques. Ces modèles cinétiques sont des modèles qui décrivent de manière détaillée et précise le fonctionnement d'une voie métabolique. Ainsi une connaissance approfondie de la voie étudiée est requise pour le développement de tels modèles. Cette connaissance inclut : la description précise de la cinétique enzymatique, mais aussi la connaissance de plusieurs autres paramètres de la voie, tels que les concentrations initiales en substrats, cosubstrats et enzymes, la présence d'effecteurs (activateurs ou inhibiteurs) contrôlant la voie métabolique et les conditions physico-chimiques de la voie (pH, température).

Notre objectif étant de relater un modèle de voie métabolique pour la production de molécule, nous avons eu le choix de nous placer soit dans un milieu cellulaire ou dans un milieu acellulaire. Notre choix s'est alors porté sur la modélisation d'une culture acellulaire qui nous a paru plus simple à mettre en place. En effet, la modélisation d'un système cellulaire demande en plus des connaissances mentionnées plus haut, d'autres informations sur l'organisme utilisé pour la production : taux d'expression des gènes codant les enzymes participant dans la voie métabolique étudiée, identification des voies parallèles à la voie de production et de leur impact sur cette même voie, connaissance des systèmes de régulation pouvant parasiter le système de production.

Les fondements de la modélisation étant posés, nous avons développé nos différents modèles cinétiques pour deux exemples d'application : i) la partie basse de la glycolyse et ii) la voie de détoxification du peroxyde. Ces deux voies métaboliques se trouvent chez des parasites et sont intéressantes à mettre en place du fait de la faible utilisation de ces organismes dans des systèmes de production. Or leur utilisation pourrait s'avérer être un avantage, notamment en raison de leur capacité à s'adapter à des changements importants de leur environnement (Zilberstein and Shapira). Nos modèles ne sont pas les premiers à être développés en utilisant cette méthode de modélisation ; néanmoins leur développement a été nécessaire pour la création du modèle hybride suivant et pour l'implémentation du système de rétrocontrôle sur la voie étudiée par la suite.

Dans l'ensemble, la modélisation par les modèles cinétiques s'avère être une technique performante pour la représentation *in-silico* de voies métaboliques, ceci, à condition d'avoir les connaissances suffisantes sur la voie de production que l'on désire mettre en place.

6.2. Modélisation des voies métaboliques par des modèles « boîte-grise »

Notre modèle cinétique réalisé au cours de ces travaux s'est montré assez récalcitrant quant à la prédiction du flux final de la voie étudiée lors de la variation de l'activité des espèces enzymatiques présentes au sein de la voie. Nous avons donc intégré aux réactions cinétiques de certaines enzymes un terme d'ajustement comportant un paramètre paramétré selon les données expérimentales que nous avions en notre possession (Lo-Thong *et al.*, 2020). La forme de ce terme est dépendante des résultats obtenus lors de la prédiction du flux final de la voie. Pour reprendre l'exemple de la voie basse de la glycolyse, ce terme a été ajouté à l'équation cinétique de l'enzyme PPDK pour permettre à ce que le flux soit mieux prédit lorsque l'on fait varier l'activité de cette enzyme.

Ce nouveau modèle hybride, baptisé « boîte-grise », a été identifié comme étant le meilleur modèle cinétique développé pour nos voies métaboliques étudiées, il a été bâti sur l'utilisation : i) d'un modèle cinétique, boîte-blanche, développé dans la première partie de ces travaux ; ii) d'un terme d'ajustement, relevant des modèles boîte-noire en raison de l'utilisation des données expérimentales pour paramétrer ledit terme.

D'un point de vue biologique, ce terme d'ajustement permet d'appréhender ou autrement dit de représenter les systèmes de régulation propres aux enzymes et aux organismes de production utilisés. L'utilisation d'un tel terme d'ajustement peut être bénéfique lorsque : les paramètres cinétiques des enzymes ne sont pas encore définis, l'équation cinétique décrivant l'enzyme est trop complexe et /ou lorsque le mécanisme réactionnel de l'enzyme n'est pas bien déterminé. Par ailleurs, l'établissement du modèle cinétique précédent et de ce nouveau modèle hybride a confirmé l'identification des enzymes les plus importantes dans le contrôle du flux de la voie métabolique.

Ce nouveau modèle hybride boîte-grise construit dans ces travaux de recherche présente deux avantages majeurs pour la modélisation de voies métaboliques. Il permet d'obtenir un modèle cinétique d'une voie métabolique de production, quand bien même la connaissance de cette voie serait incomplète pour construire un modèle mécanistique « traditionnel ». Et il facilite la représentation de réactions enzymatiques complexes et pouvant comprendre des paramètres cinétiques non définis.

Notons bien que ce type de modélisation est rendu possible grâce à un ensemble de données expérimentales permettant le paramétrage du terme d'ajustement. De ce fait, la possession d'un

ensemble de données expérimentales de bonne qualité définit la seule condition qu'un prochain utilisateur devrait remplir pour bâtir un tel modèle.

6.3. Modélisation des voies métaboliques par des modèles « boîte-noire »

Notre étude s'est intéressée également à la modélisation des voies métaboliques par le biais de modèles que nous avons surnommés « boîte-noire ». Ces modèles se basent sur l'utilisation d'ensemble de données expérimentales, qui sont issues soit : d'une reconstitution *in-vitro* de la voie concernée ou d'enregistrement de données issues d'un bioréacteur. Par conséquent, il nous permet de nous affranchir des lourdes recherches bibliographiques qui incombent aux développeurs de modèles de voie métabolique basés sur la connaissance. L'appellation « boîte-noire » provient du fait que ces constructions sont définies par des mécanismes internes dont les combinaisons qui y sont effectuées entre les entrées et la sortie sont difficiles à appréhender par l'Homme. Aussi, il est important de noter que ces relations entre les données en entrée (*e.g.*, concentrations en enzymes) ne sont pas fondées sur l'utilisation d'équation ayant un sens « biologique », à l'instar des équations de Michaelis-Menten.

Nous avons donc développé plusieurs modèles des voies métaboliques étudiées à titre d'exemples en utilisant les ensembles de données expérimentales dont nous disposions. Ces premiers modèles (Réseaux de Neurones Artificiels) nous ont donné de bons résultats de prédiction du flux final de la voie. En revanche, ils présentaient également de faible capacité de généralisation comparés aux autres modèles construits auparavant. C'est pourquoi, nous nous sommes aidés des modèles hybrides boîte-grise pour procéder à ce que nous avons appelé : l'augmentation de données, c'est-à-dire à la génération d'un ensemble de données plus grand, par ces modèles, pour compléter l'ensemble de données expérimentales initial.

Parmi les modèles réalisés, ceux qui ont donné de meilleurs résultats sont les modèles de type non-linéaires. Aussi, deux techniques de modélisation se sont démarquées pour la prédiction du flux d'une voie métabolique : les modèles de forêts aléatoires (RF) et les modèles d'apprentissage d'ensemble (XGBoost). Cette meilleure performance des modèles non-linéaires par rapport aux modèles linéaires pour nos exemples d'application serait liée au degré de non linéarité des voies métaboliques étudiées. En effet, plus la voie possède une structure non-linéaire, plus un modèle d'apprentissage automatique non-linéaire parviendra à représenter cet aspect de la voie. Ce qui n'est pas le cas des modèles linéaires qui ne peuvent appréhender et modéliser les aspects non-linéaires d'une voie métabolique.

Cette étude est la première comparaison faite dans le domaine de modélisation de voies métaboliques en vue de la prédiction du flux. Elle a permis la mise au point d'un outil d'aide à la

décision utile à tout utilisateur qui se lancerait dans la représentation d'une voie métabolique de production à l'aide de ces modèles basés sur l'utilisation de données.

6.4. Régulation de la voie par des systèmes de rétrocontrôle

Les précédentes modélisations des voies métaboliques ont servi de base pour l'implémentation d'un système de rétrocontrôle capable de maintenir le flux de la voie étudiée à un niveau optimal fixé par l'opérateur. Nous avons utilisé, plus particulièrement, le modèle hybride boîte grise comme modèle initial. L'utilisation de ce modèle hybride est plus avantageuse dans notre cas, puisqu'il permet de mettre en place des leviers de contrôle précis pour notre système de rétrocontrôle. En effet, il rend possible la modulation de certaines variables, telles que les concentrations en substrat, cosubstrat et/ou enzyme. Nous avons implémenté ce système de rétrocontrôle sur notre modèle hybride représentant la partie basse de la voie de la glycolyse du parasite *Entamoeba histolytica*.

Parmi les systèmes de rétrocontrôle existant, notre choix s'est porté sur la construction d'un régulateur PID (pour Proportionnel-Intégral-Dérivé). Ce régulateur a la particularité de pouvoir être implémenté par morceau ; avec l'ajout d'une composante à la fois. Cela permet d'adapter le système de rétrocontrôle à la voie étudiée, et de construire étape par étape le régulateur. Le régulateur a donc pour objectif le maintien du flux de la voie à une valeur consigne enregistrée, en modulant les valeurs de variables de commande, tel que le montre son architecture présentée en figure 5.3. Alors que notre modèle hybride nous proposait diverses variables de commande, nous avons sélectionné deux d'entre elles : la concentration en AMP et la concentration en PP_i. Ces deux molécules sont les cosubstrats d'une seule et même réaction catalysée par l'enzyme PPDK et sont également des inhibiteurs des deux autres réactions enzymatiques présentes dans la partie basse de la glycolyse. Ainsi, nous avons fait l'hypothèse que le contrôle de leur concentration permettrait de réguler au mieux les trois réactions enzymatiques, ce qui résulterait en une régulation optimale.

Les différents systèmes de rétrocontrôle ont montré leur performance à maintenir le flux final de la voie à la valeur consigne et ce durant les 17 premières heures de production. Néanmoins, nous relevons une limite considérable quant à son utilisation immédiate pour la modélisation de voie métabolique contrôlée pour la production de molécules dans un bioréacteur. En effet, le régulateur créé a montré des faiblesses lors de la prédiction du flux de la voie métabolique étudiée pour un temps de simulation supérieur à 42 h, ce qui est peu quand nous considérons le paramétrage généralement utilisé par les bioréacteurs dans le domaine industriel. Néanmoins, bien qu'à ce stade nous n'ayons pas encore élucidé la cause de cette limite celle-ci semble

davantage se situer du côté des limites numériques du modèle cinétique que du côté du régulateur lui-même.

Malgré ces barrières qui se dressent encore pour l'application d'un tel procédé à une plus grande échelle industrielle, ce système de rétrocontrôle implémenté sur un modèle de voie métabolique constitue la première représentation d'un système de production dont le flux en sortie est contrôlé.

6.5. Perspectives

Les différents travaux initiés dans ces présents travaux de recherche ont permis le franchissement de certaines barrières existantes dans le cadre de la modélisation des voies métaboliques pour la production de molécules. Il n'en reste pas moins que de nouvelles limites s'imposent dans ce cadre de recherche et sont à dépasser. Voyons dans cette dernière partie les principales perspectives de recherche pouvant conduire au franchissement de ces nouvelles limites.

Lorsque nous nous intéressons à la voie de production du pyruvate, qui rappelons-le se situe au niveau de la voie de la glycolyse du parasite *Entamoeba histolytica*, la première limitation est la représentation partielle de la voie. En effet, si l'on considère son utilisation au niveau industriel, il est plus simple et préférable d'alimenter le système de production avec du glucose qu'avec du 3–PG. Une amélioration de ce système de production consisterait alors à rajouter la partie haute de la voie de la glycolyse au sein du modèle cinétique. Cela permettrait à la fois de considérer le système de production complet, d'identifier les enzymes qui influencent le flux de production dans la voie intégrale ainsi que les régulations pouvant modifier le flux. Cette nouvelle modélisation ouvrirait la porte pour l'ajout d'un régulateur adapté, cette fois, à une voie métabolique complète.

