Universidade do Minho
Escola de Engenharia

Mariana Marques Pereira

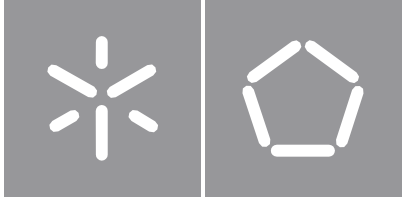
Integrating Kinetic and Constraint-Based Models of Metabolism

October of 2022

Integrating Kinetic and Constraint-Based Models of Metabolism

Mariana Pereira

UMinho | 2022



Universidade do Minho
Escola de Engenharia

Mariana Marques Pereira

Integrating Kinetic and Constraint-Based Models of Metabolism

Master's Dissertation
Master in Bioinformatics

Work carried out under the guidance of
Professor Miguel Rocha
Professor Vítor Pereira

October of 2022

Direitos de autor

Licença concedida aos utilizadores deste trabalho

Agradecimentos

Em primeiro lugar, gostaria de agradecer à minha família pelo suporte que me deram nesta viagem, bem como pelos conselhos para tornar esta dissertação o melhor que poderia ser.

Em segundo lugar, gostaria de agradecer ao aluno de douturamento Diego Troitiño Jordedo do Instituto de Investigação Marinha (IIM-CSIC) em Vigo, Espanha por toda a ajuda e conselhos na realização desta tese. Também gostaria de agradecer ao Dr. Lucas Carvalho da Universidade de Campinas, Brasil por todo o tempo, ajuda, conselhos e paciência no desempenho desta tarefa.

Agradeço ainda ao estudante de mestrado e colega de sala João Monteiro pelo companheirismo e amizade demonstrada nesta caminhada.

Finalmente, gostaria de deixar um agradecimento especial ao professor Vítor Pereira do grupo BIOSYSTEMS por toda a ajuda, compreensão e disponibilidade demonstrada ao longo deste percurso e pelos conhecimentos transmitidos em programação e algoritmos ao longo deste mestrado. Por último um sincero, agradecimento ao professor Miguel Rocha do grupo BIOSYSTEMS por me propor realizar esta tese e por todo o conhecimento transmitido durante este mestrado.

Muito obrigado a todos!

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Abstract

Mathematical models are fundamental tools for explaining biological behaviors. Dynamical and constraint-based models are two different formulations that attempt to capture the phenotypic capabilities of organisms.

Dynamic models are formulated as ordinary differential equations (ODEs) that simulate metabolic concentration over time. These models, however, only depict changes in metabolic concentration and rely on mechanistic details and kinetic parameters that are not always available. Constraint-based models, on the other hand, have a better cellular perspective. By performing constraint-based optimizations, they simulate cell behavior under different genetic and environmental conditions. Metabolic models also have some drawbacks. In addition to providing no mechanical knowledge of any chemical reactions (beyond their stoichiometry) and no information regarding metabolic concentrations or reaction flux dynamics, they are based on a steady-state assumption that production and consumption of metabolites are balanced within the cell. Constraint-based optimizations, Flux Balance Analysis (FBA) methods, generally return an infinite set of solutions, requiring the imposition of additional assumptions to identify unique flux distributions.

While individually, both modeling approaches have several advantages, one lacks the benefits provided by the other. With this in mind, we implemented a tool in MEWpy capable of hybridizing kinetic and constraint-based models. With it, we were able to reduce the constraint-based model solution space by overlapping the kinetic solution space and sampling the kinetic model, analyze the impact of different standard deviation values on the sampling, perform hybridization of enzymatic constrained models, and further compare distinct hybridization approaches. To demonstrate the potential of our tool and its applicability in strain optimization, we performed hybrid optimization of succinate production, where we discovered a set of genetic mutations that boosted its production.

Keywords: Constraint-Based Model, Kinetic Model, Hybrid model, MEWpy, Sampling, Hybrid Simulation, FBA, Succinate Production, GECKO.

Resumo

Os modelos matemáticos são ferramentas fundamentais para explicar os comportamentos biológicos. Os modelos dinâmicos e com base em restrições são duas formulações diferentes que tentam captar as capacidades fenotípicas dos organismos.

Os modelos dinâmicos são formulados como equações diferenciais comuns (ODEs) que simulam a concentração metabólica ao longo do tempo. Estes modelos, contudo, apenas retratam as alterações da concentração metabólica e dependem de detalhes mecanicistas e parâmetros cinéticos nem sempre disponíveis. Os modelos com base em restrições, por outro lado, têm uma melhor perspetiva celular. Ao efetuarem otimizações baseadas em restrições, simulam o comportamento celular sob diferentes condições genéticas e ambientais. Os modelos metabólicos também têm alguns inconvenientes. Para além de não fornecerem qualquer conhecimento mecânico de quaisquer reações químicas (para além da sua estequiometria) e nenhuma informação relativa a concentrações metabólicas ou dinâmicas de fluxos de reação, baseiam-se numa suposição de estado estável de que a produção e consumo de metabolitos são equilibrados dentro da célula. As otimizações baseadas em restrições, métodos de Análise de Equilíbrio de Fluxo (FBA), devolvem geralmente um conjunto infinito de soluções, exigindo a imposição de pressupostos adicionais para identificar distribuições de fluxo únicas.

Embora individualmente, ambas as abordagens de modelação tenham várias vantagens, uma carece dos benefícios proporcionados pela outra. Com isto em mente, implementámos uma ferramenta em MEWpy capaz de hibridizar modelos cinéticos e baseados em constrangimentos. Com ela, conseguimos reduzir o espaço de solução do modelo baseado em restrições através da sobreposição do espaço de solução cinética e da amostragem do modelo cinético, da análise do impacto de diferentes valores de desvio padrão na amostragem, da realização da hibridação de modelos com restrições enzimáticas, e ainda da comparação com mais abordagens de hibridação distintas. Para demonstrar o potencial da nossa ferramenta e a sua aplicabilidade na otimização de estirpes, realizámos a otimização híbrida da produção de succinato, onde descobrimos um conjunto de mutações genéticas que impulsionaram a sua produção.

Palavras-chave: Modelos com Base em Restrições, Modelos Cinéticos, Modelos Híbridos, Amostragem, Simulação Hibrida, FBA, Produção de Succinato, GECKO.

Table of Contents

List of Figures

List of Tables

Metabolite Acronyms

2-3PG 2-Phosphoglycerate + 3-Phosphoglycerate.

2PG *2-Phosphoglycerate*.

3PG *3-Phosphoglycerate*.

6PG *6-Phosphogluconate*.

ACCOA Acetyl-Coenzime A.

ADP *Adenosin Diphosphate*.

AMP *Adenosin Monophosphate*.

ATP *Adenosin Triphosphate*.

cAMP *Cyclic AMP*.

CGLCex Extracellular Glucose.

DHAP *Dihydroxyacetone Phosphate*.

E4P *Erythrose-4-Phosphate*.

F6P *Frutose-6-Phosphate*.

FDP *Frutose-1,6-Biphosphate*.

FUM Fumarate.

G1P *Glucose-1-Phosphate*.

G6P *Glucose-6-Phosphate*.

GAP *Glyceraldehyde-3-Phosphate*.

MAL Malate.

NAD *Diphosphopyridine Nucleotide, oxidized*.

NADH *Diphosphopyridine Nucleotide, reduced*.

NADP *Diphosphopyridine Nucleotide-Phosphate, oxidized*.

NADPH *Diphosphopyridine Nucleotide-Phosphate, reduced*.

PEP *Phosphoenolpyruvate*.

PGP *1,3-Diphosphoglycerate*.

PYR *Pyruvate*.

RIB5P *Ribose-5-Phosphate*.

RIBU5P *Ribulose-5-Phosphate*.

SED7P *Sedoheptulose-7-Phosphate*.

XYL5p *Xylulose-5-Phosphate*.

1 Introduction

1.1 Context and Motivation

Systems Biology and Bioinformatics tools enhance biological and biomedical research by analyzing relevant data and properties (e.g., genome sequencing) that can result in model-driven discoveries. This possibility has captivated the construction of genome-scale networks, enabling *in silico* simulations of complex biological systems and a greater understanding of how metabolic flux distributions may change in a specific biological network to predict cellular phenotypes (McCloskey, Palsson, & Feist, 2013).

By 2011, more than 50 genome-scale network reconstructions had been performed thanks to the continued advancement of genome sequencing (Oberhardt, Puchałka, Martins dos Santos, & Papin, 2011), and this number continues to grow to this day (Mendoza, Olivier, Molenaar, & Teusink, 2019). These models simulate cellular behavior under different genetic and environmental conditions through constraint-based optimizations, taking only their stoichiometry and reversibility of reactions into account (Machado, Costa, Ferreira, Rocha, & Tidor, 2012; Stephanopoulos, Aristidou, & Nielsen, 1998). The analysis of these models is based on the steady-state assumption whose solutions, based on constraints, is underdetermined. This results in a broad spectrum of solutions (called a flow cone) that requires additional conditions to identify unique solutions and predictions. This often presents itself as an optimization of a particular hypothesis, such as optimal biomass growth for wild-type (Edwards & Palsson, 2000) or minimizing cellular adjustments for knock-out strains (Segre, Vitkup, & Church, 2002; Shlomi, Berkman, & Ruppin, 2005). Despite this, these models cannot express the transient behavior of organisms. For example, methods were developed in simulations of fermentation profiles to integrate external concentration variations while maintaining the assumption of an internal pseudo-steady-state (Oddone, Mills, & Block, 2009; Leighty & Antoniewicz, 2011).

On the other hand, there is another type of model whose formulation differs entirely from the constraint-based models. Kinetic models of metabolism are predictors of temporal metabolic behavior of living organisms. They use prior knowledge, such as experimental data and enzyme mechanisms to construct a dynamical system that follows metabolic dynamics using ordinary differential equations (ODEs). These ODEs have all the necessary initial system values, such as metabolic concentrations, reaction rate equations, and kinetic parameters (Kim, Rocha, & Maia, 2018). However, dynamic models only depict changes in metabolic concentration and rely on mechanistic details and kinetic parameters that are not always available. Because of such problems, these models are only performed on well-known small-scale organisms such as *Saccharomyces cerevisiae* (Rizzi, Baltes, Theobald, & Reuss,

1997) and *Escherichia coli* (Chassagnole, Noisommit-Rizzi, Schmid, Mauch, & Reuss, 2002) and only consider a reduced set of metabolic pathways.

Although these two types of models have completely different formulations, one lacks the benefits provided by the other and vice versa. Hence, there is a growing interest in creating a hybrid system that can combine both formulations to have the best of both worlds. Several attempts have been made to combine both formulations (e.g. (Yugi, Nakayama, Kinoshita, & Tomita, 2005; Moulin, Tournier, & Peres, 2021)), however there are still large avenues to be explored that might help to shed some light on the dynamics of metabolic networks.

1.2 Objectives and Motivation

In order to obtain more robust strains, the main goal of this master thesis is to create a Python tool capable of implementing hybrid systems by combining kinetic and metabolic models. The approaches explored in this work integrate information retrieved from kinetic simulation into constraint-based metabolic formulations. To achieve this, we will:

- survey the state-of-the-art on different approaches to hybridize kinetic and constraint-based formulations;
- validate and explore the dynamic simulation capabilities of MEWpy, a metabolic engineering workbench developed at the Center for Biological Engineering of the University of Minho;
- propose and implement novel hybrid mechanisms able to integrate kinetic information into constraint-based models by reducing the boundaries of the steady-state solution space;
- explore the new hybrid simulation approaches in strain design optimization tasks for the increased production of targeted compound. In particular, this work, will make use of *Escherichia coli* models to optimize the production of Succinate.

1.3 Text Organization

The document is organized in five chapters. A concise description of the chapters follows. Chapter 1 has a brief description of the context, motivation and goals of this work. Chapter 2 has the state of the art of this master thesis. Here we go into more depth of dynamic, metabolic and hybrid models as well as strain optimizations and MEWpy. Chapter 3 contains the methods of this work. Here is addressed the kinetic and metabolic models used, as well as, the hybrid simulation and hybrid optimization approaches.

Chapter 4 presents the results and their discussion. Here we discuss the work that has been done and the examples that highlight the capabilities of our tool. Finally, Chapter 5 contains the conclusions and highlights future directions.

2 Mathematical models

In order to explain biological systems, biologists resort to creating models that demonstrate what happens in nature. However, these models turn out to be abstract models of reality since they focus on the objective of the study and ignore other aspects that, although relevant, are not of great interest in that model. We can take as an example a ball-and-stick model of chemical structure. These models are intended to represent molecular chemical bonds. However, other relevant information, such as the polarity resulting from these bonds in the atoms of the molecule, is not present in these models as it diverges from its primary goal (Ingalls, 2013). Depending on the purpose of the model, biologists can build models of greater or lesser complexity, such as model-organisms or ball-and-stick models, respectively. Despite this, most existing representations of our knowledge of cellular processes come from illustrating conceptual models in diagrams of interactions. Figure 1 is an example of an interaction diagram.

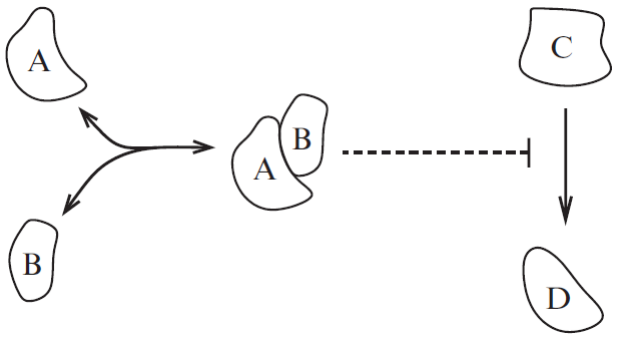


Figure 1: Interaction diagram of a regulatory interaction in which the complex A-B is not consumed. Molecular species A and B reversibly bind into molecular complex A-B which inhibits the conversion of specie C into specie D (Ingalls, 2013).

Nevertheless, these models are ambiguous about the system's behavior, especially when the network of interactions is reversible. By using mathematical descriptions of the system, we can solve these ambiguous problems since each system interaction will have a quantitative representation in the model (Ingalls, 2013).

Nowadays, mathematical models are used to understand both cells' inner and outer interactions and to predict and optimize the properties and behaviors of cell factories, which are of great importance in the biotechnological industry. Nonetheless, despite several studies on metabolic systems, this is still a very complex and challenging matter to understand.

Achieving a desired metabolic state by using a genetically modified strain is a difficult and poorly understood endeavor. Subsequently, to be able to reach this metabolic state, numerous modifications

are needed in order to reach the most desirable productivity. However, being able to pinpoint the most appropriate set of genetic changes from an immense number of combinations is very difficult (Ohno, Shimizu, & Furusawa, 2014). For this, *in silico* tools based on computational simulation and mathematical modelling have been created to overcome this problem. They enable the screening of the best combination of genetic alterations capable of increasing the target productivity (Patil, Rocha, Förster, & Nielsen, 2005; von Kamp & Klamt, 2014).