En ce qui concerne la voie de détoxification du peroxyde, des expériences supplémentaires pourraient être bénéfiques pour la construction d'un modèle cinétique plus réaliste. En effet, le modèle cinétique développé pour cette voie métabolique contient deux équations cinétiques qui ont été modifiées par l'ajout d'un terme d'ajustement. Toutefois, le modèle présente encore des limites de prédiction qui pourraient être résolues par l'utilisation de données expérimentales de bonne qualité. Cette voie métabolique a suscité notre intérêt puisqu'elle pourrait être développée de manière parallèle dans un système acellulaire afin de lutter contre l'apparition des espèces oxydantes qui ont la capacité d'altérer notre système de production. Aussi, une perspective intéressante de recherche concernant cette voie consisterait à implémenter un système de contrôle ayant pour objectif l'évaluation de l'ajout de perturbations sur le système, comme par exemple l'effet de la variation de température ou de pH qui peut survenir habituellement dans un bioréacteur.

Nous nous sommes également intéressés au système de production de la pénicilline par un organisme au sein d'un bioréacteur. Cette production nous sert d'exemple de modélisation d'une culture cellulaire à l'échelle industrielle. Nous avons remarqué que les modèles de ce système de production étaient performants pour la production de la concentration finale en pénicilline, mais

que ces résultats étaient moins bons que ceux obtenus pour les deux autres voies métaboliques étudiées. Il serait alors intéressant de conduire une modélisation de cette voie qui prend en compte non seulement des paramètres du bioréacteur, tel que nous l'avons fait dans cette étude, mais aussi les paramètres internes de la voie de fermentation de la pénicilline. Un tel modèle pourrait en effet servir à la mise en place d'un régulateur PID qui pourrait agir à la fois sur les paramètres du bioréacteur (pH, température, volume, taux d'oxygène…) et sur les paramètres de la voie métabolique (concentrations en substrats/cosubstrats/enzymes).

Cette étude, comme nous l'avons exprimé auparavant, a permis la mise en place d'une méthodologie pour la modélisation de voie métabolique pour la production de molécules selon les données de départ que possède un utilisateur lambda. Nous pouvons alors envisager la construction d'une seule et même voie métabolique d'intérêt industriel avec ces différentes techniques, et comparer chacune de ces méthodes avec l'aide d'un large ensemble de données expérimentales issues de procédés industriels.

Par ailleurs, concernant la modélisation d'un système de rétrocontrôle au sein d'une voie métabolique de production, beaucoup reste à faire. En effet, nous avons constaté que nos régulateurs étaient capables d'amener un système de production un état d'équilibre où le flux et constant, mais qu'ils n'étaient pas en mesure de répondre avec efficacité à certaines perturbations ajoutées au modèle, comme par exemple le phénomène de dégradation naturelle des enzymes. Des travaux supplémentaires sont alors nécessaires, d'une part pour assurer la bonne représentation de la perturbation, d'autre part pour identifier les bons leviers à utiliser lors du rétrocontrôle en vue de surmonter ces nouvelles perturbations. Parmi ces perturbations, il serait utile de considérer celles qui sont présentes au sein d'un bioréacteur telles que la déplétion en substrat/cosubstrat, les variations du pH et de la température, la dégradation des catalyseurs et la toxicité des produits ou des métabolites, si tel est le cas. L'addition de système de rétrocontrôle différent du régulateur PID peut être envisagée également ; à l'issue de laquelle une comparaison des résultats pourrait être effectuée pour déterminer la meilleure méthode à considérer lors de la mise en place d'un système de production contrôlé. Il va sans dire que des données issues d'un réel bioréacteur avec un système de contrôle serait un avantage pour l'amélioration des systèmes de contrôle que nous souhaitons mettre en place.

ANNEXE A (*APPENDIX A*)

Ensemble de données de la voie de la glycolyse chez *Entamoeba histolytica* représentant les flux pour différents profils d'activité enzymatique

Entamoeba histolytica glycolysis pathway dataset representing fluxes for different sets of enzyme activities

PGAM (mU)	ENO (mU)	PPDK (mU)	J_{obs} (nmol·min^{-1})
0	328.5	196.5	0
36.02	328.5	196.5	17.37
51.05	328.5	196.5	19.17
58.05	328.5	196.5	22.82
62.93	328.5	196.5	21.48
70	328.5	196.5	22.5
75.11	328.5	196.5	25.17
83.08	328.5	196.5	24.96
90.08	328.5	196.5	28.97
108.2	328.5	196.5	31.95
75	0	196.5	0
75	71.78	196.5	14.3
75	143.27	196.5	21.07
75	200.74	196.5	20.95
75	250.9	196.5	21.69
75	286.49	196.5	20.88
75	328.42	196.5	22.25
75	372.63	196.5	24.26
75	458.36	196.5	24.93
75	328.5	0	0
75	328.5	77.54	18.39
75	328.5	115.9	21.85
75	328.5	134.95	22.63
75	328.5	155.18	22.11
75	328.5	174.25	24.72
75	328.5	186.15	23.87
75	328.5	197.09	21.85
75	328.5	213.29	23.8
75	328.5	232.13	27.85

TABLE A.1

Measured pathway flux (J_{obs}) for different sets of enzyme activities (experimental dots). The experimental dots from Fig. 2 of Ref. (Moreno-Sánchez *et al.*, 2008) were digitized to obtain the data shown in the table. For each dataset, only one enzyme was varied and the other two were kept constant.

PGAM (mU)	ENO (mU)	PPDK (mU)	J (nmol·min^{-1})
0	328.5	196.5	0
0.93	328.5	196.5	0.50
1.36	328.5	196.5	0.38
3.21	328.5	196.5	1.40
5.11	328.5	196.5	2.42
7.01	328.5	196.5	3.43
8.91	328.5	196.5	4.42
10.82	328.5	196.5	5.35
12.72	328.5	196.5	6.26
14.62	328.5	196.5	7.12
16.52	328.5	196.5	7.95
18.42	328.5	196.5	8.75
20.32	328.5	196.5	9.51
22.22	328.5	196.5	10.26
24.12	328.5	196.5	11.02
26.02	328.5	196.5	11.80
27.92	328.5	196.5	12.49
29.82	328.5	196.5	13.20
31.72	328.5	196.5	13.91
33.62	328.5	196.5	14.55
35.52	328.5	196.5	15.15
37.43	328.5	196.5	15.74
39.33	328.5	196.5	16.37
41.23	328.5	196.5	16.90
43.13	328.5	196.5	17.45
45.03	328.5	196.5	18.02
46.93	328.5	196.5	18.57
48.83	328.5	196.5	19.10
50.73	328.5	196.5	19.59
52.63	328.5	196.5	20.03

54.53	328.5	196.5	20.61
56.43	328.5	196.5	21.12
58.33	328.5	196.5	21.68
60.23	328.5	196.5	22.07
62.13	328.5	196.5	22.53
64.04	328.5	196.5	22.97
65.94	328.5	196.5	23.34
67.84	328.5	196.5	23.78
69.74	328.5	196.5	24.21
71.64	328.5	196.5	24.64
73.54	328.5	196.5	25.00
75.44	328.5	196.5	25.44
77.34	328.5	196.5	25.88
79.24	328.5	196.5	26.10
81.14	328.5	196.5	26.44
83.04	328.5	196.5	26.75
84.94	328.5	196.5	27.14
86.84	328.5	196.5	27.46
88.75	328.5	196.5	27.78
90.65	328.5	196.5	28.15

TABLE A.2

Extract of 50 data from pathway flux (J) for different sets of enzyme activities (from fitting curves). The complete fitting curves from Fig. 2 of Ref. (Moreno-Sánchez et al., 2008) were digitized here to obtain the values shown in the table. For each dataset, only one enzyme was varied and the other two were kept constant. This dataset contains 184 data.

PGAM (mU)	ENO (mU)	PPDK (mU)	J_{pred} (nmol·min^{-1})
0	328.5	196.5	0
4.5	328.5	196.5	1.64543
6	328.5	196.5	2.18995
7.5	328.5	196.5	2.7325
9	328.5	196.5	3.27309
12	328.5	196.5	4.34838
13.5	328.5	196.5	4.88309
15	328.5	196.5	5.41585
16.5	328.5	196.5	5.94666
18	328.5	196.5	6.47553
21	328.5	196.5	7.52745
22.5	328.5	196.5	8.05052
24	328.5	196.5	8.57166
25.5	328.5	196.5	9.09086
27	328.5	196.5	9.60816
30	328.5	196.5	10.637
31.5	328.5	196.5	11.1486
33	328.5	196.5	11.6582
34.5	328.5	196.5	12.166
36	328.5	196.5	12.6718
39	328.5	196.5	13.6779
40.5	328.5	196.5	14.1781
42	328.5	196.5	14.6764
43.5	328.5	196.5	15.1729
45	328.5	196.5	15.6674
48	328.5	196.5	16.651
49.5	328.5	196.5	17.14
51	328.5	196.5	17.6271
52.5	328.5	196.5	18.1124
54	328.5	196.5	18.5959

57	328.5	196.5	19.5573
58.5	328.5	196.5	20.0352
60	328.5	196.5	20.5114
61.5	328.5	196.5	20.9857
63	328.5	196.5	21.4582
66	328.5	196.5	22.3977
67.5	328.5	196.5	22.8647
69	328.5	196.5	23.33
70.5	328.5	196.5	23.7935
72	328.5	196.5	24.2551
75	328.5	196.5	25.1731
76.5	328.5	196.5	25.6295
78	328.5	196.5	26.084
79.5	328.5	196.5	26.5368
81	328.5	196.5	26.9878
84	328.5	196.5	27.8846
85.5	328.5	196.5	28.3304
87	328.5	196.5	28.7744
88.5	328.5	196.5	29.2167
90	328.5	196.5	29.6572

TABLE A.3

Extract of 50 data from the generated dataset of enzyme activity ratios and their predicted pathway flux (J_{pred}) by COPASI model with the added adjustment term (with $\alpha = 3{,}09.10^{6}$). For each dataset, only one enzyme activity was varied and the other two were kept constant. This dataset contains 184 data.

PGAM (mU)	ENO (mU)	PPDK (mU)	J (nmol·min^{-1})
0	161.1	97.2	0
0.1	241.9	425.1	0.0333158882783883
0.2	328.5	196.5	0.0734970238095239
106.1	252	0.2	0.0770005036630037
0.5	92.6	67.1	0.0892887820512821
119.5	22.2	1.2	0.122123992673993
141.3	0.4	158.7	0.153096657509158
75	328.5	0.7	0.241716391941392
75	0.9	196.5	0.247131456043956
4.9	11.8	167.6	0.265629761904762
75	328.5	0.9	0.310547573260074
0.8	533.3	157.8	0.311017857142858
1.1	209.6	361.2	0.355145146520147
1.2	169.8	264.5	0.371701923076924
16.2	37.6	15.1	0.438894597069598
1.4	307.9	328.7	0.496580586080587
154.9	245.9	1.1	0.498368131868133
3.7	37.8	144.3	0.504131410256411
5.6	23	152.4	0.522016941391942
1.5	328.5	196.5	0.550367216117217
1.9	286.5	354.2	0.661469780219781
1.7	684.6	399.9	0.671381868131869
2.1	328.5	196.5	0.769954212454214
12.7	750.5	7.6	0.815059523809525
2.3	328.5	196.5	0.843079670329671
75	328.5	2.6	0.891486263736265
10.3	39.2	60.6	0.913054029304031
75	3.6	196.5	0.972346611721614
75	328.5	3.1	1.06093360805861

2.9	328.5	196.5	1.06224771062271
2.7	599.7	238	1.06376968864469
75	328.5	3.3	1.12853296703297
3.1	585.2	172.6	1.21915842490843
4	170	209.4	1.25570695970696
6.9	244.7	35.7	1.26750274725275
3.9	643	67.2	1.32660943223443
7.9	160.5	41.2	1.33644276556777
4	328.5	196.5	1.46322893772894
75	328.5	4.5	1.53195192307693
4.2	328.5	196.5	1.53602152014652
12.9	548.1	16.7	1.57998992673993
22.8	418.8	10.9	1.64716071428572
75	6.3	196.5	1.67390521978022
124.5	55.3	8.6	1.70038782051282
75	328.5	5.2	1.76556043956044
126.5	5.1	209	1.86419276556777
75	7.1	196.5	1.87734249084249
62.4	8.3	186.2	1.88727426739927
17.7	41.1	82.6	1.89106272893773
75	7.2	196.5	1.9026304945055

TABLE A.4

Extract of 50 data from the dataset of 2,000 simulated enzyme activity ratios and their corresponding pathway flux (*J*).