Currently, there are two main distinct model approaches in metabolic modelling based on different representations of metabolisms used in the study of cellular metabolisms: constraint-based and kinetic modelling (Llaneras & Picó, 2008).

2.1 Kinetic models

Kinetic models, also identified as dynamic models, have the potential to predict the metabolic network dynamics over time accurately (Soh, Miskovic, & Hatzimanikatis, 2012; Heinrich & Schuster, 1998). Due to this capability, they are considered of great importance in biotechnology. In this industry, they can be used to improve the ability to produce a desirable product in cell factories, determine to which degree the properties of a model can change, improve a product quality or process design and optimize problems (von Kamp & Klamt, 2014; Phair & Misteli, 2001).

These models are based on previous knowledge, such as experimental data and enzyme mechanisms, which is used to build a dynamic system that tracks metabolic dynamics using ordinary differential equations (ODEs). These have all initial values of the system, such as metabolite concentrations, reaction rate equations and kinetic parameters (Kim et al., 2018). Nevertheless, dynamic models possess some disadvantages that are worthy of mentioning, such as i) the unavailability of all data required to build a dynamic model in the literature since they depend on mechanistic details and kinetic parameters; ii)the fact that it is computationally intensive especially when realizing parameter estimations, iii) the fact that dynamic models only portray changes in metabolite concentration and iv) the time-consuming task of parameterizing large scale dynamic models (Dromms, Lee, & Styczynski, 2020; Kim et al., 2018; Pettigrew, Resat, & Petzold, 2009; Smallbone, Simeonidis, Swainston, & Mendes, 2010; Smallbone et al., 2013). Due to such problems, these models are only performed on well-known organisms on a small scale, like *Saccharomyces cerevisiae* (Rizzi et al., 1997) and *Escherichia coli* (Chassagnole et al., 2002).

2.1.1 Reaction Kinetics

Under kinetic models, a network of mechanistic ODEs exists to describe the evolution of metabolite concentrations and metabolic fluxes over time in response to unsteady dynamics (Klipp, Herwig, Kowald, Wierling, & Lehrach, 2005). This permits the study of the metabolic responses to perturbations over-time before reaching steady-state (Smallbone et al., 2010). A differential equation of the kinetics of metabolic systems is formulated as

$$\frac{dX}{dt} = F\left(K, X(t)\right) \tag{2.1}$$

where K is a vector of parameters and F is a vector field. The stated system ought to be defined with the initial conditions:

$$\frac{dX}{dt} = Sv(X, P) - \mu X; X(0) = X_0 \tag{2.2}$$

where X denotes the vector of metabolite concentrations, X_0 the vector of initial concentrations, S the stoichiometric matrix, v the vector of reaction rates and P kinetic parameters. The reaction rates v vary according to metabolite concentrations X and kinetic parameters P.

After determining the network interaction mechanism and the type of kinetic rate expressions used in the dynamic model, can we choose the best mathematical foundation of the model (Kim et al., 2018). Two hypotheses are the Mass Action kinetics and the Michaelis-Menten kinetics (irreversible, reversible and with different types of inhibition). Both will be briefly addressed below, where the type of enzyme kinetics can change the reaction rate vector v.

Mass Action Kinetics

One possible mathematical foundation for kinetic models is the Mass Action Law being this law the foundation of many other mathematical formulations used in kinetic models. Guldberg and Waage were the first ones to introduce this mathematical model (Waage & Guldberg, 1986) that describes and predicts the solution's behavior in a dynamic equilibrium. In general chemistry, this law is used to describe how chemical equilibrium can be altered by adding reactants/products to the mixture if more products/reactants are needed (Turányi & Tomlin, 2014). If we consider, for example, the reaction:

$$2A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} S$$

the final reaction rate vector is

$$v_f = k_1 A^2 - k_{-1} S \tag{2.3}$$

with the equilibrium constant

$$K_{eq} = \frac{k_1}{k_{-1}} \tag{2.4}$$

(Jamshidi & Palsson, 2010) where k_1 and k_{-1} represent the kinetic constants. In this example, the mass action law explains that, in reaction direction k_1, 2 reagents A will originate 1 product S (equation 2.5) and, in reaction direction k_{-1}, 1 reagent S will originate 2 products A (equation 2.6).

$$2A \xrightarrow{k_1} S \tag{2.5}$$

$$S \xrightarrow{k_{-1}} 2A \tag{2.6}$$

Michaelis-Menten Kinetics

Other possible mathematical foundation for kinetic models is the Michaelis-Menten kinetics. However, depending on the nature of the problem, a different type of approach can be used. Three of those approaches are Irreversible Michaelis-Menten, Reversible Michaelis-Menten and Michaelis-Menten regarding different types of inhibition.

In 1913, Leonor Michaelis and Maud Leonora Menten published their most known work in enzyme kinetics, showing that enzyme-catalyzed reaction rates are proportional to enzyme-substrate complex concentrations (Johnson & Goody, 2011; Michaelis & Menten, 1913). In this work, they do not consider what happens in practice, the reversibility of biochemical reactions. The fact that enzyme reactions may happen in both directions is the main difference between Irreversible and reversible Michaelis-Menten formulations.

Table 1: Reaction rate and kinetic scheme of Irreversible Michaelis-Menten vs Reversible Michaelis-Menten.

	Irreversible Michaelis-Menten	Reversible Michaelis-Menten
Kinetic scheme	$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P$	$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E + P$
Rate equation	$v = \frac{V_{max} S}{S + K_m}$	$v = \frac{V_{max}(S - \frac{P}{K_{eq}})}{K_s(1 + \frac{P}{K_p}) + S}$

In Table 1 are the reaction rate formula of both formulations and their respective kinetic scheme (Michaelis & Menten, 1913; Johnson & Goody, 2011; Keleti, 1986; Feng, Fu, & Sun, 2010). In there, v is the reaction rate, S is substrate concentration, V_{max} is maximum reaction velocity (described in the

equation 2.7), k_{cat} is the catalytic constant, E is the enzyme concentration, K_m represents the Michaelis-Menten constant (defined in equation 2.8), P represents the products, K_s and K_p represent the Michaelis-Menten constants of substrate and product, respectively, K_{eq} is the equilibrium constant and finally, k_1 and k_{-2} are second order rate constants and k_{-1} and k_2 are first-rate constants.

$$V_{max} = k_{cat} E_{total} \tag{2.7}$$

$$K_m = \frac{k_{-1} + k_{cat}}{k_1} \tag{2.8}$$

One needs to make measurements at different initial rates for different substrate concentrations in order to be able to evaluate the parameters V_{max} and K_m in a single enzyme. Knowing that the rate is a non-linear function of substrate concentration, there are only two possible options to determine the values of the parameters. One option is to maintain the non-linear rate function and determine these parameters using a non-linear regression. The other option is to transform the reaction rate equation into a linear equation and then apply a linear equation. This last option is particularly advantageous in reading the parameter values, since these values can be read more or less directly from the graph obtained in the linear regression. Some linear approaches are summarized in Table 2

Table 2: Resume of the linear approaches Lineweaver-Burk, Eadie-Hofstee, Hanes-Woolf (adapted from Klipp et al. (2005)).

Linearization	Transformed equations	New Variables
Lineweaver-Burk	$\frac{1}{v} = \frac{k_m}{V_{max}}\frac{1}{S} + \frac{1}{V_{max}}$	$\frac{1}{V'}, \frac{1}{S}$
Eadie-Hofstee	$v = V_{max} - K_m \frac{v}{S}$	$v, \frac{v}{S}$
Hanes-Woolf	$\frac{S}{v} v = \frac{S}{V_{max}} + \frac{K_m}{V_{max}}$	$\frac{S}{V'}, S$

In the last Michaelis-Menten approach there are three distinct types of inhibition: competitive inhibition, non-competitive inhibition and uncompetitive inhibition (considered no inhibition in Figure 2) (Chou & Talaly, 1977).

In competitive inhibition, the substrate and the inhibitor compete to bind with the enzyme. In this type of inhibition, the enzyme cannot bind with the substrate and the inhibitor simultaneously. It only binds

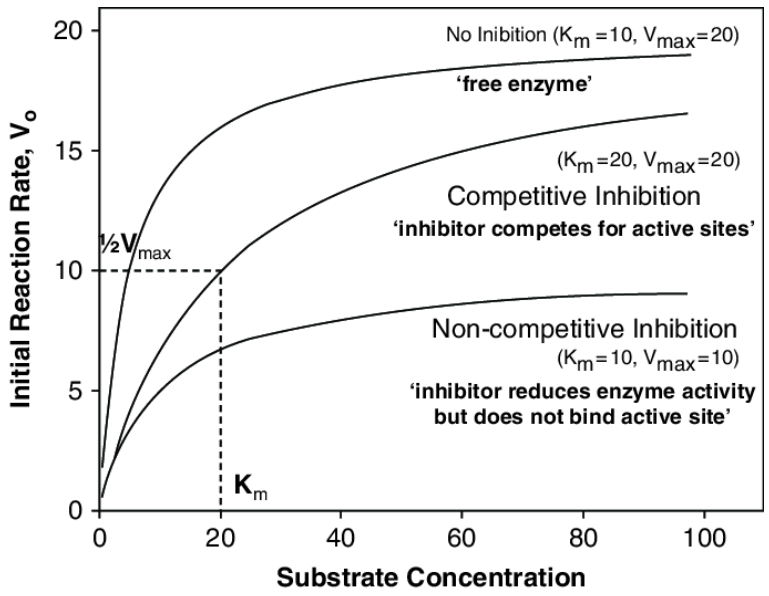


Figure 2: Plot of the Michaelis-Menten model's inhibition types showing the predicted kinetic changes with enzyme inhibition (Zimmerman & Ahn, 2010).

with one of them (substrate or inhibitor) or with neither (Cornish-Bowden, 1974). The temporary binding of organic matter to an enzyme's active site can be analogous to this type of inhibition (Zimmerman & Ahn, 2010).

In non-competitive inhibition, an inhibitor can only bind either with an enzyme or an enzyme-substrate complex, therefore lowering enzymatic efficiency (Zimmerman & Ahn, 2010; Chou & Talaly, 1977). This can be compared to an enzyme conformational change, which occurs when interacting with minerals or organic matter.

Finally, in uncompetitive inhibition, the inhibitor only binds to an enzyme-substrate complex. This can only happen because of the conformational change that occurs in the complex after the substrate binds to the enzyme. This conformational change will open a new binding site that allows the inhibitor to bind.

All three types of inhibitions are applied in reversible and irreversible Michaelis-Menten mechanisms. In competitive inhibition, the V_{max} remains unchanged while K_m values change. Opposite to this is the non-competitive inhibition where the reverse happens, K_m remain unchanged, and V_{max} values change. Lastly, both V_{max} and K_m values change in uncompetitive inhibition (Zimmerman & Ahn, 2010; Klipp et al., 2005).

2.2 Constraint-based models

Another mathematical representation that can describe metabolic network behavior is the constraint-based models. This stationary representation possesses a better cellular perspective when compared to dynamic models that have a better intracellular insight.

Metabolic models assume that the system is working in equilibrium. With only information on reaction stoichiometry and directionality, these models are able to perform predictions that are usually close to experimental observations (Kim et al., 2018). There are two methodologies to analyze stoichiometric models: when in conjunction with experimental measurements (Metabolic Flux Analysis) or *in silico* (Flux Balance Analysis).

Nevertheless, metabolic models have some drawbacks. Besides providing no mechanical knowledge of any chemical reactions (beyond their stoichiometry) and no information regarding metabolite concentrations or reaction fluxes dynamics, they rely on a steady-state assumption that the production and consumption of metabolites are balanced inside the cell. Also, the constraint optimizations usually return an infinite set of solutions, requiring the imposition of additional assumptions to pinpoint unique flux distributions.

Besides all this, one crucial fact remains. Due to the vast amount of stoichiometric information obtained from genetic annotations, several large-scale metabolic reconstructions are available compared with kinetic modelling. The unique characteristics of constraint-based modelling allow the construction of large-scale metabolic representations more efficiently and less time-consuming when in comparison to dynamic modelling.

2.2.1 Flux Balance Analysis

One methodology used to analyze metabolic models is Flux Balance Analysis (FBA). This technique has a more *in silico* approach since by only using the structural information of the system, it can predict biological behavior. Generally, the systems analyzed by this methodology are under determined (more unknown fluxes than equations). This means that the solution returned is not a number but a space of feasible flux distributions. In table 3 are presented some optimization methods used by this methodology and respective characteristics that return a space of feasible flux distributions.

FBA

For the simulation of phenotypes, the most common approach is flow balance analysis (FBA). This approach tries to find a unique solution for an under determined system (Varma & Palsson, 1994). Through a linear programming problem, it is possible to determine the particular metabolic flux distribution where the purpose is to maximize an objective function Z. This maximization can be subject to stoichiometric and capacity constraints.

$$max\ Z = \sum_{i=1}^{r} c_i v_i, \qquad \text{subject to: } Sv = 0;\ lb < v < ub \tag{2.9}$$

where c_i contains the vector weights for all individual rates.

FBA can solve many different problems, each with its respective objective function. Some examples of such problems are: maximization of biomass; determination of optimal energy efficiency through maximization/minimization of ATP production; minimization of protein distribution with an efficient channeling of metabolites in metabolic pathways (this is called minimization of absolute fluxes); maximization of the production of a particular important metabolite, among several others.

For studies demonstrating several perturbations such as genetic manipulations or growth of organisms on different substrates, FBA is a good method to obtain quick results for the different perturbations. However, one of the major disadvantages of this method is that it does not take into account the kinetics of the reactions, it ends up having limitations. This means that this method cannot predict metabolic concentrations and does not consider regulatory effects, so the predictions are not always accurate (Orth, Thiele, & Palsson, 2010).

FVA

Flux Variability Analysis (FVA) is an optimization method that determines the maximum and minimum flux values that satisfy the model constraints while maintaining the optimal objective value (equation 2.10), despite rarely having unique optimal solutions. FVA can be performed independently, such as any other optimization method, or after a FBA. The purpose of performing an FVA after a FBA is to analyze the FBA results, since several constraint-based models were identified as having unboundedness of the optimal flux space. This was mainly identified in genome-scale metabolic models (Müller & Bockmayr, 2013).

$$\begin{aligned} max_v &= v_i \\ min_v &= v_i \\ \text{subject to: } Sv &= 0;\ lb\ <\ v\ <\ ub \end{aligned} \tag{2.10}$$

The applications of this method are diverse, such as studying flux distributions under suboptimal conditions or determining reactions that are not necessary or blocked, among others (Gudmundsson & Thiele, 2010; Müller & Bockmayr, 2013).

MOMA

The prediction of phenotypic effects resulting from metabolic gene deletions is one of the applications of FBA. Here, the model will assume that the mutant strain exhibits an optimal metabolic state. The problem is that the artificially generated mutants have not been exposed to the wild strains nor to the same evolutionary pressure, and therefore flux regulation will not be done immediately to achieve optimal growth.

In an attempt to better understand the reaction rates of mutants, the metabolic adjustment minimization method (MOMA) has been proposed. This method assumes, by stoichiometry, a steady state with constraints. However, it is less demanding when it comes to optimal growth (Segre et al., 2002). So, in initial conditions, a mutant will present a suboptimal flux distribution, being an intermediate between a wild-type strain and an optimal mutant.

The distance minimization in flux space is the optimization problem that MOMA presents. In this approach we have a quadratic function defining the distance minimization, as we can see in the following equation:

$$min \parallel v_w - v_m \parallel^2 \tag{2.11}$$

where v_m and v_w represent the flux distribution of the mutant strain and the flux distribution of the wild strain, respectively. Compared with the most common formulation of FBA, in MOMA the objective function does not clearly depend on the biomass production.

ROOM

Another method based on constraints predicting the metabolic steady state in case of gene knockout is ROOM (Regulatory on/off minimization). This method aims at minimizing of changes in the fluxes of a

wild strain through a Mixed Integer Linear Programming (MILP) formulation (Shlomi et al., 2005).

What this method does is to find a mutated strain that can satisfy both stoichiometrically and the capacity constraints of the fluxes so that it is possible to minimize the total number of fluxes in a wild strain to obtain the desired product. The heuristic distance metric underlying ROOM is motivated by the assumption that the cell minimizes its flux changes after gene knockouts, being this regulatory changes parsimoniously described by Boolean on/off dynamics.

Below will be described the principles behind stoichiometric models.

Stoichiometric principles

As mentioned before, constraint-based models only require information about the system's reaction network. The reaction network structure of a system usually is characterized by two different sets (metabolites and reactions) and a stoichiometric matrix. A reaction network is represented in Figure 3. This network comprises 5 internal metabolites and 9 reactions, where 2 of them are reversible (v_2, v_4). As visible in the network, all internal species (A, B, C, D, E) must be distinguishable from the external species (C_{ext}, E_{ext}) because internal metabolites are considered for the mass balance constraints while external metabolites are considered as sources or sinks of the system.

FBA attempts to solve a problem of order $m \times n$, where m and n represent the number of metabolites and network reactions, respectively (matrix S). It does this by generating a vector space based on the limiting or non-limiting equations. As shown in the Figure 3, for each metabolite the consumption and production reactions are identified. Given its stoichiometry, when the metabolite is consumed in the reaction r_i, its value is negative, and if produced, its value is positive.

The application of the pseudo-steady-state assumption (2.12) to the mass balance of internal metabolites (2.13) leads to the fundamental metabolite balancing equation (2.14).

$$S \times v = 0 \tag{2.12}$$

$$\frac{d_x}{d_t} = S \times v \tag{2.13}$$

$$\frac{d_x}{d_t} = S \times v = 0 \tag{2.14}$$

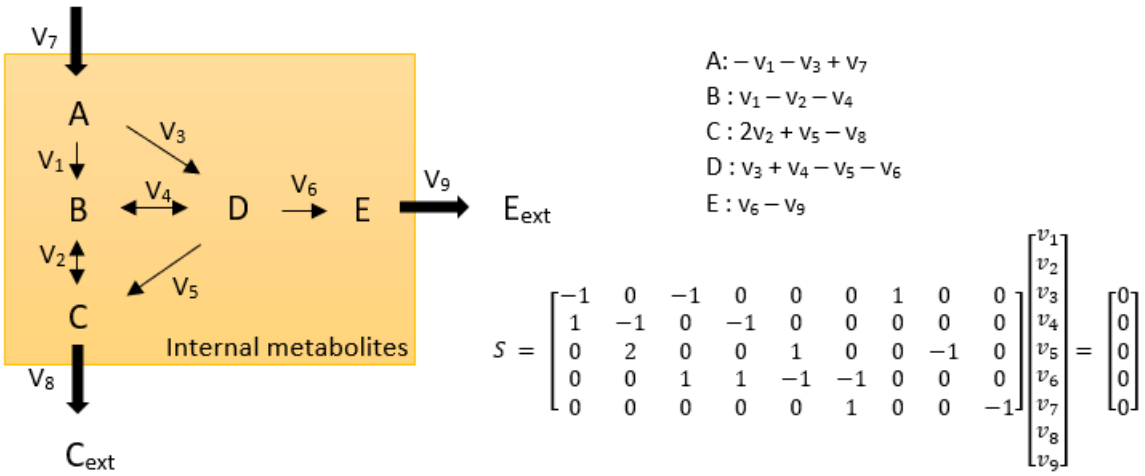


Figure 3: Mathematical representation, in the stoichiometric matrix, of the metabolic network.

The linear equation system resulted from equation 2.14 and described by the stoichiometric matrix S demands that the production and consumption must be zero. In other words, the sum of all numbers presented in the stoichiometric matrix S must be zero.

Thus, equation 2.14 defines the region where a possible flow distribution exists. However, this does not yield a single solution since there are generally more reactions (columns) than metabolites (rows) in the stoichiometric matrix, which ultimately makes sense since cells may behave differently when, for example, changing environmental conditions occur (Llaneras & Picó, 2008). Assuming that all rows in S are linearly independent and that there are m independent equations, we can obtain the number of degrees of freedom (2.15).

$$F = n - m \tag{2.15}$$

2.2.2 Metabolic Flux Analysis

Other methodology used to analyze metabolic models is the Metabolic Flux Analysis (MFA). MFA is a technique capable of calculating intracellular fluxes within an assumed stoichiometric network, where measured external rates are accounted as restrains of the system. Glucose uptake rates, growth rates or CO_2 evolution rates can be used as external restraints (Antoniewicz, 2015), reducing the number of degrees of freedom. This reduction can make the model entirely determined or over-determined if sufficient external restraints are measured. The output of the model is a metabolic flux map that shows the diagram of the biochemical reactions included in the calculations, along with an estimate of the flux at each reaction present in that diagram occurs (Klipp et al., 2005).

In MFA, the equation 2.2 is also applied. In it, the dilution term is ignored, resulting in equation 2.16.

$$\begin{aligned} S \times v = 0 \Leftrightarrow \\ |S_i S_e|\,|v_i v_e| = 0 \Leftrightarrow \\ S_i \times v_i + S_e \times v_e = 0 \Leftrightarrow \\ S_i \times v_i = -S_e \times v_e \end{aligned} \tag{2.16}$$

where (S_i) is the intracellular metabolite stoichiometric matrix, (S_e) is extracellular metabolite stoichiometric matrix, (v_i) is intracellular metabolite flux and (v_e) is intracellular metabolite flux. If we consider the metabolic network in Figure 3, $v_i = [\, \mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3, \mathrm{v}_4, \mathrm{v}_5, \mathrm{v}_6]$ and $v_e = [\, \mathrm{v}_7, \mathrm{v}_8, \mathrm{v}_9]$.

2.3 Integration of MET and KIN models

As previously mentioned, constraint-based models and kinetic models are two different types of modulation that can predict biological behavior. At the genome-scale, metabolic models use structural information about the metabolic system to be able to make predictions of behavior, even though, transcendent behavior cannot be inferred. This is no longer the case in kinetic models, where detailed information about interactions at the metabolite and enzyme level (Smallbone et al., 2010) enables the model to be analyzed dynamically. However, it is impossible to know all the kinetic parameters of a large-scale model, which thus makes the task of building genome-scale dynamic models unfeasible.

At this impasse and due to the characteristics of both approaches, several attempts have been made to combine kinetic models and metabolic models into a hybrid system (e.g., (Moulin et al., 2021; Yugi et al., 2005)). One modelling approach that falls within this hybrid system is dynamic flux balance analysis (dFBA).

2.3.1 Dynamic Flux Balance Analysis

By assuming that organisms reach steady-state rapidly, dynamic flux balance analysis (dFBA) allows the simulation of dynamic biological systems when they respond to changes in the extracellular environment. (Gomez, Höffner, & Barton, 2014). It offers a structured model of the biochemical process, in which the reaction pathways alter depending on environmental conditions. This is well represented by changes in the functional dependence on substrates (Höffner, Harwood, & Barton, 2013).

This hybrid system possesses two different optimization approaches that: the Dynamic Optimization Approach (DOA) and the Static Optimization Approach (SOA). The DOA approach looks at the problem as

a whole. It optimizes over the entire time of interest, discarding the FBA steady-state assumption to obtain time profiles of fluxes and metabolite levels (Valverde, Gullón, García-Herrero, Campoy, & Mellado, 2019; Mahadevan, Edwards, & Doyle III, 2002; Dromms et al., 2020). To do this, the problem is transformed into a non-linear problem (NLP), solving it all at once (Mahadevan et al., 2002; Valverde et al., 2019; Dromms et al., 2020). However, the mathematical properties that most characterized FBA are lost (Dromms et al., 2020).

Contrary to the DOA approach, the SOA approach divides the problem into several time intervals, solving this instantaneous optimization problem at the beginning of each time interval. Its mathematical approach is a blend of the linear problem (LP) with an ODE model that repeats the LP throughout the optimization time, making it computationally expensive (but tractable). (Mahadevan et al., 2002; Höffner et al., 2013; Dromms et al., 2020).

Since dFBA is a growing subject, there are several dFBA applications based on genome-scale models. There are small scale models ($<$ 200 fluxes per FBA model and 1-2 compartments) of bacteria (Anesiadis, Cluett, & Mahadevan, 2008; Lequeux et al., 2010; Meadows, Karnik, Lam, Forestell, & Snedecor, 2010; Mahadevan et al., 2002; Varma & Palsson, 1994), *Saccharomyces cerevisiae* yeast (Sainz, Pizarro, Pérez-Correa, & Agosin, 2003; Pizarro et al., 2007; Lee, Gianchandani, Eddy, & Papin, 2008), plants (R.-Y. Luo et al., 2006) and in animals (R. Luo et al., 2009), and genome-scale applications (Vargas, Pizarro, Pérez, & Agosin, 2011).

2.3.2 Hybrid approaches

In order to overcome the difficulties of kinetic and constraint-based models, hybrid models are models that attempt to predict and describe a system using a mechanism that integrates both metabolic and dynamic models. (Yugi et al., 2005; Machado et al., 2012) are some of the works that formulated a hybrid model.

Yugi et al. proposed a hybrid simulated method capable of simulating the metabolism of a cell (Yugi et al., 2005). This hybrid method can reduce the number of biochemical experiments by dividing the metabolic pathways. In some parts, a kinetic-based dynamic method is applied, and at the rest is applied the MFA method.

In 2010, integrative omics metabolic analysis (IOMA) was created. This new method uses quadratic programming to seek a steady-state in which measured proteomic and metabolomic data is kinetically

connected to the large scale model (Yizhak, Benyamini, Liebermeister, Ruppin, & Shlomi, 2010).

Another study made in this area was formulated by Machado et al. (Machado et al., 2012). The authors compared Chassagnole dynamic and metabolic model of *Escherichia coli* central carbon metabolism by randomly sampling kinetic parameters and solving the respective ODE system. These solutions were then mapped into the metabolic model flux bounds to further restrict the solution space. As a result, the authors observed that all kinetic solutions overlapped the metabolic model feasibility space. The authors in (Moulin et al., 2021) after replicating the experiment, analyzed the ODE resorting to Elementary Flux Modes (EFM). EFM are non-decomposable (non-zero) steady-state pathways in metabolic networks that operate as minimal functional units, i.e., they represent a minimal finite set of possible states that can generate all the possible states of the network by using convex combinations (Schuster & Hilgetag, 1994). The authors computed all EFM and analyzed how they were affected by different kinetic parameter values.

2.4 Strain Optimization

The study of computer algorithms to determine the best gene knockdown strategies for achieving better phenotypes is expanding quickly in recent years, mainly based on constraint-based model methodologies, *in vivo* proofs of concept and design tools' predictions (Maia, Rocha, & Rocha, 2016).

By relying exclusively on a solution selected from several existing ones, as are the FBA simulations, the conclusions about their results are incomplete and often ineffective with reality. When performing strain optimization strategies, more factive results are selected using methodologies based on evolutionary algorithms, which apply stochastic search and optimization techniques, inspired by the natural mechanisms of evolution and genetics.

2.4.1 Overview

On the base of the information stored in metabolic models, several constraint-based methods may be used to predict phenotypic behavior. The flux hypercone of permitted flux distributions is defined as the point at where the available biological restrictions (such as steady state, reversibility, and flux capacity) connect. Under normal (wild-type) settings, the assumption of maximal growth is reasonable, but when the organism is subjected to genetic perturbations, such as when modelling the phenotypes of gene deletion mutant strains, it is hotly contested.

Computational strain optimization methods (CSOMs) may be viewed as methods that attempt to provide a realistic response to a subject or group of questions pertinent to strain design. These issues may be

expressed mathematically and can be solved using various optimization techniques. These methods automatically or semi-automatically look for solutions to problems like which genes should be removed from the model to couple the production of compound versus to growth, powered by phenotype prediction methods and guided by genome-scale metabolic models (GSMMs) (Maia et al., 2016). Gene deletion, gene over- or under expression, heterologous insertion, and, more recently, cofactor specificity modulation are the CSOMs' most frequent tasks, as shown in Figure 4. The brief description of each methodology is found in the next paragraphs.

Gene deletion: The majority of the time, *in silico* CSOMs that take gene deletion into consideration (Fig. 4 A) look for combinations of metabolic function suppression that produce favorable phenotypes. Commonly, to complete this task, constraints that restrict the flux of the disabled reactions to zero are imposed, the fluxes over those reactions are prevented, and the effect of the perturbation is then assessed.

Heterologous insertion: Adding non-native functions by the addition of genes or pathways may increase the metabolic flexibility of ideal hosts, either by increasing the yields of native molecules or by enabling the synthesis of wholly new ones (Fig. 4 B). For the intended functionality, algorithms with this kind of capabilities typically search databases of balanced responses and attempt to reconcile them with the original network.

Gene over- or under expression: Using promoter libraries or synthetic biology methods, it is possible to fine-tune enzyme levels and related fluxes rates in order to control gene over- or under expression. This strategy can be helpful in cases when gene deletion is fatal but down regulation is not, as well as a way to get around flux bottlenecks in some processes leading to a desired biological function (Fig. 4 C).

Modulation of cofactor binding specificity: A novel method is to modulate the cofactor binding specificities to address the paucity of key cofactors necessary for critical stages in various metabolic engineering attempts. This technique may be computationally simulated by switching the cofactor specificities of particular reactions in the network, then using a phenotypic prediction tool to assess the consequences of the perturbations. Because of their role in catabolic and anabolic processes, modulating NAD(H) or NADP(H) availability is a common example (Fig. 4 D).