PGAM (mU)	ENO (mU)	PPDK (mU)	J (nmol·min^{-1})
0	0	0	0
0	0	25	0
0	0	50	0
0	0	75	0
0	0	100	0
0	0	125	0
0	0	150	0
0	0	175	0
0	0	200	0
0	0	225	0
25	25	25	0.798158
25	25	50	1.35803
25	25	75	1.74591
25	25	100	2.0156
25	25	125	2.20703
25	25	150	2.34741
25	25	175	2.45412
25	25	200	2.51681
25	25	225	2.45302
25	25	250	2.40359
25	25	275	2.36416
25	25	300	2.33197
25	25	325	2.30519
25	25	350	2.28257
25	25	375	2.26321
25	25	400	2.24645
25	25	425	2.23179
25	25	450	2.21889
25	25	475	2.20742
25	25	500	2.19717

36.0150375939873	328.5	196.5	17.3731799481551
50	25	25	1.43468
50	25	50	2.44244
50	25	75	3.13833
50	25	100	3.61907
50	25	125	3.9578
50	25	150	4.20468
50	25	175	4.39155
50	25	200	4.50127
50	25	225	4.39121
50	25	250	4.30611
50	25	275	4.23831
50	25	300	4.18302
50	25	325	4.13706
50	25	350	4.09823
50	25	375	4.06502
50	25	400	4.03627
50	25	425	4.01114
50	25	450	3.989
50	25	475	3.96933

TABLE A.5

Extract of 50 data from the Dataset 1 of experimental and simulated enzyme activity ratios and their corresponding pathway flux (J). This dataset contains 68,950 data.

ANNEXE B (*APPENDIX B*)

Ensemble de données de la voie de détoxification du peroxyde chez *Trypanosoma cruzi* représentant les flux pour différents profils d'activité enzymatique

Trypanosoma cruzi peroxide detoxification pathway dataset representing fluxes for different sets of enzyme activities

TryR (mU)	TXN (mU)	TXNPx (mU)	J (nmol·min^{-1})
25	47.6	112.6	2.16031
26	88	179	2.69536
28	51.7	141.6	2.87708
30	53.3	123	3.09102
31	88	179	3.82096
40	88	179	5.44658
40	94.7	205.6	5.63746
42	31.9	166.3	3.49036
45	88	179	6.14293
47	88	179	6.38520
52	88	179	6.91195
54	66.8	65.5	1.11232
54	88	179	7.09462
58	90.9	76.8	2.79701
58	88	179	7.41894
59	88	179	7.49233
60	98.3	152.7	7.40529
62	25.3	170.2	3.11290
63	88	179	7.75909
68	70.5	113.5	5.17005
74	7.9	123.6	0.86127
79	22.4	149.4	2.63108
81	94.1	93.3	4.73100
81	88	179	8.57953
82	83	148.5	7.48772
85	23.5	84.9	1.45118
89	88	179	8.81566
93	88	179	8.91458
94	88	179	8.93763
95	53.9	103.9	3.97088

100	88	179	9.06350
101	88	179	9.08259
102	88	179	9.10119
105	88	179	9.15419
107	88	179	9.18734
109	88	179	9.21885
110	88	179	9.23403
113	50.1	126.6	4.73165
118	101	155.8	9.36906
122	88	179	9.39003
125	45.3	192	5.74591
126	88	179	9.43285
127	88	179	9.44295
134	88	179	9.50740
137	88	179	9.53199
141	44.5	92.9	2.91306
145	88	179	9.58995
149	88	179	9.61524
152	73.5	186.4	8.58508
152	88	179	9.63279

TABLE B.1

Extract of 50 data from the Dataset 2 of experimental and simulated enzyme activity ratios and their corresponding pathway flux (J). This dataset contains 1,671 data.

ANNEXE C (*APPENDIX C*)

Ensemble de données du processus de fermentation de la pénicilline chez *Penicillium chrysogenum* dans un bioréacteur

Dataset of penicillin fermentation process in Penicillium chrysogenum in a bioreactor

Time (h)	Oil flow ($L \cdot h^{-1}$)	Aeration rate ($L \cdot h^{-1}$)	Vessel Weight (kg)	Carbon evolution rate ($g \cdot h^{-1}$)	Vessel Volume (L)	CO2 percent in off-gas (%)	[Pen] ($g \cdot L^{-1}$)
0.2	22	30	62574	0.034045	58479	0.089514	1.0178e-25
0.4	22	30	62585	0.038702	58487	0.10176	0.001
0.6	22	30	62598	0.04024	58495	0.1058	0.00099934
0.8	22	30	62607	0.041149	58499	0.10819	0.00099874
1	22	30	62613	0.041951	58501	0.1103	0.00099821
1.2	22	30	62617	0.042758	58500	0.11242	0.00099771
1.4	22	30	62620	0.043599	58498	0.11463	0.00099722
1.6	22	30	62622	0.044478	58497	0.11694	0.00099674
1.8	22	30	62624	0.045398	58494	0.11936	0.00099627
2	22	30	62626	0.046361	58491	0.12189	0.0009958
2.2	22	30	62629	0.047367	58490	0.12453	0.0009953
2.4	22	30	62634	0.048418	58491	0.12729	0.00099477
2.6	22	30	62642	0.049515	58494	0.13018	0.00099421
2.8	22	30	62650	0.050658	58497	0.13318	0.00099363
3	22	30	62657	0.05185	58500	0.13631	0.00099306
3.2	22	30	62665	0.053091	58503	0.13957	0.00099251
3.4	22	30	62670	0.054382	58503	0.14297	0.00099198
3.6	22	30	62675	0.055722	58503	0.14649	0.00099147
3.8	22	30	62681	0.057114	58504	0.15015	0.00099095
4	22	30	62688	0.058562	58506	0.15395	0.00099039
4.2	30	30	62698	0.060065	58510	0.1579	0.00098981

4.4	30	30	62709	0.061618	58515	0.16198	0.00098922
4.6	30	30	62720	0.063224	58519	0.1662	0.00098863
4.8	30	30	62729	0.064885	58522	0.17057	0.00098806
5	30	30	62737	0.066604	58524	0.17508	0.00098753
5.2	30	30	62743	0.068385	58523	0.17976	0.00098702
5.4	30	30	62749	0.070228	58523	0.1846	0.00098651
5.6	30	30	62757	0.07213	58525	0.1896	0.00098597
5.8	30	30	62767	0.074088	58528	0.19475	0.000985 4
6	30	30	62778	0.076105	58533	0.20005	0.00098481
6.2	30	30	62789	0.078189	58537	0.20552	0.00098423
6.4	30	30	62798	0.080341	58540	0.21118	0.00098366
6.6	30	30	62806	0.082563	58542	0.21701	0.00098312
6.8	30	30	62814	0.084854	58543	0.22303	0.00098259
7	30	30	62821	0.087217	58544	0.22924	0.00098207
7.2	30	30	62829	0.089656	58546	0.23564	0.00098153
7.4	30	30	62839	0.092166	58549	0.24224	0.00098096
7.6	30	30	62850	0.094753	58554	0.24903	0.00098037
7.8	30	30	62861	0.097418	58559	0.25603	0.00097977
8	30	30	62872	0.10016	58564	0.26324	0.00097918
8.2	30	42	62883	0.11245	58568	0.21112	0.00097861
8.4	30	42	62892	0.11193	58571	0.21015	0.00097805
8.6	30	42	62900	0.11456	58573	0.21507	0.00097751
8.8	30	42	62909	0.11765	58576	0.22089	0.00097695

9	30	42	62919	0.1209	58580	0.22698	0.00097638
9.2	30	42	62931	0.12427	58585	0.2333	0.00097579
9.4	30	42	62943	0.12773	58591	0.23979	0.00097518
9.6	30	42	62957	0.13129	58598	0.24647	0.00097454
9.8	30	42	62971	0.13495	58606	0.25333	0.00097391
10	30	42	62985	0.13871	58613	0.2604	0.00097327

TABLE C.1

Extract of 50 data from the Dataset 3 of experimental recordings of the process of penicillin fermentation in a bioreactor. This dataset contains 113,935 data.

Liste des tableaux

Liste des figures

Bibliographie

Aguilar,R., Poznyak,A., Martínez-Guerra,R., and Maya-Yescas,R. (2002) Temperature control in catalytic cracking reactors via a robust PID controller. *Journal of Process Control*, **12**, 695–705.

Ajjolli Nagaraja,A., Fontaine,N., Delsaut,M., Charton,P., Damour,C., Offmann,B., Grondin-Perez,B., and Cadet,F. (2019) Flux prediction using artificial neural network (ANN) for the upper part of glycolysis. *PLoS ONE*, **14**, e0216178.

Al Imran,A., Rahman,A., Kabir,H., and Rahim,S. (2018) The Impact of Feature Selection Techniques on the Performance of Predicting Parkinson's Disease. *IJITCS*, **10**, 14–29.

Alargt,F.S. and Ashur,A.S. (2013) Analysis and Simulation of Interleaved Boost Converter for Automotive applications. **2**, 9.

Albracht,S.P.J. (1980) The prosthetic groups in Succinate Dehydrogenase number and stoichiometry. **612**, 11–28.

Ali,V. and Nozaki,T. (2007) Current Therapeutics, Their Problems, and Sulfur-Containing-Amino-Acid Metabolism as a Novel Target against Infections by "Amitochondriate" Protozoan Parasites. *CMR*, **20**, 164–187.

Alkim,C., Cam,Y., Trichez,D., Auriol,C., Spina,L., Vax,A., Bartolo,F., Besse,P., François,J.M., and Walther,T. (2015) Optimization of ethylene glycol production from (d)-xylose via a synthetic pathway implemented in Escherichia coli. *Microb Cell Fact*, **14**, 127.

Almquist,J., Cvijovic,M., Hatzimanikatis,V., Nielsen,J., and Jirstrand,M. (2014) Kinetic models in industrial biotechnology – Improving cell factory performance. *Metabolic Engineering*, **24**, 38–60.

Alves,R., Vilaprinyo,E., Hernández-Bermejo,B., and Sorribas,A. (2008) Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways. *Biotechnology and Genetic Engineering Reviews*, **25**, 1–40.

Alzghoul,A., Backe,B., Löfstrand,M., Byström,A., and Liljedahl,B. (2014) Comparing a knowledge-based and a data-driven method in querying data streams for system fault detection: A hydraulic drive system application. *Computers in Industry*, **65**, 1126–1135.

Ambrosen,K.S., Skjerbæk,M.W., Foldager,J., Axelsen,M.C., Bak,N., Arvastson,L., Christensen,S.R., Johansen,L.B., Raghava,J.M., Oranje,B., *et al.* (2020) A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data. *Transl Psychiatry*, **10**, 276.

Amri,Y., Fadhilah,A.L., Fatmawati, Setiani,N., and Rani,S. (2016) Analysis Clustering of Electricity Usage Profile Using K-Means Algorithm. *IOP Conf. Ser.: Mater. Sci. Eng.*, **105**, 012020.

Angermueller,C., Pärnamaa,T., Parts,L., and Stegle,O. (2016) Deep learning for computational biology. *Mol Syst Biol*, **12**, 878.

Antoniewicz,M., Kraynie,D., Laffend,L., Gonzalezlergier,J., Kelleher,J., and Stephanopoulos,G. (2007) Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of E. coli producing 1,3-propanediol. *Metabolic Engineering*, **9**, 277–292.

Antoniewicz,M.R. (2015) Methods and advances in metabolic flux analysis: a mini-review. *J Ind Microbiol Biotechnol*, 9.

Antoniewicz,M.R., Stephanopoulos,G., and Kelleher,J.K. (2006) Evaluation of regression models in metabolic physiology: predicting fluxes from isotopic data without knowledge of the pathway. *Metabolomics*, **2**, 41–52.