Finding the ideal collection of genes (or reactions) to inactivate (or over/under express) in order to maximize a set of objective functions is the challenge as it pertains to strain optimization (Maia, Rocha, & Rocha, 2013). Because they provide the optimization aims and objectives, these objective functions

are the most significant user input, e.g. the overproduction of a particular chemical of interest while maximizing the flow of the reaction.

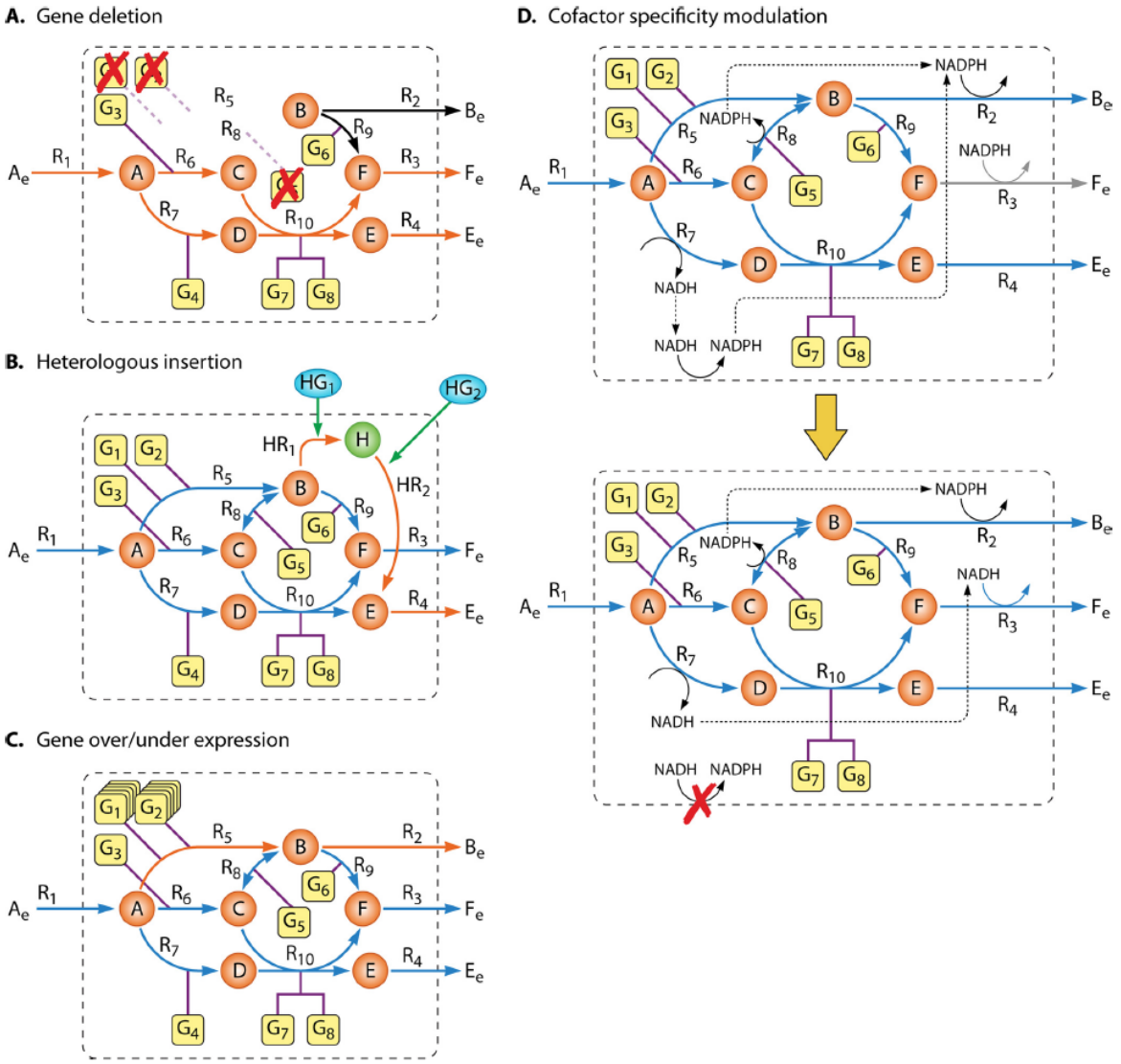


Figure 4: Computational strain optimization tasks. In the illustration, three gene deletions force the flux through reactions that produce the desired compounds (A), two heterologous genes are included to produce an intermediary compound and then excrete the desired product (B), two enzymes are over expressed to produce excess amounts of compound B, which is then excreted (C), and the enzyme catalyzing transport reaction R3 is replaced by a heterologous enzyme using NADH (D). An excess of NADH is produced when a membrane oxidoreductase enzyme is deleted, and this extra NADH can be utilized by the new transport process to eliminate chemical F (Maia et al., 2016).

2.4.2 MEWpy

Given the lack of integrative tools for the increasing number of modelling approaches, the Centre of Biological Engineering (CEB - UMinho) developed a python package called MEWpy (Pereira et al., 2021), which provides ways to investigate several kinds of constraint-based models, such as those with metabolic, enzymatic, or regulatory constraints. Different phenotypic prediction techniques may be performed using MEWpy to help strain optimization.

Overview

MEWpy seeks to offer a Python implementation of Computational Strain Optimization (CSO) algorithms that can be used with both the previously mentioned improved modeling techniques as well as GSMMs creating Gene-Protein-Reaction (GPR) relationships. MEWpy's conceptual architecture, which is shown in Figure 5.

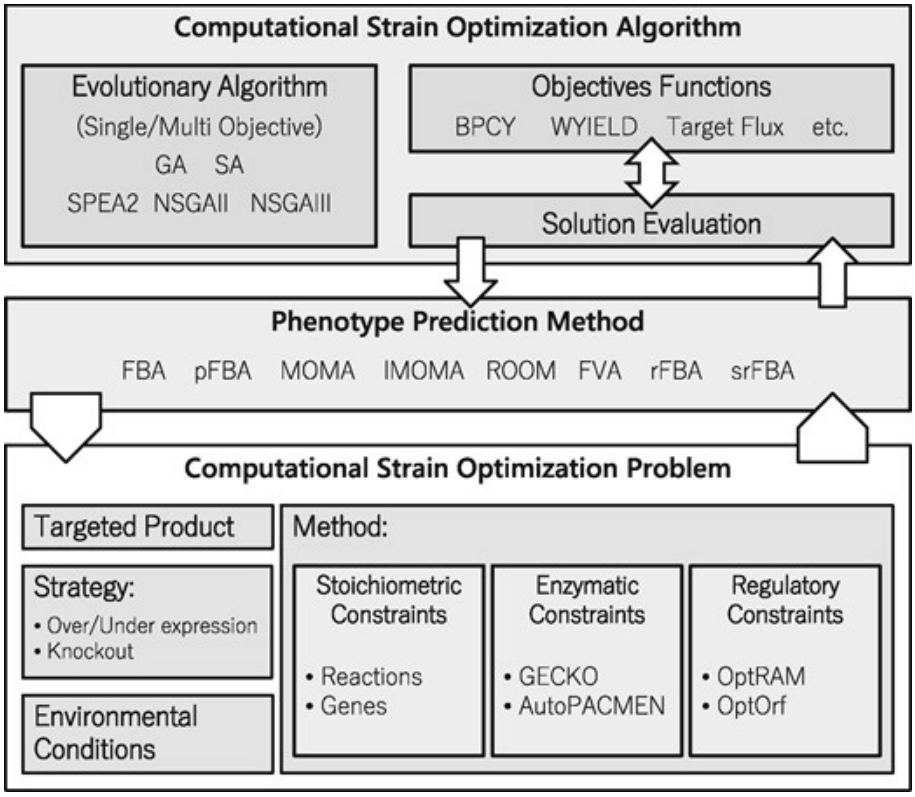


Figure 5: Conceptual architecture of the MEWpy framework. (Pereira et al., 2021).

MEWpy has three main layers: (i) Problem definition layer; (ii) Phenotype simulation layer; and (iii) Optimization layer. The next paragraphs describe each one of them.

Problem definition layer: the layer that defines the problem for the CSO problem includes the modeling framework that has been chosen, the definition of the modification targets (reactions, genes, proteins, or regulatory variables present in the model), the modification strategy (deletion, over-/under expression), along with the target product and the environmental conditions.

Phenotype simulation layer: the techniques used to assess the mutant strains produced by the CSO algorithm are gathered in this layer. These techniques include FBA, pFBA, ROOM, MOMA and its linear version (lMOMA), as well as FVA. To add transcriptional regulation, MEWpy also implements the regulatory FBA (rFBA) and the steady-state rFBA (srFBA) (Pereira et al., 2021).

Optimization layer: the strain optimization techniques and associated goal functions are included in this layer. Running the phenotypic simulations needed by the objective functions at each iteration of the optimization algorithms demonstrates the fitness of potential solutions. The objective functions present in this layer. Moreover, this layer allows the use of different optimization functions, such as:

- Product Yield with minimum biomass (YIELD): A minimal amount of biomass (compared to the wild-type) must be generated by computing the yield of the intended product flow (P) in relation to the substrate consumption (S), as $YIELD = \frac{P}{S}$.
- Biomass-Product Coupled Yield (BPCY) (Patil et al., 2005): this objective function multiplies the fluxes of biomass (B) and a chosen product (P) in relation to the substrate (S) consumption: $BCPY = \frac{B \times P}{S}$.
- Target Flux: Maximize or minimize the value of a certain reaction flux.
- Weighed Yield (WYIELD): Although a BPCY score might be high, the flux value of the target reaction may be unstable when maximizing growth. To guide the EA to more robust solutions, that is solutions with non null minimal guaranteed target production, MEWpy includes a weight yield objective. WYIELD leverages the target flux variability by maximizing the trade-off between the minimum and maximum theoretical rates: $WYIELD = \alpha \times FVA_{min} + (1 - \alpha) \times FVA_{max}$.

Also, given their versatility in the formulation of objective functions, MEWpy turns to evolutionary algorithms (EAs) for the CSO algorithms. EAs are stochastic algorithms inspired by nature (Pereira et al., 2021). They operate by developing a group of operators who encode answers to a certain challenge. Each individual is assessed using a fitness function that, based on the effectiveness of the solution, gives

it a numerical value. The fittest are chosen to integrate the following population from the new solution set that is produced by mating and mutation operators after each generation.

MEWpy also has some EAs available for optimization. As it is mostly a multi-objective problem, in theory, the presence of multiple objectives in a problem results in a collection of optimum solutions rather than a single one. Moreover, multi-objective EAs seek to optimize (i) their computational complexity (where is the number of objectives and is the population size); (ii) non-elitism approach; and (iii) the need for specifying a sharing parameter.

One of the available algorithms in MEWpy is the non-dominated sorting genetic algorithm II (NSGA-II) (Deb, Pratap, Agarwal, & Meyarivan, 2002), which makes all three of the aforementioned problems easier. The best K (in terms of fitness and spread) solutions are chosen by mixing the parent and offspring populations in the selection operator, which then generates a mating pool.

Also, MEWpy performs Evolutionary Computation based strain design optimization by knocking out (KO) or over/under expressing (OU) reactions, genes or enzymes. The procedure of these problems is based on the transformation of transcriptional and translational information into restrictions over the fluxes of the reactions by a combinatorial solution that over- or under-expresses a collection of genes. To that end, the Boolean operators [OR/AND] in GPR rules are translated into functional operators (by default, MAX and MIN functions) asserting new reaction flux constraints, using the flux distribution obtained by applying the deletions contained in the genetic modifications as a reference.

The package can evaluate the deletion of a gene, and apply the gene knockout (GKO) problem, which through a *BooleanEvaluator*, evaluates the impact on the reaction after the gene deletion. For Gene over/under (GOU) Problem the procedure is identical, except the boolean operators are substituted with functions. To do this, MEWpy employ a *GeneEvaluator*, which identifies the functions to replace the [AND, OR] operators as well as the gene expression.

Table 3: Types of optimization and respective characteristics.

Optimization	Description	References
FBA	Flux balance analysis (FBA) is an optimization method usually used to predict flux distributions while optimizing a given cellular function or a combination of them. Usually, the FBA results in reasonable solutions for wild-type cells when the given cellular function of maximizing cellular growth.	(Orth et al., 2010)
pFBA	The parsimonious Flux Balance Analysis (pFBA) was developed for minimizing the flux associated with each reaction in a metabolic model while maintaining an optimum flux through the objective function. This approach makes it possible to identify the least biologically "expensive" usage of an organism's metabolism to achieve high growth rates.	(Lewis et al., 2010)
FVA	Flux Variability Analysis (FVA) is an optimization method that determines the maximum and minimum flux values that satisfy the model constraints while maintaining the optimal objective value. This analysis is usually performed after a FBA, in order to, analyze its results. This happens especially when the FBA is performed on a genome-scale metabolic model, since in several constraint-based models it identified unboundedness of the optimal flux space.	(Müller & Bockmayr, 2013)
MOMA	The Minimization Of Metabolic Adjustment (MOMA) is an optimization method that relies on the assumption that optimal growth flux for gene deletions displays a suboptimal flux distribution that is intermediate between the wild-type optimum and the mutant optimum. This projection can be interpreted as the FBA projection onto the feasible space of the related mutant.	(Segre et al., 2002)
ROOM	The Regulatory On/Off Minimization, in similarity to MOMA, aims to predict the steady metabolic state of the organism after gene knockout. However, it uses a different norm than MOMA, minimizing the total number of significant flux changes from the wild-type flux distribution.	(Shlomi et al., 2005)

Table 4: dFBA methods advantages and disadvantages.

Methods	Advantages	Disadvantages	References
SOA	Capable of tracking metabolite dynamics. Don't need non-linear constraints. Scalable to larger systems. Supports the usage of an instantaneous objective function. Easier implementation of an interactive intervention at arbitrary simulation points.	Computationally expensive. It cannot include kinetic or regulatory information.	(Dromms et al., 2020; Gomez et al., 2014; Höffner et al., 2013; Valverde et al., 2019)
DOA	Optimizes during the entire time by solving NLP. Capable of tracking metabolite dynamics.	It is limited to small-scale metabolic models. Possible that non-linear constraints make optimization problems harder to solve. Needs previous knowledge of simulation conditions to transform dFBA into a NLP.	(Dromms et al., 2020; Gomez et al., 2014; Höffner et al., 2013; Valverde et al., 2019)

3 Methods

3.1 Escherichia coli Models

In biology, one of the best characterized metabolic networks belongs to *Escherichia coli*. One of its most modeled metabolisms is the Central Carbon Metabolism (CCM) (Kurata, Maeda, & Matsuoka, 2014) because of (1) the usage of *E. coli* in the production of valuable materials in the industry and as a model microbe; and (2) CCM is nearly completely conserved across organisms and the core upon which almost all catabolic and biosynthetic processes are built (Jahan, Maeda, Matsuoka, Sugimoto, & Kurata, 2016; Noor, Eden, Milo, & Alon, 2010).

The CCM in *E. coli* consists of three main metabolic pathways being Glycolysis, Tricarboxylic Acid Cycle (TCA), and the pentose-phosphate (PP) pathway.

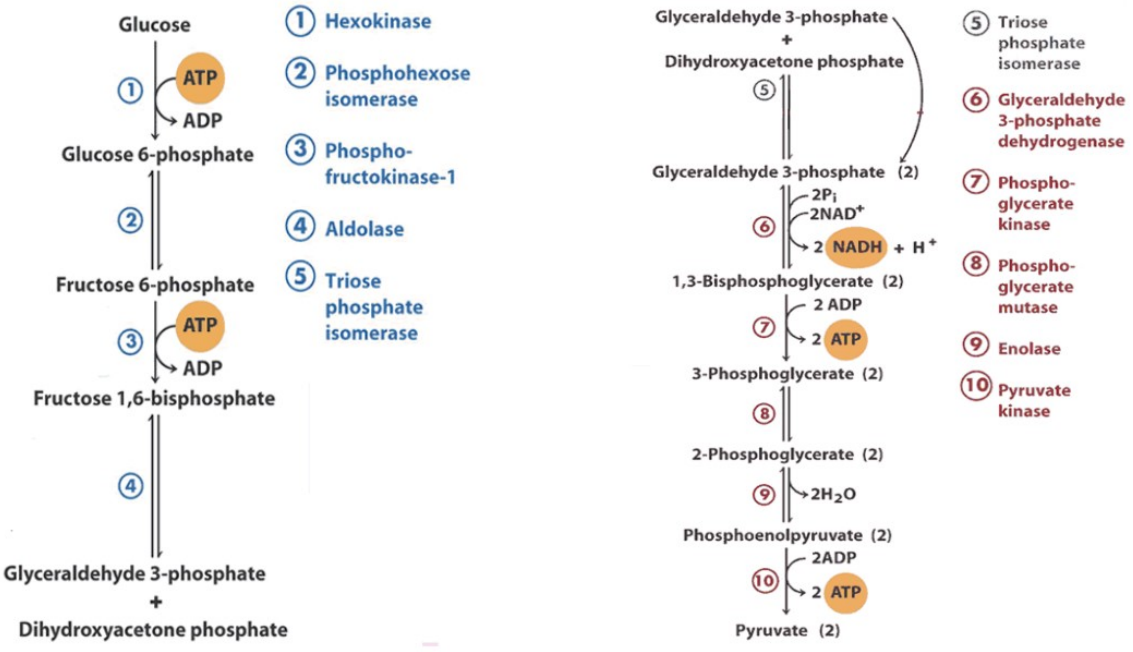
3.1.1 Central Carbon Metabolism Pathways

a) Glycolysis

Glycolysis, also known as the Embden-Meyerhof-Parnas pathway, is a sequence of chemical reactions where glucose turns into pyruvate (a crucial intermediate molecule of all metabolism). This is the main pathway of glucose degradation for energy production in aerobic and anaerobic organisms, the latter also producing metabolites of no interest to the organism but sometimes of industrial interest (Sousa, Faia, Ferreira, & Castro, 2010; Nelson & Cox, 2012).

It is possible to divide glycolysis into two distinct phases, the preparatory phase and the pay-off phase. The preparatory phase (Figure 6a) is where phosphorylations occur (steps 1 and 3 of Figure 6a) where 2 ATP molecules are consumed) to increase the intermediates' free energy until the formation of two molecules of glyceraldehyde-3-phosphate (Nelson & Cox, 2012; Sousa et al., 2010; Røe, 2006).

The pay-off phase consists of the final steps of the pathway shown in red in Figure 6b and ends with the production of two pyruvate molecules. In this step occurs the oxidation of glyceraldehyde-3-phosphate (step 6 where 2 molecules of NADH are generated), two phosphorylations at the substrate level (steps 7 and 10), and the production of one water molecule (step 9). Thus the energy balance of glycolysis is 2 ATP and 2 NADH (Nelson & Cox, 2012; Sousa et al., 2010; Røe, 2006).



(a) Preparatory phase (b) Pay-off phase

Figure 6: Glycolysis phases. The two phases of glycolysis are (a) preparatory phase where glucose is transformed into three-carbon sugar phosphates and 2 ATP molecules are consumed, and (b) pay-off phase where pyruvate, 4 molecules of ATP and 2 molecules of NADH are formed. Image from Røe (2006)

b) Tricarboxylic acid cycle

For most aerobic organisms, glycolysis is only the first step in producing energy by breaking down glucose. The next step is called the TCA cycle or Krebs cycle. This step is responsible for obtaining most of the energy possible for these organisms from simple sugars (Nelson & Cox, 2012; Sousa et al., 2010).

Before the TCA cycle begins, pyruvate from glycolysis and many other metabolic pathways must be converted into acetyl-Coenzyme A (acetyl-CoA). The multienzyme pyruvate dehydrogenase complex catalyzes the pyruvate using thiamine pyrophosphate (TPP), lipoic acid (LA), coenzyme A (CoA), flavin adenine dinucleotide (FAD), NAD^+, and manganese (Mn) as co-factors. This catalytic reaction produces one NADH and one CO_2 molecule (Nelson & Cox, 2012; Sousa et al., 2010).

The Krebs cycle consists of a set of eight reactions, which are represented in Figure 7. First, the oxaloacetate fuses with acetyl-CoA to form citrate and regenerates CoA. Next, citrate is dehydrated into aconitate and then rehydrated into isocitrate. Following this, isocitrate suffers oxidative decarboxylation turning into α-ketoglutarate, where NADH is formed and CO_2 is released. This undergoes oxidative decarboxylation with CoA entry, forming succinyl-CoA and NADH. In the subsequent reaction, CoA is

rereleased due to the scission of succinyl-CoA into succinate. In this reaction, phosphorylation occurs at the substrate level of guanosine diphosphate (GDP) into guanosine triphosphate (GTP, which is used to produce ATP). Afterward, succinate is oxidized to fumarate, forming reduced flavin adenine dinucleotide ($FADH_2$). With the hydration of the fumarate, malate is formed, which is then oxidized to oxaloacetate, thus completing the cycle and forming the last NADH (Nelson & Cox, 2012; Sousa et al., 2010).

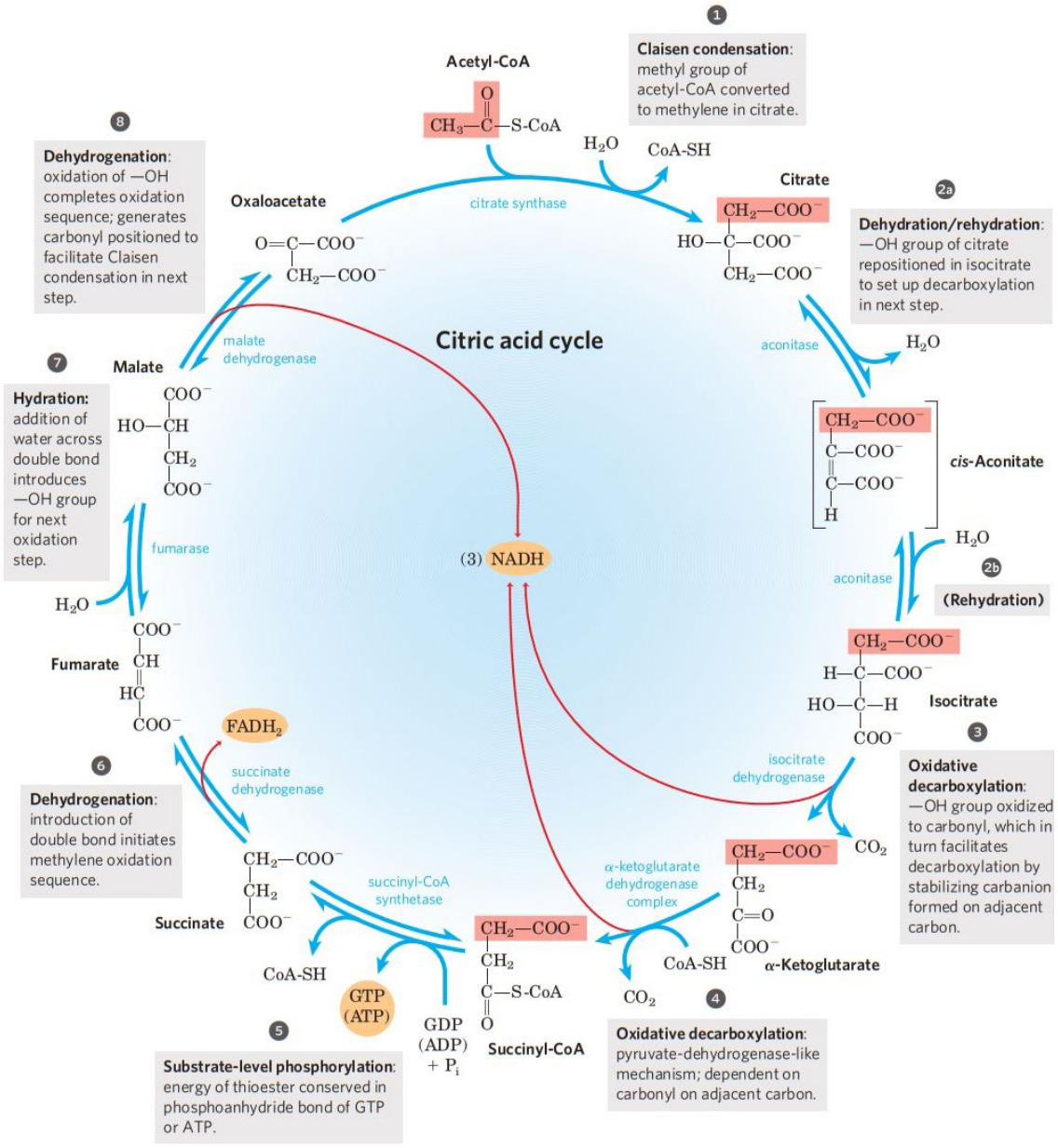


Figure 7: Representation of the citric acid cycle. In this cycle, pyruvate firstly is converted into acetyl-CoA and NADPH. The acetyl-CoA will then enter the cycle where, through a series of reaction, it will produce 3 NADPH, 1 $FADH_2$ and 1 GTP. Shaded in pink are the carbon atoms derived from the acetate of acetyl-CoA. Image from Nelson and Cox (2012).

c) Pentose-phosphate pathway

An alternative pathway for the metabolism of glucose is the PP pathway. This pathway is subdivided into oxidative phase and non-oxidative phase (Figure 8). The oxidative phase is where glucose degradation deviates from the glycolysis pathway and produces ribulose-5-phosphate. In this phase, there is the production of 2 molecules of NADPH and 1 of CO_2 (Figure 8). Ribulose-5-phosphate is an essential intermediate compound in this pathway because (1) this is the compound that initiates the non-oxidative phase, and (2) it is an intermediate in synthesizing nucleotides, coenzymes, DNA, and RNA (Nelson & Cox, 2012; Sousa et al., 2010).

The non-oxidative phase consists of the regeneration of glucose-6-phosphate (Figure 8). In this phase, the reactions convert phosphate pentoses to phosphate hexoses allowing oxidative reactions to continue (Nelson & Cox, 2012).

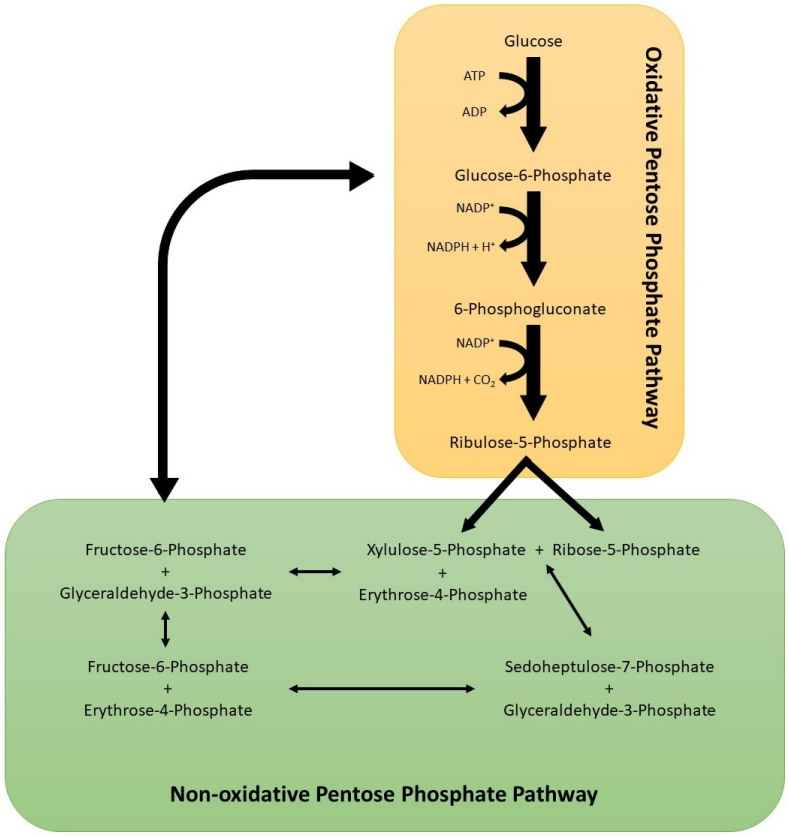


Figure 8: The pentose-phosphate pathway. The pentose-phosphate pathway has an oxidative and a non-oxidative phase. In the oxidative phase, glucose transforms into ribulose-5-phosphate, which produces 2 NADPHs. The non-oxidative phase transforms ribulose-5-phosphate into glucose-6-phosphate, regenerating it. Adapted from Nelson and Cox (2012) and Jiang et al. (2014).

3.1.2 Kinetic model

Over time several dynamic models have been developed. One of the most important is the CCM model of *Escherichia coli* that Chassagnole and his partners developed. This dynamic model was the first to link the sugar transport system (the phosphotransferase system (PTS)) to the CCM of *Escherichia coli* with experimental support. However, this link to the CCM does not comprise one of its most important pathways, the TCA cycle, making it a significant limitation of this model (Chassagnole et al., 2002). In Figure 9 is presented the metabolic of Chassagnole's model.