Arabzadeh,V., Sohrabi,M.R., Goudarzi,N., and Davallo,M. (2019) Using artificial neural network and multivariate calibration methods for simultaneous spectrophotometric analysis of Emtricitabine and Tenofovir alafenamide fumarate in pharmaceutical formulation of HIV drug. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **215**, 266–275.

Arendt,K., Jradi,M., Shaker,H.R., and Veje,C.T. (2018) COMPARATIVE ANALYSIS OF WHITE-, GRAY- AND BLACK-BOX MODELS FOR THERMAL SIMULATION OF INDOOR ENVIRONMENT: TEACHING BUILDING CASE STUDY. 8.

Aslam,F. and Kaur,G. (2011) Comparative Analysis of Conventional, P, PI, PID and Fuzzy Logic Controllers for the Efficient Control of Concentration in CSTR. *IJCA*, **17**, 12–16.

Awan,S.E., Bennamoun,M., Sohel,F., Sanfilippo,F.M., Chow,B.J., and Dwivedi,G. (2019) Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PLoS ONE*, **14**, e0218760.

Aydogdu,M., Bagirova,M., Allahverdiyev,A., Abamor,E.S., Ozyilmaz,O.A., Dinparvar,S., and Kocagoz,T. (2019) Large-scale cultivation of Leishmania infantum promastigotes in stirred bioreactor. *J Vector Borne Dis*, 6.

Azodi,C.B., Tang,J., and Shiu,S.-H. (2020) Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics*, **36**, 442–455.

Babajide Mustapha,I. and Saeed,F. (2016) Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules*, **21**, 983.

Banas,K., Banas,A.M., Gajda,M., Kwiatek,W.M., Pawlicki,B., and Breese,M.B.H. (2013) Analysis of synchrotron radiation induced X-ray emission spectra with R environment. *Radiation Physics and Chemistry*, **93**, 82–86.

Bansal,D., Sehgal,R., Chawla,Y., Mahajan,R.C., and Malla,N. (2004) In-vitro activity of antiamoebic drugs against clinical isolates of Entamoeba histolytica and Entamoeba dispar. *Annals of Clinical Microbiology and Antimicrobials*, 5.

Bansal,H.O., Sharma,R., and Shreeraman,P.R. (2012) PID Controller Tuning Techniques: A Review. *Journal of Control Engineering and Technology*, **2**, 10.

Baranowska,B., Terlecki,G., and Baranowski,T. (1984) The influence of inorganic phosphate and ATP on the kinetics of bovine heart muscle pyruvate kinase. *Mol Cell Biochem*, **64**, 45–50.

Baranwal,M., Magner,A., Elvati,P., Saldinger,J., Violi,A., and Hero,A.O. (2020) A deep learning architecture for metabolic pathway prediction. *Bioinformatics*, **36**, 2547–2553.

Barcroft,J. and Hill,A.V. (1910) The nature of oxyhaemoglobin, with a note on its molecular weight. *The Journal of Physiology*, **39**, 411–428.

Basheer,I.A. and Hajmeer,M. (2000) Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, **43**, 3–31.

Battiti,R. and Masulli,F. (1990) BFGS optimization for faster and automated supervised learning. In, *International neural network conference*. Springer, pp. 757–760.

Bei,H., Wang,L., Sun,J., and Zhang,L. (2019) A Multistage Feedback Control Strategy for Producing 1,3-Propanediol in Microbial Continuous Fermentation. *Complexity*, **2019**, 1–9.

Bideaux,C., Montheard,J., Cameleyre,X., Molina-Jouve,C., and Alfenore,S. (2016) Metabolic flux analysis model for optimizing xylose conversion into ethanol by the natural C5-fermenting yeast Candida shehatae. *Appl Microbiol Biotechnol*, **100**, 1489–1499.

Boran,E., Özgür,E., Yücel,M., Gündüz,U., and Eroglu,I. (2012) Biohydrogen production by Rhodobacter capsulatus in solar tubular photobioreactor on thick juice dark fermenter effluent. *Journal of Cleaner Production*, **31**, 150–157.

Botstein,D. and Fink,G. (1988) Yeast: an experimental organism for modern biology. *Science*, **240**, 1439–1443.

Bowie,J.U., Sherkhanov,S., Korman,T.P., Valliere,M.A., Opgenorth,P.H., and Liu,H. (2020) Synthetic Biochemistry: The Bio-inspired Cell-Free Approach to Commodity Chemical Production. *Trends in Biotechnology*, **38**, 766–778.

Braun,J. and Chaturvedi,N. (2002) An Inverse Gray-Box Model for Transient Building Load Prediction. *HVAC&R Res.*, **8**, 73–99.

Breiman,L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Brougham,D.F., Ivanova,G., Gottschalk,M., Collins,D.M., Eustace,A.J., O'Connor,R., and Havel,J. (2011) Artificial Neural Networks for Classification in Metabolomic Studies of Whole Cells Using ^{1}H Nuclear Magnetic Resonance. *Journal of Biomedicine and Biotechnology*, **2011**, 1–8.

Buchner,E. and Rapp,R. (1897) Alkoholische Gährung ohne Hefezellen. *Ber. Dtsch. Chem. Ges.*, **30**, 2668–2678.

Bucz,Š. and Kozáková,A. (2018) Advanced Methods of PID Controller Tuning for Specified Performance. In, Shamsuzzoha,M. (ed), *PID Control for Industrial Processes*. InTech.

Burbidge,R., Trotter,M., Buxton,B., and Holden,S. (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, **26**, 5–14.

Burgard,A., Burk,M.J., Osterhout,R., Van Dien,S., and Yim,H. (2016) Development of a commercial scale process for production of 1,4-butanediol from sugar. *Current Opinion in Biotechnology*, **42**, 118–125.

Cakit,E., Durgun,B., and Cetik,O. (2015) A Neural Network Approach for Assessing the Relationship between Grip Strength and Hand Anthropometry. *Neural Network World*, **25**, 603–622.

Calli,B., Schoenmaekers,K., Vanbroekhoven,K., and Diels,L. (2008) Dark fermentative H2H2 production from xylose and lactose—Effects of on-line pH control. *International Journal of Hydrogen Energy*, **33**, 522–530.

Camacho,D.M., Collins,K.M., Powers,R.K., Costello,J.C., and Collins,J.J. (2018) Next-Generation Machine Learning for Biological Networks. *Cell*, **173**, 1581–1592.

Cascante,M., Boros,L.G., Comin-Anduix,B., de Atauri,P., Centelles,J.J., and Lee,P.W.-N. (2002) Metabolic control analysis in drug discovery and disease. *Nat Biotechnol*, **20**, 243–249.

Chai,T. and Draxler,R.R. (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, **7**, 1247–1250.

Chance,Britton (1943) The kinetics of the enzyme-substrate compound of peroxidase. *Journal of Biological Chemistry*, **151**, 553–577.

Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen,Z., He,N., Huang,Y., Qin,W.T., Liu,X., and Li,L. (2018) Integration of A Deep Learning Classifier with A Random Forest Approach for Predicting Malonylation Sites. *Genomics, Proteomics & Bioinformatics*, **16**, 451–459.

Cheng,F., Luozhong,S., Guo,Z., Yu,H., and Stephanopoulos,G. (2017) Enhanced Biosynthesis of Hyaluronic Acid Using Engineered *Corynebacterium glutamicum* Via Metabolic Pathway Regulation. *Biotechnol. J.*, **12**, 1700191.

Chevalier,M., Gomez-Schiavon,M., Ng,A., and El-Samad,H. (2018) Design and analysis of a Proportional-Integral-Derivative controller with biological molecules. 35.

Chicco,D. (2017) Ten quick tips for machine learning in computational biology. *BioData Mining*, **10**, 35.

Chong,J. and Xia,J. (2017) Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. *Metabolites*, **7**, 62.

Church,G.M. and Regis,E. (2012) Regenesis: how synthetic biology will reinvent nature and ourselves Basic Books, New York.

Cifuentes,J., Marulanda,G., Bello,A., and Reneses,J. (2020) Air Temperature Forecasting Using Machine Learning Techniques: A Review. *Energies*, **13**, 4215.

Cornish-Bowden,A. ed. (1997) New beer in an old bottle: Eduard Buchner and the growth of biochemical knowledge Universitat de Valencia, Valencia.

Costa,R.S., Hartmann,A., and Vinga,S. (2016) Kinetic modeling of cell metabolism for microbial production. *Journal of Biotechnology*, **219**, 126–141.

Costello,Z. and Martin,H.G. (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Syst Biol Appl*, **4**, 19.

Cotten,C. and Reed,J.L. (2013) Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC Bioinformatics*, **14**, 32.

Culley,C., Vijayakumar,S., Zampieri,G., and Angione,C. (2020) A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc Natl Acad Sci USA*, **117**, 18869–18879.

Cuperlovic-Culf,M. (2018) Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. *Metabolites*, **8**, 4.

Curto,R., O. Voit,E., Sorribas,A., and Cascante,M. (1998) Mathematical models of purine metabolism in man. *Mathematical Biosciences*, **151**, 1–49.

Curto,R., Voit,E.O., Sorribas,A., and Cascante,M. (1997) Validation and steady-state analysis of a power-law model of purine metabolism in man. *Biochemical Journal*, **324**, 761–775.

Dahl,G.E., Jaitly,N., and Salakhutdinov,R. (2014) Multi-task Neural Networks for QSAR Predictions. *arXiv:1406.1231 [cs, stat]*.

DiStefano,J.J. (2013) Dynamic systems biology modeling and simulation First edition. Elsevier, Academic Press, Amsterdam.

Doran,P.M. (2013) Heat Transfer. In, *Bioprocess Engineering Principles*. Elsevier, pp. 333–377.

Dorronsoro,I., Chana,A., Abasolo,M.I., Castro,A., Gil,C., Stud,M., and Martinez,A. (2004) CODES/Neural Network Model: a Useful Tool for in Silico Prediction of Oral Absorption and Blood-Brain Barrier Permeability of Structurally Diverse Drugs. *QSAR & Combinatorial Science*, **23**, 89–98.

Drysch,A., El Massaoudi,M., Mack,C., Takors,R., de Graaf,A.A., and Sahm,H. (2003) Production process monitoring by serial mapping of microbial carbon flux distributions using a novel Sensor Reactor approach: II—13C-labeling-based metabolic flux analysis and l-lysine production. *Metabolic Engineering*, **5**, 96–107.

Duchêne,M. (2015) Metronidazole and the Redox Biochemistry of *Entamoeba histolytica*. In, Nozaki,T. and Bhattacharya,A. (eds), *Amebiasis*. Springer Japan, Tokyo, pp. 523–541.

Dutta,N. and Saha,M.K. (2018) Immobilization of a Mesophilic Lipase on Graphene Oxide: Stability, Activity, and Reusability Insights. In, *Methods in Enzymology*. Elsevier, pp. 247–272.

El Seoud,O.A., Baader,W.J., and Bastos,E.L. (2016) Practical Chemical Kinetics in Solution. In, Wang,Z. (ed), *Encyclopedia of Physical Organic Chemistry, 5 Volume Set*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 1–68.

El-Din,A.G., Smith,D.W., and El-Din,M.G. (2004) Application of artificial neural networks in wastewater treatment. *Journal of Environmental Engineering and Science*, **3**, S81–S95.

Ellis,G. (2012) Four Types of Controllers. In, *Control System Design Guide*. Elsevier, pp. 97–119.

Erb,T.J., Frerichs-Revermann,L., Fuchs,G., and Alber,B.E. (2010) The Apparent Malate Synthase Activity of Rhodobacter sphaeroides Is Due to Two Paralogous Enzymes, (3S)-

Malyl-Coenzyme A (CoA)/β-Methylmalyl-CoA Lyase and (3S)- Malyl-CoA Thioesterase. *JB*, **192**, 1249–1258.

Eriksen,D.T., Lian,J., and Zhao,H. (2014) Protein design for pathway engineering. *Journal of Structural Biology*, **185**, 234–242.