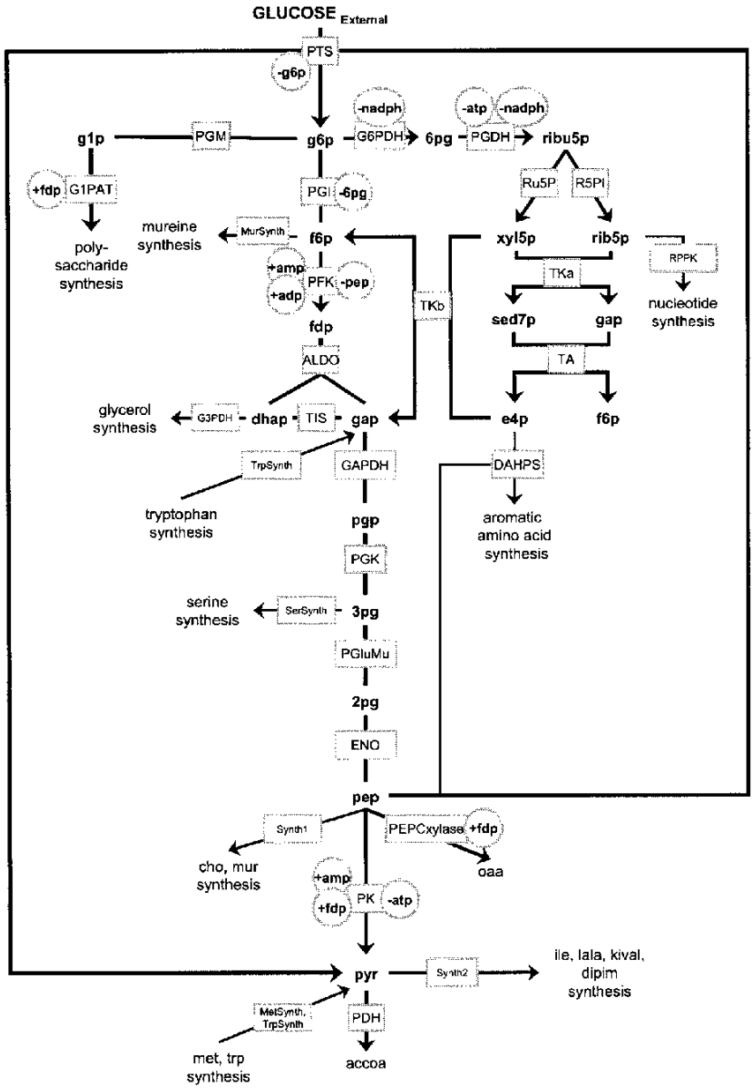


Figure 9: Structure of the Chassagnole model. In the squares are the enzyme activities, and in the circles are the regulatory effects. Image from Chassagnole et al. (2002).

In designing this model, Chassagnole used the intracellular concentrations of metabolites and cometabolites under experimental observed transient conditions to validate this model structure and

estimate kinetic parameters (Chassagnole et al., 2002). This kinetic model will be extensively explored in the result chapter 3.3.3 of this dissertation.

3.1.3 Constraint-based model

With the increased information available in databases and biochemical reviews, there has been the evolution of GSMMs of *Escherichia coli* over the past three decades (Majewski & Domach, 1990; Monk et al., 2017). Two models that represent this evolution are the iJR904 model and the iML1515 model. Both models model the *Escherichia coli* K-12 MG1655 strain, the first strain to be sequenced (Monk et al., 2017; Reed, Vo, Schilling, & Palsson, 2003).

Comparing the two models, we can see in Table 5 that the iML1515 model covers more genes, reactions, metabolites, and open reading frames (ORFs) than the iJR904 model. In addition, all genes in the iML1515 model are connected to a protein product, catalytic domain, and enzyme transformation (Monk et al., 2017; Reed et al., 2003).

Table 5: Comparison of GSMM iJR904 and iML1515 in number of genes, reactions, metabolites, open reading frames (ORFs) and protein domains.

	iJR904	iML1515
Genes	904	1515
Reactions	931	2719
Metabolites	625	1192
ORFs	55	1515
Domains	—	1888
Publication year	2003	2017

3.1.4 Reactions Mapping from Kinetic to Constraint-based Models

In order to realize a hybrid simulation, it is first necessary to have a reaction mapping. This is important because (1) the reaction IDs differ when comparing two different types of models, so it is necessary to identify for each mapped reaction the ID to which it corresponds in each of the models; (2) the directionality of each reaction may differ when compared to the other model. This can happen in reversible reactions where, for example, in the reaction described below, the dynamic model reacts in the k_1 direction, and

the constraint-based model reacts in the k_{-1} direction. In mapping, this translates into the numbers 1 and -1. The number 1 indicates that the reaction is set to the same reaction in both models, and the number -1 indicates that the constraint-based model (where usually the reverse of the reaction direction happens) has to reverse the direction of the reaction to be equal to the dynamic model.

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B$$

Table 6: Reaction mapping from the Chassagnole kinetic model to the iJR904 and iML1515 metabolic models.

Chassagnoles	iJR904		Chassagnoles	iML1515	
Reaction	Reaction	Reaction Direction	Reaction	Reaction	Reaction Direction
vPGI	PGI	1	vPGI	PGI	1
vPFK	PFK	1	vPFK	PFK	1
vALDO	FBA	1	vALDO	FBA	1
vGAPDH	GAPD	1	vGAPDH	GAPD	1
vTIS	TPI	1	vTIS	TPI	1
vPGK	PGK	-1	vPGK	PGK	-1
vrpGluMu	PGM	-1	vrpGluMu	PGM	-1
vENO	ENO	1	vENO	ENO	1
vPK	PYK	1	vPK	PYK	1
vPDH	PDH	1	vPDH	PDH	1
vG6PDH	G6PDH2r	1	vG6PDH	G6PDH2r	1
vPGDH	GND	1	vPGDH	GND	1
vR5PI	RPI	-1	vR5PI	RPI	-1
vRu5P	RPE	1	vRu5P	RPE	1
vTKA	TKT1	1	vTKA	TKT1	1
vTKB	TKT2	1	vTKB	TKT2	1
vTA	TALA	1	vTA	TALA	1
vpepCxylase	PPC	1	vpepCxylase	PPC	1
vPGM	PGMT	-1	vPGM	PGMT	-1
vPTS	GLCpts	1	vPTS	GLCptspp	1

The mapping was done manually using the KEGG database (Kyoto Encyclopedia of Genes and Genomes) since it contains a detailed description of the metabolic reactions (Kanehisa & Goto, 2000) and the BiGG database (Biochemical Genetics and Genomics), which contains the iJR904 and iML1515 models and information on their reactions. The maps originated from the mapping are presented in Table 6.

3.2 Hybrid Modeling Simulation Approach

We devised two approaches for kinetic/constraint-based simulations. The first approach, based on the work described by Machado et al. (2012), reduced the solution space of the constraint-based model by sampling the kinetic space. A second novel approach explores the use of metabolic models with enzymatic constraints to impose restrictions on enzymes usage derived from kinetic simulations.

3.2.1 Sampling approach

The sampling approach consist of varying the maximum velocity or v_{max} parameters of the kinetic model as a proxy for simulating different concentrations of enzymes. Indeed, as the v_{max} is equal to the catalytic rate of the enzyme i for reaction j, k_{cat}^{ij}, times the enzyme concentration $[E_i]$:

$$V_{max}^{ij} = k_{cat}^{ij} \times [E_i], \tag{3.17}$$

by modifying the maximum velocity, and given that the catalytic rates remain constant, we may 'change' the enzyme concentration. This approach allows to obtain a distribution of kinetic fluxes at steady-state to be used to constrain the constraint-based model space.

To define a kinetic solution space for varying enzyme concentrations, we sample the maximum velocities, V_{max}^{ij}, using a log-normal distribution with N(0, σ), and collect the kinetic reaction fluxes rates at steady-state. In the experiments, distinct σ values were considered, notably, $\sigma \in \{1, 2, 4\}$.

The distribution of kinetic reaction rates allows to define a lower (lb) and an upper bound (ub) to be imposed to the respective constraint-based model reactions. Since the sampled reaction rates (values obtained in the sampling) are very susceptible to outliers, we define as lower bound the 10th percentile of the sampled flows, and the 90th percentile as upper bound.

To convert kinetic bounds to constraint-based bounds, we employ a dictionary mapping the reactions and directions. This last is relevant in cases where kinetic and constraint-based reactions have opposite directions. Also, it is necessary to convert the original units to constraint-based units. For example,

the reaction rates in the Chassagnole's model are defined in $mmol/s$ which need to be converted to $mmol/gDW/h$. The formula below demonstrates how the conversion was done:

$$v_j^{cb} = \frac{v_j^k \times direction \times 3600}{gGW} \tag{3.18}$$

where $direction$ is 1 or -1 depending on the value of reaction direction in the map, v_j^k is the rate being converted and gDW is the organism cell volume in $gDWL^-1$. In the Chassagnole's model, this last value is equal to $564.0\ gDWL^-1$.

3.2.2 Enzymatic constraints approach

The GECKO formulation (Sánchez et al., 2017) allows to implement an alternative approach by limiting enzyme usage in constraint-based models. GECKO models take as assumptions that the organism will always try to use enzyme that are more efficient (have a higher catalytic rate) and with a lower cost (have a lower molecular weight). As such, the GECKO formulation adds to reactions stoichiometry, $A \xrightarrow{e_i} B$, the enzyme usage $A + 1/k_{cat} \times e_i \rightarrow B$. The new stoichiometric matrix assumes the form:

$$S' = \left[\begin{array}{ccc:ccc} S_{11} & \dots & S_{1n} & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ S_{m1} & \dots & S_{mn} & 0 & \dots & 0 \\ \hdashline -1/k_{cat}^{11} & \dots & 0 & 1 & \dots & 0 \\ \dots & \ddots & \dots & \dots & \ddots & \vdots \\ 0 & \dots & -1/k_{cat}^{pn} & 0 & \dots & 1 \end{array}\right] \tag{3.19}$$

where new lines are added corresponding to each enzyme e_i. Enzyme usages e_i are limited by their concentration $[E_i]$, modeled by the bottom right sub-matrix of S':

$$0 < e_i < [E_i] \tag{3.20}$$

In the absence of proteomic data, enzyme usages are further limited by a protein pool modeling the fraction of protein mass that corresponds to metabolic enzymes $\sum MW_i.e_i \leq \sigma.f.P_{total}$.

Importantly, the formulation does not alter the steady-state assumption $\sum_j s_{ij}v_i = 0$ (top lines of S'), and adds new enzyme mass balance constants (bottom lines of S'):

$$-\frac{1}{k_{cat}^{ij}}v_j + e_i = 0 \tag{3.21}$$

From equations 3.20 and 3.21 the following constraints may be derived:

$$v_j \leq k_{cat}^{ij}.[E_i] \tag{3.22}$$

reflecting the definition of maximum velocity equation 3.17.

The GECKO formulation allows integrating results from kinetic simulations by modifying enzyme usage bounds. Let us suppose that enzyme i only catalyses reaction j. The constraint on the enzyme usage, $0 < e_i < [E_i]$, may be altered to reflect the kinetic sampling. Indeed, and considering equation 3.17, we may infer an upper bound for the enzyme usage by setting, after units conversion, $[E_i] = V_{max}^{ij}/k_{cat}^{ij}$, where V_{max}^{ij} is sampled from the kinetic space.

Conversely, we may decrease the amplitude of the enzyme usage range by setting a lower bound to the constraint. Considering that $v_j \leq k_{cat}^{ij}.[E_i]$, equation 3.22, and that $v_j = e_i/k_{cat}^{ij}$, equation 3.21, we may impose that $v_j^k/k_{cat}^{ij} \leq e_i \leq [E_i]$, where the v_j^k is the flux rate resulting from the kinetic sampling after units conversion.

Combining both the definition of upper and lower bounds for enzyme usage in GECKO models, and for a kinetic simulation considering maximum reaction rates V_{max}^{ij} and respective fluxes v_j^k, we obtain:

$$v_j^k/k_{cat}^{ij} \leq e_i \leq V_{max}^{ij}/k_{cat}^{ij} \tag{3.23}$$

If the enzyme i is promiscuous, that is, catalyses more than one reaction, the reactions on the GECKO model need to be altered by renaming the proteins in mapped reactions, and by adding the respective pseudo-reactions drawing proteins from the pool. It is important to note that the *arm* reactions, used to preserve the original model reaction bounds in the presence of isozymes, continue to be sufficient. When a promiscuous enzyme i catalyzes more than one mapped reactions j, with j in the set of mapped reactions, equation 3.24 needs to be altered to:

$$min_j(v_j^k/k_{cat}^{ij}) \leq e_i \leq \sum_j (V_{max}^{ij}/k_{cat}^{ij}) \tag{3.24}$$

that is, the enzyme usage is limited by the minimum of the lower bound and by the sum of the upper bound usages from all the mapped reactions j catalyzed by enzyme i.

3.3 Hybrid Optimization Approach

A first approach to a hybrid CSO that constrains the metabolic space is to impose reaction-bound constraints to the metabolic model that results from sampling the kinetic solution space. The distribution

of kinetic reaction flux rates at steady-state that result from varying the v_{maxs} within a log-norm distribution defines reaction constraints on the metabolic model. We will refer to such constraints as *kconstraints*. The followings are extensions of CSO methods already available in MEWpy, and are further described in the tool documentation[1].

3.3.1 Reactions KO/OU on Constrained Solution Space

Using MEWpy, we may run strain design tasks by deleting (KO), "under" or "over-regulating" reactions (OU). The modifications are imposed by altering the metabolic reaction bounds. Integrating the *kconstraints* into a reaction KO (RKO) or OU (ROU) can be achieved by adding the constraints to the model and by defining the respective reactions as non-targets for modification. As such, the optimization procedure will aim to find a set of modifications on reaction bounds (reactions not included in the *kconstraints*) that favor an optimization objective, such as increasing the production of a metabolite of interest.

A second approach consists in running RKO/ROU CSO without imposing prior constraints on the metabolic model and by imposing the *kconstraints* at simulation time. Although the two approaches seem equivalent, they are not. Indeed, in the first approach, the *kconstraints* are imposed when computing reference flux values to realize over/down regulations. In the last, the reference fluxes are computed solely considering the medium definition and the method. Also, the reactions included in the *kconstraints* are considered modification targets. The *kconstraints* are applied at a later time, during the phenotype evaluation of solutions, by intersecting each suggested solution with the *kconstaints*. For example, if constraints are applied in the solution to a reaction R_j, $a < v_j < b$, whose bounds are also modified by the *kconstraints*, $c < v_j < d$. The resulting constraints will be $max(a, c) < v_j < min(c, d)$. If the two ranges are disjoint, the *kconstraint* will prevail; that is, $c < v_j < d$ is applied.

3.3.2 Genes KO/OU on Constrained Solution Space

Integrating the kconstraints in MEWpy gene deletion or gene over/down regulation has an additional complexity due to constraints imposed at different modeling formulation levels (genes vs. reactions). However, distinct approaches may be explored.

[1] https://mewpy.readthedocs.io

The first approach for GKO/GOU problems mirrors the first approach of hybrid RKO/ROU; that is, the kconstraints are imposed beforehand on the model, and the genes associated with the included reactions are not considered modification targets.

A second approach may also follow the second approach used in RKO/ROU, run a MEWpy gene deletion or over/under expression optimization task, and impose the kconstraints when simulating phenotypes. There is however an issue with such an approach: it would impose constraints on reaction bounds that may conflict with modifications on gene expression folds.

3.3.3 Enzymes KO/OU

The logical procedure to conduct computational strain optimization using hybrid GECKO simulations, is to modify the kinetic V_{max} values and run an hybrid simulation. As such, at each iteration of the EA, distinct combinations of V_{max} folds are evaluated defining a candidate solution whose fitness is asserted against the objective based on the resulting phenotypes.

The method can be extended to additionally include modifications on the expression of genes not associated with reactions modeled by the kinetic formulation. The optimization method would allow a mixed os v_{maxs} and proteins solutions as an extension of the GECKO GOU problem already available in MEWpy. For illustrative purposes, a solution might be of the form $\{'rmaxPTS' : 2, 'rmaxG3PDH' : 4, 'P00350' : 0, 'P2925' : 0.5\}$. Running phenotype predictions on the mutant involves:

1. compute the flux distribution of the kinetic model at steady-state by multiplying the model v_{max} values by the respective folds in the solution while considering the remaining default v_{max} values;
2. translate the $vmax$ values and kinetic reaction fluxes into enzymatic usage constraints for the GECKO model;
3. run a pFBA to compute the protein usages in the GECKO model considering, or not (giving rise to a second method), the previous constraints for reference;
4. apply usage constraints on the P00350 and P00393 proteins draw pseudo reactions considering the reference values and respective folds, that is, set the upper (lower) bond to the fold of the reference values for down (up) regulation;
5. run the intended phenotypic simulation (FBA, pFBA, MOMA, IMOMA or ROOM).

The EA is subsequently responsible for computing the fitness of the solution, selecting the best solutions, applying genetic operators, and for driving the optimization.

4 Results

4.1 Validation of MEWpy Kinetic Simulation

In order to be able to implement the hybrid simulation, it was first necessary to validate the kinetic simulation that had already been built. To do this, we used the Chassagnole model and second version of Millard model of central carbon metabolism of *Escherichia coli* (Chassagnole et al., 2002; Millard, Smallbone, & Mendes, 2017) and compared it with the results obtained in the kinetic simulation. To do this, we evaluated the Chassagnole model with two different solvers available in MEWpy: LSODA and Radau; and the Millard model using MEWpy solver LSODA. In Table 7 and Table 8, we can observe that the reaction rates obtained from the kinetic simulation are identical to those presented in the Chassagnole paper and the Millard paper, respectively.

Table 7: Comparison between simulation results using Radau and LSODA solvers and Chassagnole results (Chassagnole et al., 2002).

Metabolite	Chassagnole 2002	LSODA	Radau
CGLCex	0.0556	0.0557029206	0.0557029572
G6P	3.48	3.4818959363	3.4818964616
F6P	0.60	0.6003265013	0.6003265920
FDP	0.272	0.2729634034	0.2729633300
GAP	0.218	0.2184020435	0.2184020359
DHAP	0.167	0.1673057485	0.1673057420
PGP	0.008	0.0080089558	0.0080089593
3PG	2.13	2.1333722481	2.1333731774
2PG	0.399	0.3994430917	0.3994432660
PEP	2.67	2.6729543278	2.6729554980
PYR	2.67	2.6706153349	2.6706153287
6PG	0.808	0.8078619028	0.8078619877
RIBU5P	0.111	0.1110926220	0.1110926321
XYL5p	0.138	0.1381231012	0.1381231135
SED7P	0.276	0.2760076835	0.2760097306
RIB5P	0.398	0.3983460730	0.3983461088
E4P	0.098	0.0981392567	0.0981392603
G1P	0.653	0.6527589402	0.6527590314

Table 8: Comparison between simulation results using LSODA solver and Millard results (Millard et al., 2017).

Metabolite	Millard	Optimization results
2-3PG	1.075	1.0745707555
6PG	0,1316	0.1315997812
G6P	0,861	0.8611291907
PYR	0.237	0.2368914360
PEP	0.997	0.9970378435
FUM	0.213	0.2131148695
MAL	1.032	1.0321525770
F6P	0.262	0.2617663697
ACCOA	0.155	0.1547431071
AMP	0.186	0.1862537413
ADP	0.598	0.5983167527
ATP	2.572	2.5721990425
cAMP	0.923	0.9231306619
NAD	1.412	1.4115435803
NADP	1.168	0.1678371641
NADH	0.158	0.1584564204
NADPH	0.089	0.0891628355

With this, we can consider the MEWpy kinetic simulation tool validated and able to simulate with different types of solvers.

4.2 Impact of Different Sigmas on the Sampling

The sampling approach, described in the previous chapter, derives the kinetic fluxes rate space at steady-state by varying enzymes catalytic maximum velocity, V_{maxs}. This approach, that acts as a proxy for different enzymes concentrations in the model, was conducted on the *E. coli* Chassagnole model by sampling V_{maxs} from log-norm distributions centered on the model optimized parameters and with standard deviations σ in $\{1, 2, 4\}$. As the deviation of V_{maxs} values from the mean increases, the higher the probability of the ODE system to be unsolvable or to take longer to find a solution. Consequently, the number of samples for each σ varies as showcased in Table 9.

Table 9: Number of samples.

σ	N.° of samples
1	9322
2	3188
4	3421

We decided to investigate how the different sigmas affect the sampling results and therefore effects on the steady-state kinetic fluxes. The script used in this investigation can be found in `https://github.com/mpereira19/HybridMEWpy`.

In order to explore the impact of different sigmas on the sampling, we decided to conduct an unsupervised analysis of the sampling, a Principal Component Analysis (PCA) using 2 components, after removing outliers and standardize the samples. The principal components (PC1 and PC2), plotted in Figure 10, can explain approximately 16% and 13% of the data, respectively.

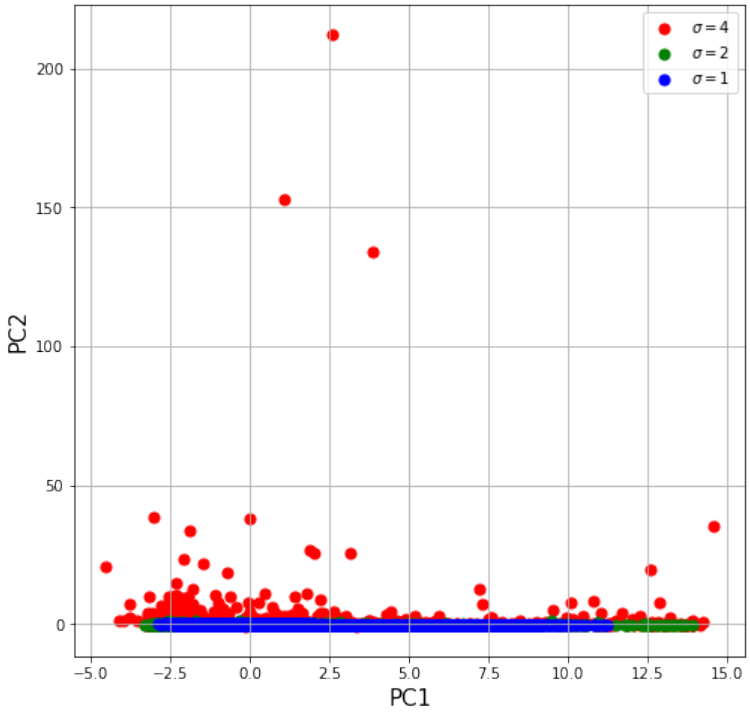


Figure 10: PCA of kinetic flux values using 2 components.

When looking at the PCA plot, we cannot find any well-defined clusters of the data as the second component is highly compacted. This is mainly due to the higher range of flux rate values obtained when increasing the sampling standard deviation. Indeed, by increasing σ as a proxy of enzymes concentrations

the maximum rate of a set of fluxes increases significantly, for example, the maximum rates for Pyruvate kinase are 0.178295 ($\sigma = 1$), 0.257157 ($\sigma = 2$) and 4.930824 ($\sigma = 4$). To mitigate the effects on the PCA of such high values, we may apply a hyperbolic tangent ($tanh$) to all the values, allowing to retain the scale of lower values, closer to 0, while condensing the higher ones. This results in better variance explained values, notably, 42% and 26%, and in a more readable plot (Figure 11).

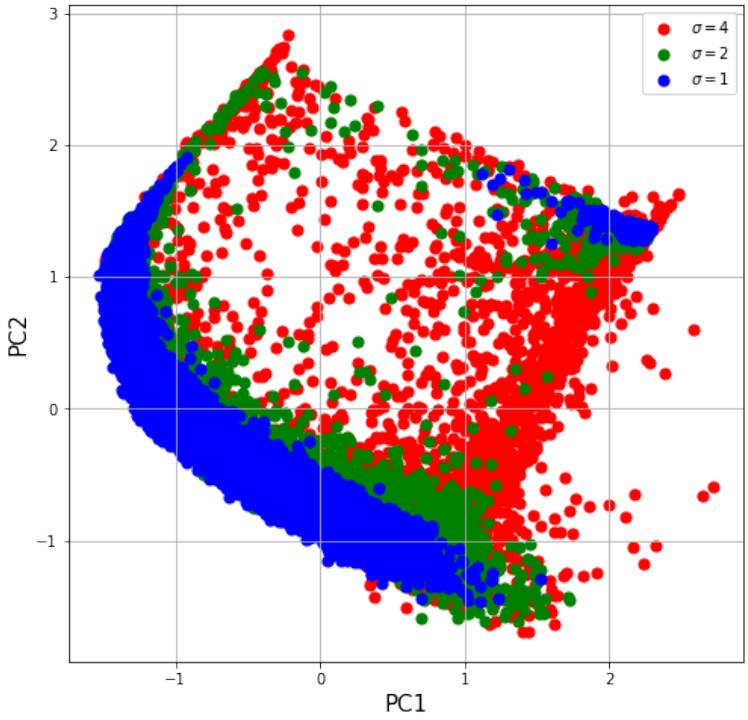


Figure 11: PCA of kinetic flux values using 2 components and tanh.

To interpret each component, we must compute the correlations between the original data and each principal component. This task was accomplished without using the $tanh$, that is, using the components obtained using a standard scaler, which does not affect the correlation analysis. We decided to look at what the correlation was between the original data and each principal component. Table 10 presents the reactions with higher correlation to each of the principal components.

The first principal component (PC1) is strongly correlated with seven original reactions: 6-Phosphogluconate dehydrogenase (vPGDH), Glucose-6-phosphate dehydrogenase (vG6PDH), Ribulose-phosphate epimerase (vRu5P), Transketolase a and b (vTKA and vTKB), Transaldolase (vTA), and Ribose-phosphate isomerase (VR5PI). The second principal component (PC2) is strongly correlated with six original reactions: Glyceraldehyde-3-phosphate dehydrogenase (vGAPDH), Triptophan synthesis (vTRPSYNTH), Phosphoglycerate kinase (vPGK), Phosphoglycerate mutase (vrpGluMu), Enolase (vENO), and Phosphoenolpyruvate carboxylase (vpepCxylase).

Table 10: Reactions with flux rates more correlated with the principal components.

	Reaction	Correlation
PC1	vPGDH	0.997112
	vG6PDH	0.996379
	vRu5P	0.988552
	vTKA	0.985892
	vTA	0.985065
	vTKB	0.965565
	vR5PI	0.920645
PC2	vGAPDH	0.992710
	vTRPSYNTH	0.992374
	vPGK	0.989853
	vrpGluMu	0.987239
	vENO	0.987096
	vpepCxylase	0.981477

By mapping the reactions with higher correlation to the PC1 and PC2 into the Chassagnole model structure, Figure 12 shows the structure of the Chassagnole model, we can better understand why such relations emerged. The reactions more correlated to PC1 are identified in blue and to PC2 in green.

Looking at the first principal component, we can see that the seven reactions that are highly correlated components belong to the pentose-phosphate pathway described in this model. This makes sense because if we increase the flux of the first reaction of this pathway (Glucose-6-phosphate dehydrogenase (vG6PDH)), all the fluxes of the reactions of this pathway will increase accordingly. In a biological perspective, the increase in these reactions fluxes implies that *Escherichia coli* prioritizes glucose degradation by the PP pathway (Figure 8) rather than the glycolysis preparative phase (Figure 6a). This may happen, for example, if it needs to catabolize sugars like D-xylose, D-ribose, or L-arabinose that cannot be catabolized in other pathways. The deficit of reducing energy can also prioritize the PP pathway since NADPH is needed in biosynthesis, in the recruitment of essential metabolites (*e.g.*, nucleic acids, amino acids, and vitamins), and in the production of components of the lipopolysaccharide layer of the cell (Sprenger, 1995; Lin, 1987; Fraenkel, 1987).

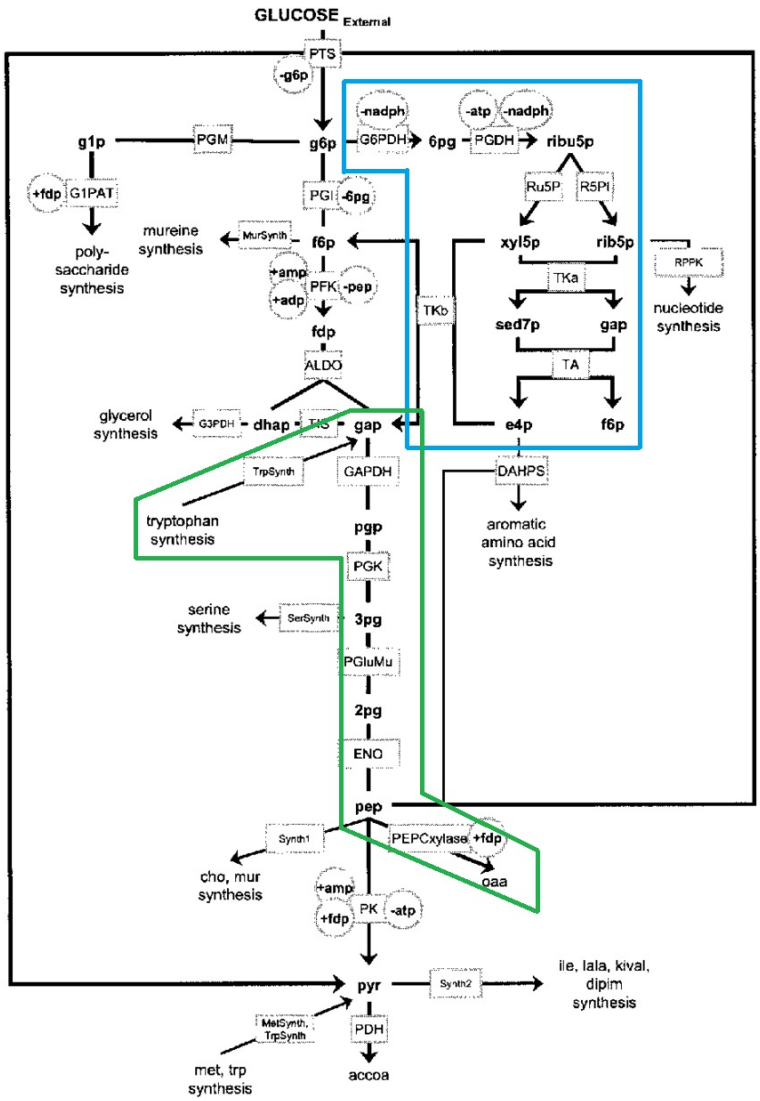


Figure 12: Structure of Chassagnole's kinetic model with reactions identified in the correlation delimited. In blue are the reactions identified in PC1 and in green, the reactions in PC2. Image adapted from Chassagnole et al. (2002).