Espenshade,P.J. (2013) Cholesterol Synthesis and Regulation. In, *Encyclopedia of Biological Chemistry*. Elsevier, pp. 516–520.

Eubank,W.B. and Reeves,R.E. (1982) Analog Inhibitors for the Pyrophosphate-Dependent Phosphofructokinase of Entamoeba histolytica and their Effect on Culture Growth. *The Journal of Parasitology*, **68**, 599.

van Eunen,K., Bouwman,J., Daran-Lapujade,P., Postmus,J., Canelas,A.B., Mensonides,F.I.C., Orij,R., Tuzun,I., van den Brink,J., Smits,G.J., *et al.* (2010) Measuring enzyme activities under standardized in vivo-like conditions for systems biology: Standardized enzyme assays for systems biology. *FEBS Journal*, **277**, 749–760.

Fang,X., Yu,S.X., Lu,Y., Bast,R.C., Woodgett,J.R., and Mills,G.B. (2000) Phosphorylation and inactivation of glycogen synthase kinase 3 by protein kinase A. *Proceedings of the National Academy of Sciences*, **97**, 11960–11965.

Fell,D.A. (1992) Metabolic control analysis: a survey of its theoretical and experimental development. *Biochemical Journal*, **286**, 313–330.

Fell,D.A. and Small,J.R. (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochemical Journal*, **238**, 781–786.

Fortman,J.L., Chhabra,S., Mukhopadhyay,A., Chou,H., Lee,T.S., Steen,E., and Keasling,J.D. (2008) Biofuel alternatives to ethanol: pumping the microbial well. *Trends in Biotechnology*, **26**, 375–381.

Fouchard,S., Pruvost,J., Degrenne,B., Titica,M., and Legrand,J. (2009) Kinetic modeling of light limitation and sulfur deprivation effects in the induction of hydrogen production with *Chlamydomonas reinhardtii* : Part I. Model development and parameter identification. *Biotechnol. Bioeng.*, **102**, 232–245.

Frahm,B., Brod,H., and Langer,U. (2009) Improving bioreactor cultivation conditions for sensitive cell lines by dynamic membrane aeration. *Cytotechnology*, **59**, 17–30.

Francke,T., López-Tarazón,J.A., and Schröder,B. (2008) Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrol. Process.*, **22**, 4892–4904.

Frieden,C., Gilbert,H.R., and Bock,P.E. (1976) Phosphofructokinase. III. Correlation of the regulatory kinetic and molecular properties of the rabbit muscle enzyme. *Journal of Biological Chemistry*, **251**, 5644–5647.

Fritsch,S., Guenther,F., and Wright,M.N. (2019) Neuralnet: Training of Neural Networks. R package version 1.44.2.

Fukuda,S., Spreer,W., Yasunaga,E., Yuge,K., Sardsud,V., and Müller,J. (2013) Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management*, **116**, 142–150.

Galvanauskas,V., Simutis,R., and Vaitkus,V. (2019) Adaptive Control of Biomass Specific Growth Rate in Fed-Batch Biotechnological Processes. A Comparative Study. 18.

Garde,R., Ibrahim,B., Kovács,Á.T., and Schuster,S. (2020) Differential equation-based minimal model describing metabolic oscillations in Bacillus subtilis biofilms. 14.

Garfinkel,D., Garfinkel,L., Pring,M., Green,S.B., and Chance,B. (1970) Computer Applications to Biochemical Kinetics. *Annu. Rev. Biochem.*, **39**, 473–498.

Genuer,R., Poggi,J.-M., and Tuleau-Malot,C. Variable selection using Random Forests. 11.

Gerdes,H. (2021) Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. 15.

Gernaey,K.V., van Loosdrecht,M.C.M., Henze,M., Lind,M., and Jørgensen,S.B. (2004) Activated sludge wastewater treatment plant modelling and simulation: state of the art. *Environmental Modelling & Software*, **19**, 763–783.

Gerson,D.F., Kole,M.M., Ozum,B., and Oguztoreli,M.N. (1988) Substrate Concentration Control in Bioreactors. *Biotechnology and Genetic Engineering Reviews*, **6**, 67–150.

Giusti,S., Mazzei,D., Cacopardo,L., Mattei,G., Domenici,C., and Ahluwalia,A. (2017) Environmental Control in Flow Bioreactors. *Processes*, **5**, 16.

Goldrick,S., Ştefan,A., Lovett,D., Montague,G., and Lennox,B. (2015) The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology*, **193**, 70–82.

Gonzales,M.L.M., Dans,L.F., and Sio-Aguilar,J. (2019) Antiamoebic drugs for treating amoebic colitis. *Cochrane Database of Systematic Reviews*.

González-Chávez,Z., Olin-Sandoval,V., Rodíguez-Zavala,J.S., Moreno-Sánchez,R., and Saavedra,E. (2015) Metabolic control analysis of the Trypanosoma cruzi peroxide detoxification pathway identifies tryparedoxin as a suitable drug target. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1850**, 263–273.

González-Chávez,Z., Vázquez,C., Mejia-Tlachi,M., Márquez-Dueñas,C., Manning-Cela,R., Encalada,R., Rodríguez-Enríquez,S., Michels,P.A.M., Moreno-Sánchez,R., and Saavedra,E. (2019) Gamma-glutamylcysteine synthetase and tryparedoxin 1 exert high control on the antioxidant system in Trypanosoma cruzi contributing to drug resistance and infectivity. *Redox Biology*, **26**, 101231.

Grissa,D., Pétéra,M., Brandolini,M., Napoli,A., Comte,B., and Pujos-Guillot,E. (2016) Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Front. Mol. Biosci.*, **3**.

Guldberg,C. and Waage,P. (1867) Études sur les affinités chimiques. Christiania : Imprimerie de Brøgger&Christie.

Hagan,M.T., Demuth,H.B., Beale,M.H., and De Jesús,O. (2014) Neural Network Design Martin Hagan.

Hanif,A., Yasmeen,A., and Rajoka,M.I. (2004) Induction, production, repression, and de-repression of exoglucanase synthesis in Aspergillus niger. *Bioresource Technology*, **94**, 311–319.

Haque,M. (2020) The COVID-19 Pandemic - A Global Public Health Crisis: A Brief Overview Regarding Pharmacological Interventions. *Pesqui. Bras. Odontopediatria Clín. Integr.*, **20**, e0146.

Hartwell,L.H., Hopfield,J.J., Leibler,S., and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Hatzimanikatis,V., Emmerling,M., Sauer,U., and Bailey,J.E. (1998) Application of mathematical tools for metabolic design of microbial ethanol production. *BIOTECHNOLOGY AND BIOENGINEERING*, **58**, 8.

Hatzimanikatis,V. et al (1997) Effects of spatiotemporal variations on metabolic control: Approximate analysis using (log)linear kinetic models. *BIOTECHNOLOGY AND BIOENGINEERING*, **54**, 14.

Heckmann,D. (2018) Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *NATURE COMMUNICATIONS*, 10.

Heijnen,J.J. (2005) Approximative kinetic formats used in metabolic network modeling. *Biotechnol. Bioeng.*, **91**, 534–545.

Hein,J.E. (2021) Machine learning made easy for optimizing chemical reactions. *Nature*, **590**, 40–41.

Hoops,S., Sahle,S., Gauges,R., Lee,C., Pahle,J., Simus,N., Singhal,M., Xu,L., Mendes,P., and Kummer,U. (2006) COPASI--a COmplex PAthway SImulator. *Bioinformatics*, **22**, 3067–3074.

Hou,J., Acharya,L., Zhu,D., and Cheng,J. (2016) An overview of bioinformatics methods for modeling biological pathways in yeast. *Briefings in Functional Genomics*, **15**, 95–108.

Hu,R., Lan,D., Cui,R., Hong,H., Niu,Y., and Wang,Y. (2020) Controlling methanol feeding for recombinant protein production by Pichia pastoris under oxidation stress in fed-batch fermentation In Review.

Huang,H. and Buekens,A. (2001) Chemical kinetic modeling of *de novo* synthesis of PCDD/F in municipal waste incinerators. *Chemosphere*, **44**, 1505–1510.

Huang,K.-P. (1989) The mechanism of protein kinase C activation. *Trends in Neurosciences*, **12**, 425–432.

Hussien,S.Y.S., Jaafar,H.I., Ghazali,R., and Razif,N.R.A. (2015) The Effects of Auto-Tuned Method in PID and PD Control Scheme for Gantry Crane System. **4**, 6.

Huyett,L.M., Dassau,E., Zisser,H.C., and Doyle,F.J. (2015) Design and Evaluation of a Robust PID Controller for a Fully Implantable Artificial Pancreas. *Ind. Eng. Chem. Res.*, **54**, 10311–10321.

Imtiaz,U., Jamuar,S.S., Sahu,J.N., and Ganesan,P.B. (2014) Bioreactor profile control by a nonlinear auto regressive moving average neuro and two degree of freedom PID controllers. *Journal of Process Control*, **24**, 1761–1777.

Jaenicke,L. (2007) Centenary of the Award of a Nobel Prize to Eduard Buchner, the Father of Biochemistry in a Test Tube and Thus of Experimental Molecular Bioscience. *Angew. Chem. Int. Ed.*, **46**, 6776–6782.

Jain,A.K., Mao,J., and Mohiuddin,K.M. (1996) Artificial Neural Networks: A Tutorial. *Computer*, **29**, 31–44.

Jia,G., Stephanopoulos,G.N., and Gunawan,R. (2011) Parameter estimation of kinetic models from metabolic profiles: two-phase dynamic decoupling method. *Bioinformatics*, **27**, 1964–1970.

Jiang,X., Meng,X., and Xian,M. (2009) Biosynthetic pathways for 3-hydroxypropionic acid production. *Appl Microbiol Biotechnol*, **82**, 995–1003.

Johnsen,U. and Schönheit,P. (2007) Characterization of cofactor-dependent and cofactor-independent phosphoglycerate mutases from Archaea. *Extremophiles*, **11**, 647–657.

Ju,S., Shaltiel,G., Shamir,A., Agam,G., and Greenberg,M.L. (2004) Human 1-D-myo-Inositol-3-phosphate Synthase Is Functional in Yeast. **279**, 21759–21765.

Jurica,M.S., Mesecar,A., Heath,P.J., Shi,W., Nowak,T., and Stoddard,B.L. (1998) The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure*, **6**, 195–210.

Kadarmideen,H.N. (2016) Systems Biology in Animal Production and Health, Vol. 2. Springer International Publishing, Cham, pp. 136–143.

Kang,J.-H., McRoberts,J., Shi,F., Moreno,J.E., Jones,A.D., and Howe,G.A. (2014) The Flavonoid Biosynthetic Enzyme Chalcone Isomerase Modulates Terpenoid Production in Glandular Trichomes of Tomato. *Plant Physiology*, **164**, 1161–1174.

Kantor,M., Abrantes,A., Estevez,A., Schiller,A., Torrent,J., Gascon,J., Hernandez,R., and Ochner,C. (2018) *Entamoeba histolytica*: Updates in Clinical Manifestation, Pathogenesis, and Vaccine Development. *Canadian Journal of Gastroenterology and Hepatology*, **2018**, 1–6.

Kelleher,J.K. (2001) Flux Estimation Using Isotopic Tracers: Common Ground for Metabolic Physiology and Metabolic Engineering. *Metabolic Engineering*, **3**, 100–110.

Kerkhoven,E.J., Lahtvee,P.-J., and Nielsen,J. (2014) Applications of computational modeling in metabolic engineering of yeast. *FEMS Yeast Res*, n/a-n/a.

Khodayari,A. and Maranas,C.D. (2016) A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat Commun*, **7**, 13806.

Kim,G.B., Kim,W.J., Kim,H.U., and Lee,S.Y. (2020) Machine learning applications in systems metabolic engineering. *Current Opinion in Biotechnology*, **64**, 1–9.

Kim,O.D., Rocha,M., and Maia,P. (2018) A Review of Dynamic Modeling Approaches and Their Application in Computational Strain Optimization for Metabolic Engineering. *Front. Microbiol.*, **9**, 1690.

Kolanoski,H. (1995) Application of artificial neural networks in particle physics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **367**, 14–20.

Kotera,M., Tabei,Y., Yamanishi,Y., Tokimatsu,T., and Goto,S. (2013) Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets. *Bioinformatics*, **29**, i135–i144.

Küçükönder,H., Boyaci,S., and Akyüz,A. (2016) A modeling study with an artificial neural network: developing estimationmodels for the tomato plant leaf area. *TURKISH JOURNAL OF AGRICULTURE AND FORESTRY*, **40**, 203–212.

Kuhn,M. (2020) caret : Classification and Regression Training.

Kumar,M., Prasad,D., Giri,B.S., and Singh,R.S. (2019) Temperature control of fermentation bioreactor for ethanol production using IMC-PID controller. *Biotechnology Reports*, **22**, e00319.

Kumar,S. and Barth,A. (2010) Phosphoenolpyruvate and Mg2+ Binding to Pyruvate Kinase Monitored by Infrared Spectroscopy. *Biophysical Journal*, 10.

Lancashire,L.J., Lemetre,C., and Ball,G.R. (2008) An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics*, **10**, 315–329.

Lavric,V., Ofiţeru,I.D., and Woinaroschy,A. (2005) A sensitivity analysis of the fed-batch animal-cell bioreactor with respect to some control parameters. *Biotechnol. Appl. Biochem.*, **41**, 29.

Layeghifard,M., Hwang,D.M., and Guttman,D.S. (2018) Constructing and Analyzing Microbiome Networks in R. In, Beiko,R.G., Hsiao,W., and Parkinson,J. (eds), *Microbiome Analysis*, Methods in Molecular Biology. Springer New York, New York, NY, pp. 243–266.

Lee,S.C., Hwang,Y.B., Chang,H.N., and Chang,Y.K. (1991) Adaptive control of dissolved oxygen concentration in a bioreactor. *Biotechnol. Bioeng.*, **37**, 597–607.

Leighty,R.W. and Antoniewicz,M.R. (2011) Dynamic metabolic flux analysis (DMFA): A framework for determining fluxes at metabolic non-steady state. *Metabolic Engineering*, **13**, 745–755.

L'Heureux,A., Grolinger,K., Elyamany,H.F., and Capretz,M.A.M. (2017) Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, **5**, 7776–7797.

Li,X. and Wen,J. (2014) Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews*, **37**, 517–537.

Li,Y., Wang,S., Umarov,R., Xie,B., Fan,M., Li,L., and Gao,X. (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.

Liaw,A. and Wiener,M. (2002) Classification and Regression by randomForest. **2**, 5.

Liebermeister,W., Uhlendorf,J., and Klipp,E. (2010) Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics*, **26**, 1528–1534.

Liu,H., Ramos,K.R.M., Valdehuesa,K.N.G., Nisola,G.M., Lee,W.-K., and Chung,W.-J. (2013) Biosynthesis of ethylene glycol in Escherichia coli. *Appl Microbiol Biotechnol*, **97**, 3409–3417.

Liu,J., Brazier-Hicks,M., and Edwards,R. (2009) A kinetic model for the metabolism of the herbicide safener fenclorim in *Arabidopsis thaliana*. *Biophysical Chemistry*, **143**, 85–94.

Lopina,O.D. (2017) Enzyme Inhibitors and Activators. In, Senturk,M. (ed), *Enzyme Inhibitors and Activators*. InTech.

Lo-Thong,O., Charton,P., Cadet,X.F., Damour,C., Grondin-Perez,B., Saavedra,E., and Cadet,F. (2020) Identification of flux checkpoints in a metabolic pathway through white-box, grey-box and black-box modeling approaches. *Scientific Reports*, 19.

M. Vastrad,C. (2013) Performance Analysis of Neural Network Models for Oxazolines and Oxazoles Derivatives Descriptor Dataset. *International Journal of Information Sciences and Techniques*, **3**, 1–15.

Ma,Y., Ding,Z., Qian,Y., Shi,X., Castranova,V., Harner,E.J., and Guo,L. (2006) Predicting Cancer Drug Response by Proteomic Profiling. *Clinical Cancer Research*, **12**, 4583–4589.

Marín-Hernández,Á., Rodríguez-Zavala,J.S., Del Mazo-Monsalvo,I., Rodríguez-Enríquez,S., Moreno-Sánchez,R., and Saavedra,E. (2016) Inhibition of Non-flux-Controlling Enzymes Deters Cancer Glycolysis by Accumulation of Regulatory Metabolites of Controlling Steps. *Front. Physiol.*, **7**.

Marlin,T.E. (2000) Chapter 9: PID Controller Tuning for Dynamic Performance. In, *Process Control: Designing Processes and Control Systems for Dynamic Performance.*, p. 1056.

Martínez,J.L. (2016) The impact of respiration and oxidative stress response on recombinant α-amylase production by Saccharomyces cerevisiae. *Metabolic Engineering Communications*, 6.

Matamoros,M., Gómez-Blanco,J.C., Sánchez,Á.J., Mancha,E., Marcos,A.C., Carrasco-Amador,J.P., and Pagador,J.B. (2020) Temperature and Humidity PID Controller for a Bioprinter Atmospheric Enclosure System. *Micromachines*, **11**, 999.

Matera,S., Schneider,W.F., Heyden,A., and Savara,A. (2019) Progress in Accurate Chemical Kinetic Modeling, Simulations, and Parameter Estimation for Heterogeneous Catalysis. *ACS Catal.*, **9**, 6624–6647.

Mazzei,D., Vozzi,F., Cisternino,A., Vozzi,G., and Ahluwalia,A. (2008) A High-Throughput Bioreactor System for Simulating Physiological Environments. *IEEE Trans. Ind. Electron.*, **55**, 3273–3280.

McCloskey,D., Palsson,B.Ø., and Feist,A.M. (2013) Basic and applied uses of genome−scale metabolic network reconstructions of *Escherichia coli*. *Mol Syst Biol*, **9**, 661.

McMillan,G.K. (2012) Industrial Applications of PID Control. In, Vilanova,R. and Visioli,A. (eds), *PID Control in the Third Millennium*, Advances in Industrial Control. Springer London, London, pp. 415–461.

Medvedev,A., Zhusubaliyev,Z.T., Rosén,O., and Silva,M.M. (2019) Oscillations-free PID control of anesthetic drug delivery in neuromuscular blockade. *Computer Methods and Programs in Biomedicine*, **171**, 119–131.

Meinshausen,N. (2006) Quantile Regression Forests. 983–999.

Mendes,P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.*, **22**, 361–363.

Mendes,P. (1993) GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. **9**, 563–571.

Mendes,P. and Kell,D.B. (1996) On the analysis of the inverse problem of metabolic pathways using artificial neural networks. *Biosystems*, **38**, 15–28.

Mesquita,T.J.B., Sargo,C.R., Fuzer,J.R., Paredes,S.A.H., Giordano,R. de C., Horta,A.C.L., and Zangirolami,T.C. (2019) Metabolic fluxes-oriented control of bioreactors: a novel approach to tune micro-aeration and substrate feeding in fermentations. *Microb Cell Fact*, **18**, 150.

Meunier,J.-C., Buc,J., Navarro,A., and Ricard,J. (1974) Regulatory Behavior of Monomeric Enzymes. 2. A Wheat-Germ Hexokinase as a Mnemonical Enzyme. *Eur. J. Biochem.*, 15.

Michaelis,L. and Menten,M. (1913) Die Kinetik der Invertinwirkung. *Biochem Z*, 333–369.

Mi-ichi,F., Ishikawa,T., Tam,V.K., Deloer,S., Hamano,S., Hamada,T., and Yoshida,H. (2019) Characterization of Entamoeba histolytica adenosine 5′-phosphosulfate (APS) kinase; validation as a target and provision of leads for the development of new drugs against amoebiasis. *PLoS Negl Trop Dis*, **13**, e0007633.

Min,X., Feng,M., Guan,Y., Man,S., Fu,Y., Cheng,X., and Tachibana,H. (2016) Evaluation of the C-Terminal Fragment of Entamoeba histolytica Gal/GalNAc Lectin Intermediate Subunit as a Vaccine Candidate against Amebic Liver Abscess. *PLoS Negl Trop Dis*, **10**, e0004419.

Mitchell,J.B.O. (2014) Machine learning methods in chemoinformatics. **4**, 468–481.

Moreno-Sánchez,R., Encalada,R., Marín-Hernández,A., and Saavedra,E. (2008) Experimental validation of metabolic pathway modeling: An illustration with glycolytic segments from *Entamoeba histolytica*. *FEBS Journal*, **275**, 3454–3469.

Moreno-Sánchez,R., Saavedra,E., Rodríguez-Enríquez,S., and Olín-Sandoval,V. (2008) Metabolic Control Analysis: A Tool for Designing Strategies to Manipulate Metabolic Pathways. *Journal of Biomedicine and Biotechnology*, **2008**, 1–30.

Morgan,J.A. and Rhodes,D. (2002) Mathematical Modeling of Plant Metabolic Pathways. *Metabolic Engineering*, **4**, 80–89.

Mukhopadhyay,A., Redding,A.M., Rutherford,B.J., and Keasling,J.D. (2008) Importance of systems biology in engineering microbes for biofuel production. *Current Opinion in Biotechnology*, **19**, 228–234.

Muller,M., Mentel,M., van Hellemond,J.J., Henze,K., Woehle,C., Gould,S.B., Yu,R.-Y., van der Giezen,M., Tielens,A.G.M., and Martin,W.F. (2012) Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. *Microbiology and Molecular Biology Reviews*, **76**, 444–495.

Murdoch,W.J., Singh,C., Kumbier,K., Abbasi-Asl,R., and Yu,B. (2019) Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci USA*, **116**, 22071–22080.

Murrell,P. and Wehrens,R. (2006) Non-linear regression for optimising the separation of carboxylic acids.

Nagy,Z.K. (2007) Model based control of a yeast fermentation bioreactor using optimally designed artificial neural networks. *Chemical Engineering Journal*, **127**, 95–109.

Nakamura,M., Hachiya,T., Saito,Y., Sato,K., and Sakakibara,Y. (2012) An efficient algorithm for de novo predictions of biochemical pathways between chemical compounds. 12.

Naushad,S.M., Janaki Ramaiah,M., Pavithrakumari,M., Jayapriya,J., Hussain,T., Alrokayan,S.A., Gottumukkala,S.R., Digumarti,R., and Kutala,V.K. (2016) Artificial neural network-based exploration of gene-nutrient interactions in folate and xenobiotic metabolic pathways that modulate susceptibility to breast cancer. *Gene*, **580**, 159–168.

Neil,S.M., Lakey,T., and Tomlinson,S. (1985) Calmodulin regulation of adenylate cyclase activity. *Cell Calcium*, **6**, 213–226.

Nielsen,J. and Jorgensen,H.S. (1995) Metabolic control analysis of the penicillin biosynthetic pathway in a high-yielding strain of Penicillium chrysogenum. *Biotechnol. Prog.*, **11**, 299–305.

Nöh,K., Grönke,K., Luo,B., Takors,R., Oldiges,M., and Wiechert,W. (2007) Metabolic flux analysis at ultra short time scale: Isotopically non-stationary 13C labeling experiments. *Journal of Biotechnology*, **129**, 249–267.

Onoyovwe,A., Hagel,J.M., Chen,X., Khan,M.F., Schriemer,D.C., and Facchini,P.J. (2013) Morphine Biosynthesis in Opium Poppy Involves Two Cell Types: Sieve Elements and Laticifers. *Plant Cell*, **25**, 4110–4122.

Oreski,D., Oreski,S., and Klicek,B. (2017) Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, **52**, 109–119.

Orozco,E., Marchat,L.A., Gómez,C., López-Camarillo,C., and Pérez,D.G. (2009) Drug Resistance Mechanisms in Entamoeba histolytica, Giardia lamblia, Trichomonas vaginalis, and Opportunistic Anaerobic Protozoa. 11.

Orth,J.D., Thiele,I., and Palsson,B.Ø. (2010) What is flux balance analysis? *Nat Biotechnol*, **28**, 245–248.

Othman,N., Saidin,S., and Noordin,R. (2017) In-vitro Testing of Potential *Entamoeba histolytica* Pyruvate Phosphate Dikinase Inhibitors. *The American Journal of Tropical Medicine and Hygiene*, **97**, 1204–1213.

Oyetunde,T., Bao,F.S., Chen,J.-W., Martin,H.G., and Tang,Y.J. (2018) Leveraging knowledge engineering and machine learning for microbial bio-manufacturing. *Biotechnology Advances*, **36**, 1308–1315.

Oyetunde,T., Liu,D., Martin,H.G., and Tang,Y.J. (2019) Machine learning framework for assessment of microbial factory performance. *PLoS ONE*, **14**, e0210558.

Ozturk,S.S. and Hu,W.-S. (2006) PHARMACEUTICAL AND CELL-BASED THERAPIES. 784.

Pachauri,N., Rani,A., and Singh,V. (2017) Bioreactor temperature control using modified fractional order IMC-PID for ethanol production. *Chemical Engineering Research and Design*, **122**, 97–112.

Paddon,C.J., Westfall,P.J., Pitera,D.J., Benjamin,K., Fisher,K., McPhee,D., Leavell,M.D., Tai,A., Main,A., Eng,D., *et al.* (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, **496**, 528–532.

Pan,L., Cheng,C., Haberkorn,U., and Dimitrakopoulou-Strauss,A. (2017) Machine learning-based kinetic modeling: a robust and reproducible solution for quantitative analysis of dynamic PET data. *Phys. Med. Biol.*, 17.

Panchal,G., Ganatra,A., Kosta,Y.P., and Panchal,D. (2010) Searching Most Efficient Neural Network Architecture Using Akaike's Information Criterion (AIC). *IJCA*, **1**, 54–57.

Patnaik,P.R. (1997) Artificial intelligence as a tool for automatic state estimation and control of bioreactors. **9**, 297–304.

Pawul,M. and Śliwka,M. (2016) APPLICATION OF ARTIFICIAL NEURAL NETWORKS FOR PREDICTION OF AIR POLLUTION LEVELS IN ENVIRONMENTAL MONITORING. *J. Ecol. Eng.*, **17**, 190–196.

Persad,A., Chopda,V.R., Rathore,A.S., and Gomes,J. (2013) Comparative Performance of Decoupled Input–Output Linearizing Controller and Linear Interpolation PID Controller: Enhancing Biomass and Ethanol Production in Saccharomyces cerevisiae. *Appl Biochem Biotechnol*, **169**, 1219–1240.

Petroll,K., Kopp,D., Care,A., Bergquist,P.L., and Sunna,A. (2019) Tools and strategies for constructing cell-free enzyme pathways. *Biotechnology Advances*, **37**, 91–108.

Philibert,A., Loyce,C., and Makowski,D. (2013) Prediction of N2O emission from local information with Random Forest. *Environmental Pollution*, **177**, 156–163.

Pickering,E.M., Hossain,M.A., Mousseau,J.P., Swanson,R.A., French,R.H., and Abramson,A.R. (2017) A cross-sectional study of the temporal evolution of electricity consumption of six commercial buildings. *PLoS ONE*, **12**, e0187129.

Pillay,C.S., Hofmeyr,J.-H.S., Olivier,B.G., Snoep,J.L., and Rohwer,J.M. (2009) Enzymes or redox couples? The kinetics of thioredoxin and glutaredoxin reactions in a systems biology context. *Biochemical Journal*, **417**, 269–277.

Pillay,C.S., Hofmeyr,J.-H.S., and Rohwer,J.M. (2011) The logic of kinetic regulation in the thioredoxin system. *BMC Syst Biol*, **5**, 15.

Pineda,E., Encalada,R., Rodríguez-Zavala,J.S., Olivos-García,A., Moreno-Sánchez,R., and Saavedra,E. (2010) Pyruvate:ferredoxin oxidoreductase and bifunctional aldehyde-alcohol dehydrogenase are essential for energy metabolism under oxidative stress in Entamoeba histolytica: Fermenting enzymes and oxidative stress in Entamoeba. *FEBS Journal*, **277**, 3382–3395.

Pineda,E., Encalada,R., Vázquez,C., González,Z., Moreno-Sánchez,R., and Saavedra,E. (2015) Glucose Metabolism and Its Controlling Mechanisms in Entamoeba histolytica. In, Nozaki,T. and Bhattacharya,A. (eds), *Amebiasis*. Springer Japan, Tokyo, pp. 351–372.

Pineda,E., Encalada,R., Vázquez,C., Néquiz,M., Olivos-García,A., Moreno-Sánchez,R., and Saavedra,E. (2015) *In vivo* identification of the steps that control energy metabolism and survival of *Entamoeba histolytica*. *FEBS J*, **282**, 318–331.

Pintelas,E., Livieris,I.E., and Pintelas,P. (2020) A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. *Algorithms*, **13**, 17.

Porter,C.M. and Miller,B.G. (2012) Cooperativity in monomeric enzymes with single ligand-binding sites. *Bioorganic Chemistry*, **43**, 44–50.

Puri,M., Solanki,A., Padawer,T., Tipparaju,S.M., Moreno,W.A., and Pathak,Y. (2016) Introduction to Artificial Neural Network (ANN) as a Predictive Tool for Drug Design,

Discovery, Delivery, and Disposition. In, *Artificial Neural Network for Drug Design, Delivery and Disposition*. Elsevier, pp. 3–13.

Qi,Y., Bar-Joseph,Z., and Klein-Seetharaman,J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**, 490–500.

Quach,J., St-Pierre,J., and Chadee,K. (2014) The future for vaccine development against *Entamoeba histolytica*. *Human Vaccines & Immunotherapeutics*, **10**, 1514–1521.

Qureshi,N. (2009) Solvent Production. In, *Encyclopedia of Microbiology*. Elsevier, pp. 512–528.

Rajasethupathy,P., Vayttaden,S.J., and Bhalla,U.S. (2005) Systems modeling: a pathway to drug discovery. *Current Opinion in Chemical Biology*, **9**, 400–406.

Ramachandran,S., Chaudhuri,R., Prasad,S., Rauf,A., Paul,C., Chakraborty,S., Lal,B., and Shubhra,R. (2011) Biological Data Modelling and Scripting in R. In, Yang,N.-S. (ed), *Systems and Computational Biology - Bioinformatics and Computational Modeling*. InTech.

Raman,K. and Chandra,N. (2009) Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics*, **10**, 435–449.

Rana,P., Berry,C., Ghosh,P., and Fong,S.S. (2020) Recent advances on constraint-based models by integrating machine learning. *Current Opinion in Biotechnology*, 7.

Rapoport,T., Heinrich,R., Jacobasch,G., and Rapoport,S. (1974) A Linear Steady−State Treatment of Enzymatic Chains. *Eur. J. Biochem.*, 107–129.

Réda,C., Kaufmann,E., and Delahaye-Duriez,A. (2020) Machine learning applications in drug development. *Computational and Structural Biotechnology Journal*, **18**, 241–252.

Ricard,J., Meunier,J.-C., and Buc,J. (1974) Regulatory Behavior of Monomeric Enzymes. 1. The Mnemonical Enzyme Concept. *Eur J Biochem*, **49**, 195–208.

Riddick,G., Song,H., Ahn,S., Walling,J., Borges-Rivera,D., Zhang,W., and Fine,H.A. (2011) Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics*, **27**, 220–224.

van Riel,N.A.W., Tiemann,C.A., Hilbers,P.A.J., and Groen,A.K. (2021) Metabolic Modeling Combined With Machine Learning Integrates Longitudinal Data and Identifies the Origin of LXR-Induced Hepatic Steatosis. *Front. Bioeng. Biotechnol.*, **8**, 536957.

Ripley,B. and Venables,W. (2009) Package 'nnet'.

Rohwer,J.M. (2012) Kinetic modelling of plant metabolic pathways. *Journal of Experimental Botany*, **63**, 2275–2292.

Roncolato,E.C., Teixeira,J.E., Barbosa,J.E., Zambelli Ramalho,L.N., and Huston,C.D. (2015) Immunization with the Entamoeba histolytica Surface Metalloprotease EhMSP-1 Protects Hamsters from Amebic Liver Abscess. *Infect. Immun.*, **83**, 713–720.

del Rosario,R.C.H., Mendoza,E., and Voit,E.O. (2008) Challenges in lin-log modelling of glycolysis in *Lactococcus lactis*. *IET Syst. Biol.*, **2**, 136.

Rumelhart, David E.,H.,Geoffrey E. and Williams, Ronald J. (1986) Learning Representations by Back Propagating Errors. *Nature*, **323**, 533–536.

Saa,P.A. and Nielsen,L.K. (2016) Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. *Sci Rep*, **6**, 29635.

Saavedra,E., Encalada,R., Pineda,E., Jasso-Chávez,R., and Moreno-Sánchez,R. (2005) Glycolysis in *Entamoeba histolytica* : Biochemical characterization of recombinant glycolytic enzymes and flux control analysis. *FEBS Journal*, **272**, 1767–1783.

Saavedra,E., Encalada,R., Vázquez,C., Olivos-García,A., Michels,P.A.M., and Moreno-Sánchez,R. (2019) Control and regulation of the pyrophosphate-dependent glucose metabolism in *Entamoeba histolytica*. *Molecular and Biochemical Parasitology*, **229**, 75–87.

Saavedra,E., Gonzalez-Chavez,Z., Moreno-Sanchez,R., and Michels,P.A.M. (2019) Drug Target Selection for *Trypanosoma cruzi* Metabolism by Metabolic Control Analysis and Kinetic Modeling. *CMC*, **26**.

Saavedra,E., Marín-Hernández,A., Encalada,R., Olivos,A., Mendoza-Hernández,G., and Moreno-Sánchez,R. (2007) Kinetic modeling can describe *in vivo* glycolysis in *Entamoeba histolytica*: Modeling Entamoeba glycolysis. *FEBS Journal*, **274**, 4922–4940.

Saavedra,E., Olivos,A., Encalada,R., and Moreno-Sánchez,R. (2004) Entamoeba histolytica: kinetic and molecular evidence of a previously unidentified pyruvate kinase. *Experimental Parasitology*, **106**, 11–21.

Saidin,S., Othman,N., and Noordin,R. (2017) In Vitro Testing of Potential Entamoeba histolytica Pyruvate Phosphate Dikinase Inhibitors. *The American Journal of Tropical Medicine and Hygiene*, **97**, 1204–1213.

Saidin,S., Yunus,M.H., Zakaria,N.D., Razak,K.A., Huat,L.B., Othman,N., and Noordin,R. (2014) Production of recombinant Entamoeba histolyticapyruvate phosphate dikinase and its application in a lateral flow dipstick test for amoebic liver abscess. *BMC Infect Dis*, **14**, 182.

Samarawickrema,N. (1997) Involvement of superoxide dismutase and pyruvate:ferredoxin oxidoreductase in mechanisms of metronidazole resistance in *Entamoeba histolytica*. *Journal of Antimicrobial Chemotherapy*, **40**, 833–840.

Savageau,M.A. (1970) Biochemical Systems Analysis. *J. theor. Biol.*, **26**, 215–226.

Savageau,M.A. (1988) Introduction to S-systems and the underlying power-law formalism. *Mathematical and Computer Modelling*, **11**, 546–551.

Scherlozer,A., Orsini,M., and Patole,S. (2016) Simulation and Numerical Analysis and Comparative Study of a PID Controller Based on Ziegler-Nichols and Auto Turning Method. 17.

Schmidt,J., Marques,M.R.G., Botti,S., and Marques,M.A.L. (2019) Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater*, **5**, 83.

Schultz,M. and Reitmann,S. (2018) Prediction of aircraft boarding time using LSTM network. In, *2018 Winter Simulation Conference (WSC)*. IEEE, Gothenburg, Sweden, pp. 2330–2341.

Segel,I.H. (1975) Enzyme Kinetics Wiley. New York, USA.

Sel'Kov,E.E. (1968) Self-Oscillations in Glycolysis. 1. A Simple Kinetic Model. *Eur J Biochem*, **4**, 79–86.

Sengupta,N., Rose,S.T., and Morgan,J.A. (2011) Metabolic flux analysis of CHO cell metabolism in the late nongrowth phase. *Biotechnology and Bioengineering*, **108**, 11.

Sergei Mikhalevich, Francesco Rossi, Flavio Manenti, and Sergey Baydali (2015) Robust pi/pid controller design for the reliable control of plug flow reactor. *Chemical Engineering Transactions*, **43**, 1525–1530.

Setiawan,I. (2020) Time series air quality forecasting with R Language and R Studio. *J. Phys.: Conf. Ser.*, **1450**, 012064.

Shaked,I., Oberhardt,M.A., Atias,N., Sharan,R., and Ruppin,E. (2016) Metabolic Network Prediction of Drug Side Effects. *Cell Systems*, **2**, 209–213.

Shapiro,N.Z. and Shapley,L.S. (1965) Mass Action Laws and the Gibbs Free Energy Function. *Journal of the Society for Industrial and Applied Mathematics*, **13**, 353–375.

Shi,S., Chen,Y., Siewers,V., and Nielsen,J. (2014) Improving Production of Malonyl Coenzyme A-Derived Metabolites by Abolishing Snf1-Dependent Regulation of Acc1. *mBio*, **5**, e01130-14.

Shimizu,H., Mizuguchi,T., Tanaka,E., and Shioya,S. (1999) Nisin Production by a Mixed-Culture System Consisting of Lactococcus lactis andKluyveromyces marxianus. *APPL. ENVIRON. MICROBIOL.*, **65**, 8.

Shirley,D.-A.T., Farr,L., Watanabe,K., and Moonah,S. (2018) A Review of the Global Burden, New Diagnostics, and Current Therapeutics for Amebiasis. *Open Forum Infectious Diseases*, **5**.

Sigurdsson,M.I., Jamshidi,N., Steingrimsson,E., Thiele,I., and Palsson,B.Ø. (2010) A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst Biol*, **4**, 140.

Silva,M.M., Paz,L., Wigren,T., and Mendonça,T. (2015) Performance of an Adaptive Controller for the Neuromuscular Blockade Based on Inversion of a Wiener Model: Performance of an Adaptive Controller for the NMB Based on Inversion of a Wiener Model. *Asian Journal of Control*, **17**, 1136–1147.

Smets,I.Y., Bastin,G., and Van Impe,J. (2000) Feedback Control of Fed-Batch Bioreactors for Microbial Growth Processes with Non-Monotonic Kinetics. 6.

Somarathna,P.D.S.N., Minasny,B., and Malone,B.P. (2017) More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. *Soil Science Society of America Journal*, **81**, 1413–1426.

Stanford,N.J., Lubitz,T., Smallbone,K., Klipp,E., Mendes,P., and Liebermeister,W. (2013) Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks. *PLoS ONE*, **8**, e79195.

Stanley,S.L. (2006) Vaccines for amoebiasis: barriers and opportunities. *Parasitology*, **133**, S81–S86.

Stanley,S.L. and Li,E. (2001) Amoebiasis. In, John Wiley & Sons, Ltd (ed), *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester, UK, p. a0001944.

Stephanopoulos,G. (1999) Metabolic Fluxes and Metabolic Engineering. *Metabolic Engineering*, **1**, 1–11.

Stephen,P., Vijayan,R., Bhat,A., Subbarao,N., and Bamezai,R.N.K. (2008) Molecular modeling on pyruvate phosphate dikinase of *Entamoeba histolytica* and *in silico* virtual screening for novel inhibitors. *J Comput Aided Mol Des*, **22**, 647–660.

Sulieman,A.K., Putra,M.D., Abasaeed,A.E., Gaily,M.H., Al-Zahrani,S.M., and Zeinelabdeen,M.A. (2018) Kinetic modeling of the simultaneous production of ethanol and fructose by Saccharomyces cerevisiae. *Electronic Journal of Biotechnology*, **34**, 1–8.

Sutariya,V., Groshev,A., Sadana,P., Bhatia,D., and Pathak,Y. (2013) Artificial Neural Network in Drug Delivery and Pharmaceutical Research. *TOBIOIJ*, **7**, 49–62.

Tami,R., Soualmi,B., Doufene,A., Ibanez,J., and Dauwels,J. (2019) Machine learning method to ensure robust decision-making of AVs. In, *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, Auckland, New Zealand, pp. 1217–1222.

Tan,S.Z., Manchester,S., and Prather,K.L.J. (2016) Controlling Central Carbon Metabolism for Improved Pathway Yields in *Saccharomyces cerevisiae*. *ACS Synth. Biol.*, **5**, 116–124.

Tan,S.Z. and Prather,K.L. (2017) Dynamic pathway regulation: recent advances and methods of construction. *Current Opinion in Chemical Biology*, **41**, 28–35.

Terkelsen,T., Krogh,A., and Papaleo,E. (2020) CAncer bioMarker Prediction Pipeline (CAMPP)—A standardized framework for the analysis of quantitative biological data. *PLoS Comput Biol*, **16**, e1007665.

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

Thishya,K., Vattam,K.K., Naushad,S.M., Raju,S.B., and Kutala,V.K. (2018) Artificial neural network model for predicting the bioavailability of tacrolimus in patients with renal transplantation. *PLOS ONE*, **13**, e0191921.

Treloar,N.J., Fedorec,A.J.H., Ingalls,B., and Barnes,C.P. (2020) Deep reinforcement learning for the control of microbial co-cultures in bioreactors. *PLOS COMPUTATIONAL BIOLOGY*, 18.

Tušek,A. and Kurtanjek,Ž. (2010) Lin-log model of E coli central metabolism. 6.

Upcroft,P. and Upcroft,J.A. (2001) Drug Targets and Mechanisms of Resistance in the Anaerobic Protozoa. *Clinical Microbiology Reviews*, **14**, 150–164.

Ureta,T. (1976) The allosteric regulation of hexokinase C from amphibian liver. *Journal of Biological Chemistry*, **251**, 5035–5042.

Varela-Gómez,M., Moreno-Sánchez,R., Pardo,J.P., and Perez-Montfort,R. (2004) Kinetic Mechanism and Metabolic Role of Pyruvate Phosphate Dikinase from *Entamoeba histolytica*. *Journal of Biological Chemistry*, **279**, 54124–54130.

Villa-Vialaneix,N., Follador,M., and Leip,A. (2010) A comparison of three learning methods to predict N2O fluxes and N leaching. 10.

Visser,D. and Heijnen,J.J. (2003) Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metabolic Engineering*, **5**, 164–176.

Voit,E.O. and Almeida,J. (2004) Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, **20**, 1670–1681.

Wang,L. (2020) PID Control System Design and Automatic Tuning using MATLAB/Simulink 1st ed. Wiley.

Wehrs,M., Tanjore,D., Eng,T., Lievense,J., Pray,T.R., and Mukhopadhyay,A. (2019) Engineering Robust Production Microbes for Large-Scale Cultivation. *Trends in Microbiology*, **27**, 524–537.

Wei,Y., Zhang,X., Shi,Y., Xia,L., Pan,S., Wu,J., Han,M., and Zhao,X. (2018) A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, **82**, 1027–1047.

Werner,R.G. and Noé,W. (1993) Mammalian cell cultures. Part II: Genetic engineering, protein glycosylation, fermentation and process control. *Arzneimittelforschung*, **43**, 1242–1249.

Wess,J., Brinek,M., and Boles,E. (2019) Improving isobutanol production with the yeast Saccharomyces cerevisiae by successively blocking competing metabolic pathways as well as ethanol and glycerol formation. *Biotechnol Biofuels*, **12**, 173.

Wickham,H. and Grolemund,G. (2017) R for Data Science O'Reilly Media.

Wiechert,W. and Noack,S. (2011) Mechanistic pathway modeling for industrial biotechnology: challenging but worthwhile. *Current Opinion in Biotechnology*, **22**, 604–610.

Wiechert,W., Siefke,C., de Graaf,A.A., and Marx,A. (1997) Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. *BIOTECHNOLOGY AND BIOENGINEERING*, **55**, 18.

Wright,M.N. and Ziegler,A. (2017) ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Soft.*, **77**.

Wu,S.G., Wang,Y., Jiang,W., Oyetunde,T., Yao,R., Zhang,X., Shimizu,K., Tang,Y.J., and Bao,F.S. (2016) Rapid Prediction of Bacterial Heterotrophic Fluxomics Using Machine Learning and Constraint Programming. *PLoS Comput Biol*, **12**, e1004838.

Wu,Z., Kan,S.B.J., Lewis,R.D., Wittmann,B.J., and Arnold,F.H. (2019) Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci USA*, **116**, 8852–8858.

Xu,T., Guo,Z., Liu,S., He,X., Meng,Y., Xu,Z., Xia,Y., Xiao,J., Zhang,Y., Ma,Y., *et al.* (2018) Evaluating Different Machine Learning Methods for Upscaling Evapotranspiration from Flux Towers to the Regional Scale. *J. Geophys. Res. Atmos.*, **123**, 8674–8690.

Yan,L., Zhang,H., Xiao,Y., Wang,M., Guo,Y., Sun,C., Tang,X., Jing,L., Li,S., Zhang,M., *et al.* (2020) Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. 18.

Yang,P., Hwa Yang,Y., B. Zhou,B., and Y. Zomaya,A. (2010) A Review of Ensemble Methods in Bioinformatics. *CBIO*, **5**, 296–308.

Yang,R. and Rizzoni,G. (2016) Comparison of Model-based Vs. Data-driven Methods for Fault Detection and Isolation in Engine Idle Speed Control System. 9.

Young,B.J., Riera,N.I., Beily,M.E., Bres,P.A., Crespo,D.C., and Ronco,A.E. (2012) Toxicity of the effluent from an anaerobic bioreactor treating cereal residues on Lactuca sativa. *Ecotoxicology and Environmental Safety*, **76**, 182–186.

Yousoff,S.N.M., Baharin,A., and Abdullah,A. (2017) Differential Search Algorithm in Deep Neural Network for the Predictive Analysis of Xylitol Production in Escherichia Coli. 53–67.

Zamani Joharestani,M., Cao,C., Ni,X., Bashir,B., and Talebiesfandarani,S. (2019) PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, **10**, 373.

Zampieri,G., Vijayakumar,S., Yaneske,E., and Angione,C. (2019) Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput Biol*, **15**, e1007084.

Zelezniak,A., Vowinckel,J., Capuano,F., Messner,C.B., Demichev,V., Polowsky,N., Mülleder,M., Kamrad,S., Klaus,B., Keller,M.A., *et al.* (2018) Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cell Systems*, **7**, 269-283.e6.

Zhang,J., Petersen,S.D., Radivojevic,T., Ramirez,A., Pérez,A., Abeliuk,E., Sánchez,B.J., Costello,Z., Chen,Y., Fero,M., *et al.* (2019) Predictive engineering and optimization of tryptophan metabolism in yeast through a combination of mechanistic and machine learning models Bioengineering.

Zhou,J., Li,E., Wei,H., Li,C., Qiao,Q., and Armaghani,D.J. (2019) Random Forests and Cubist Algorithms for Predicting Shear Strengths of Rockfill Materials. 16.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods*, **12**, 931–934.

Zhou,K., Ng,W., Cortés-Peña,Y., and Wang,X. (2020) Increasing metabolic pathway flux by using machine learning models. *Current Opinion in Biotechnology*, **66**, 179–185.

Zhou,Q., Wang,S., Xu,X., and Xiao,F. (2008) A grey-box model of next-day building thermal load prediction for energy-efficient control. *Int. J. Energy Res.*, 14.

Ziegler,J.G. and Nichols,N.B. (1993) Optimum Settings for Automatic Controllers. *Journal of Dynamic Systems, Measurement, and Control*, **115**, 220–222.

Zilberstein,D. and Shapira,M. The Role of pH and Temperature in the Development of Leishmania Parasites. 22.