Similarly, this also happens in PC2, where the reactions that show a high correlation with PC2 correspond to most of the reactions present in the second phase of glycolysis (Figure 6b). The only exceptions to this are tryptophan synthesis (vTRPSYNTH) and phosphoenolpyruvate carboxylase (vpepCxylase). The former produces glyceraldehyde-3-phosphate that enters the beginning of the second phase of glycolysis, while the latter bypasses glycolysis and forms oxaloacetate (oaa). This is important because oxaloacetate is the first compound of the Krebs cycle (Figure 7). Although the Chassagnole model does not have the Krebs cycle in its implementation, the first compound of the cycle appears as a result of phosphoenolpyruvate carboxylase leaving the door open for a future implementation of this model with the Krebs cycle. This highly correlated succession of reactions has biological significance since in most organisms, including *Escherichia coli*, in the energy production process the step following glycolysis is the Krebs cycle. As mentioned in section 3.1, this cycle is responsible for obtaining most of

Table 11: ANOVA F-statistics and p-values for the reactions strongly correlated with the two components.

	Reaction	F score	p-value
PC1	vPGDH	747.769483	4.000488e-311
	vG6PDH	710.675798	2.292880e-296
	vRu5P	706.709168	8.755823e-295
	vTKA	737.870930	3.421699e-307
	vTA	781.754106	0.000000e+00
	vTKB	641.811395	8.456692e-269
	vR5PI	681.396548	1.132912e-284
PC2	vGAPDH	16.706638	5.649355e-08
	vTRPSYNTH	63.184941	4.649559e-28
	vPGK	16.158739	9.760196e-08
	vrpGluMu	12.282010	4.678424e-06
	vENO	12.238612	4.885602e-06
	vpepCxylase	26.358482	3.728712e-12

the energy of these organisms from simple sugars and can produce 3 NADPH, 1 $FADH_2$ and 1 GTP per cycle (Nelson & Cox, 2012; Sousa et al., 2010).

To infer if the individual standard deviation values (σ), used in the sampling procedure, effect the PCA, we conducted an analysis of variance (ANOVA) on the reactions that most contributed to the variability of data, that is, strongly correlated with the two components.

Since all p-values are less than $\alpha = 0.05$, we can reject the null hypothesis. This implies that at least one of the sigma values is different from the other sigmas. To further investigate if all groups are different from one another, we performed a Tukey test for pairwise mean comparisons of all strongly correlated reactions with the two components.

As we can see in Table 12, all highly correlated reactions in PC1 have $\alpha < 0.05$, therefore, we can reject the null hypothesis. This means that for all this seven reactions, there exists a significant difference of fluxes between sigma values. The same does not occur in the PC2 highly correlated reactions, where there is no significant differences between the fluxes of $\sigma = 1$ and $\sigma = 2$. With these results, we have

sufficient proof to say that there exist significant differences in the fluxes due to the sigma values.

Table 12: Tukey test for pairwise mean comparisons p-values of the reactions strongly correlated with the two components.

	Reaction	$\sigma 1$ - $\sigma 2$	$\sigma 1$ - $\sigma 4$	$\sigma 2$ - $\sigma 4$
PC1	vPGDH	0.0	0.0	0.0
	vG6PDH	0.0	0.0	0.0
	vRu5P	0.0	0.0	0.0
	vTKA	0.0	0.0	0.0
	vTA	0.0	0.0	0.0
	vTKB	0.0	0.0	0.0
	vR5PI	0.0	0.0	0.0
PC2	vGAPDH	0.955732	1.172907e-07	7.721179e-06
	vTRPSYNTH	0.944023	0.000000e+00	0.000000e+00
	vPGK	0.955427	1.972911e-07	1.107549e-05
	vrpGluMu	0.938847	8.465882e-06	1.307600e-04
	vENO	0.938845	8.822233e-06	1.346668e-04
	vpepCxylase	0.999046	6.006307e-14	4.490968e-08

4.3 Hybrid Simulation

4.3.1 Sampling Approach

The Chassagnole kinetic model considers a growth of $\mu = 0.278e-4\ s^{-1}$ at steady-state, with a glucose uptake of approximately $0.2\ mmol/s$. Taking these values as reference, we ran a set of simulations using the metabolic model of *E. coli* iML1515. As such, we converted the reference values to the constraint-based model units, respectively, $\mu = 0.1\ h^{-1}$ and $1.28\ mmol/(gDW.h)$. This last considered a cellular volume of $564\ gDWL^{-1}$, the value used to formulate the kinetic model.

To verify the consistency between both models, we conducted an FBA phenotypic simulation on the constraint-based model by imposing a maximum glucose uptake of $1.28\ mmol/(gDW.h)$. The model predicted a growth of $0.089646h^{-1}$, which is approximately 90% of the growth considered by the Chassagnole model. Resorting to a parsimonious FBA, pFBA, we were able to determine that a

maximum glucose uptake of $1.41\ mmol/(gDW.h)$ would allow to reach the intended growth. As such, we considered this last glucose uptake rate as reference.

Following the approach described in Section 3.2, we additionally imposed constraints resulting from sampling the kinetic space using a log-norm distribution with different standard deviations ($\sigma \in (1, 2, 4)$). The predict growths, depicted in Table 13, are very akin and no significant effect on growth can be observed from applying the sampling constraints.

Table 13: Predicted growth on iML1515 with limited glucose uptake. The maximum glucose uptake is set to $1.41\ mmol/(gDW.h)$

Constraints	Growth (h^{-1})
None	0.101441
HYB σ=1	0.100064
HYB σ=2	0.100693
HYB σ=4	0.100810

Given the previous results, we decided to run phenotypic simulations and evaluate cellular growth for increasing rates of the maximum glucose uptake. The results of the experiment, Figure 13, show that, as would be expected, the sampling constraints limit the *E. coli* growth that becomes constant in richer mediums. However, and contrary to intuition, higher σ values translate into lower maximum growth, despite the increasing volume defined by the constraints: (1.87e-06, 0.129 and 0.216, respectively, for $\sigma = 1, 2, 4$).

Interestingly, when identifying the growth limiting constraints for each σ value with a maximum glucose uptake of 10 $mmol/(gDW.h)$, we are able to identify several lower and upper bounds contributing to such a result (see Table 14). This observation is of extreme importance in this work since it demonstrates that the kinetic sampling performed with different σ restricts even more the reaction fluxes of the metabolic model, thus, reducing the range of values at which each reaction can vary and that, in some reactions, the optimal flux is defined by the sampled fluxes. This reduces the solution space of the simulation and therefore delivers a solution with fluxes closer to reality.

Table 14: Constraint bounds and predicted pFBA flux per σ value.

Reactions	WT	$\sigma = 1$			$\sigma = 2$			$\sigma = 4$		
	Fluxes	LB	UB	Fluxes	LB	UB	Fluxes	LB	UB	Fluxes
GLCptspp	10.000000	**1.270098e+00**	1.279681	**1.270098e+00**	0.387029	1.279784	1.139263	1.942640e-02	1.279807	8.680932e-01
PGI	7.644634	-3.446882e-01	1.040456	-3.376827e-01	**-0.135309**	1.214594	**-0.135309**	**-1.036128e-01**	1.248857	**-1.036128e-01**
PGMT	0.000000	-4.184027e-02	-0.000909	-9.086332e-04	-0.089312	**-0.000217**	**-0.000217**	-1.310213e-01	-0.000007	-7.231915e-06
G6PDH2r	2.355366	1.181491e-01	**1.607781**	**1.607781e+00**	0.022477	**1.274572**	**1.274572**	5.595319e-04	**0.971706**	**9.717060e-01**
PFK	6.256868	**4.764070e-01**	1.116531	**4.764070e-01**	**0.001187**	1.189241	**0.001187**	3.129574e-06	1.220201	3.129574e-06
TALA	0.000000	9.763464e-03	0.520805	1.713745e-01	-0.016769	0.414766	0.000000	-8.414611e-03	0.256726	0.000000e+00
TKT1	0.611922	1.140647e-02	0.520883	1.713745e-01	-0.016731	0.414856	0.101099	-8.257408e-03	0.280469	1.078086e-01
TKT2	0.278073	**6.721277e-07**	0.501020	**6.721277e-07**	**-0.062321**	0.384503	**-0.062321**	**-3.173122e-02**	0.197692	**-3.173122e-02**
FBA	6.256868	4.764041e-01	1.116516	4.764070e-01	0.001173	1.189230	0.001187	8.297872e-07	1.220189	3.129574e-06
GAPD	17.105434	1.620500e+00	**2.296688**	**2.296687e+00**	0.415735	**2.418779**	**2.418779**	4.479114e-02	**2.959086**	**2.959084e+00**
TPI	8.328232	**4.670745e-01**	1.104597	**4.670745e-01**	**0.000740**	1.172539	**0.000740**	**-1.469064e-03**	1.183972	**-1.469064e-03**
PGK	-17.105434	**-2.296687e+00**	-1.620499	**-2.296687e+00**	**-2.418779**	-0.415735	**-2.418779**	**-2.959084e+00**	-0.044791	**-2.959084e+00**
PGM	-15.598898	-2.201616e+00	-1.513538	-1.547745e+00	-2.336120	-0.406214	-1.680798	-2.671661e+00	-0.036513	-2.328943e+00
ENO	15.598898	1.513530e+00	2.201597	1.547745e+00	0.406214	2.336045	1.680798	3.651255e-02	2.671624	2.328943e+00
PYK	0.000000	**2.863634e-02**	0.561977	**2.863634e-02**	**0.000162**	0.713785	**0.000162**	**4.995319e-07**	0.654463	**4.995319e-07**
PPC	2.596107	**1.625620e-02**	0.512823	**1.625620e-02**	0.000102	0.713211	0.185516	5.384043e-07	1.083559	1.083559e+00
PDH	9.689687	2.847915e-02	1.570828	1.570828e+00	0.000218	1.816189	1.816189	1.004043e-08	3.917834	3.917834e+00
GND	2.355366	1.181261e-01	1.607527	7.808477e-01	0.022375	1.268623	0.557644	5.204043e-04	0.898687	5.406041e-01
RPI	-1.430463	**-5.915533e-01**	-0.086719	**-5.915533e-01**	**-0.501777**	-0.026539	**-0.501777**	**-4.499361e-01**	-0.001376	**-4.499361e-01**
RPE	0.889996	3.511352e-03	1.022047	1.713752e-01	-0.064383	0.789488	0.038779	-3.617166e-02	0.449654	7.607742e-02

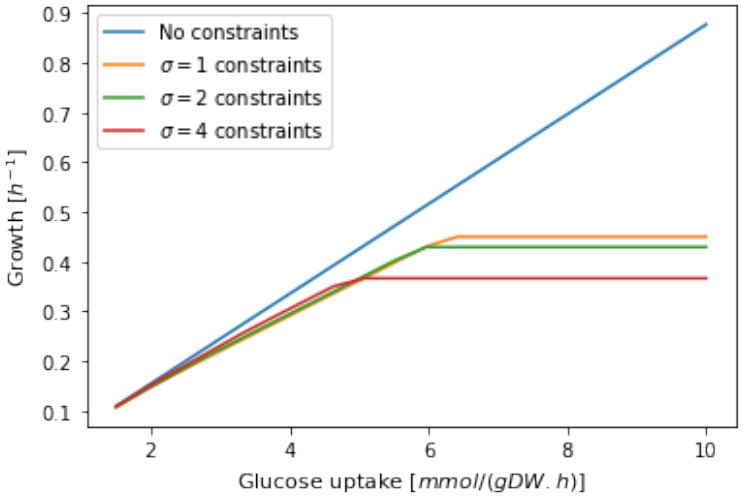


Figure 13: Predicted Growth by glucose uptake.

4.3.2 Hybrid GECKO

Besides integrating kinetic information into constraint-based models by imposing reactions bounds, we also propose a method for hybridizing kinetic and GECKO models. As discussed in Section 3.2 such a task implies modifying the iML1515 GECKO model so that promiscuous enzymes catalyzing reactions constrained and unconstrained by the kinetic model have different draw protein pseudo reactions. This implied renaming proteins falling in this category and add new species as well as new pseudo draw reactions.

The mapping between the kinetic and GECKO model requires additional information, notably the V_{max} parameter identifier, the constraints-based model reaction identifier, the set of proteins and respective K_{cats} for each direction, and a flag identifying if the reaction direction is the same in both models. This information, for ease of usage, is stored in a JSON file, as next illustrated:

```
{
...
    "vALDO":
        ["rmaxALDO",
            [   # forward reaction
                ["P0AB71A", 51119.51743175544],
                ["P0A991A", 51119.51743175544]
            ],
            [   # backward reaction
                ["P0AB71A", 7.199942400460795],
                ["P0A991A", 7.199942400460795]
            ],
```

```
            1
        ],
...
}
```

Similarly to the previous approach, we ran phenotypic simulations considering a glucose uptake of 1.41 $mmol/(gDW.h)$ on the GECKO model without and with constraints from the kinetic simulation (hybrid GECKO). This last considered the kinetic model default parameters. The biomass reaction rates were in both cases equal to 0.101 h^{-1} which is consistent with the *in vivo* observations used to optimize the kinetic model parameters. Additionally, Figure 14 presents the predict growth using the GECKO model and the hybrid GECKO approach for increasing glucose uptake rates. This last is simulated applying only the upper bounds derived from the V_{maxs} (Hybrid GECKO UB), and additionally applying lower bounds to enzyme usage derived from the kinetic flux rates (Hybrid GECKO LB UB). The imposition of enzyme usages lower bounds renders the model infeasible for glucose maximum uptakes under 3.8 $mmol/(gDW.h)$.

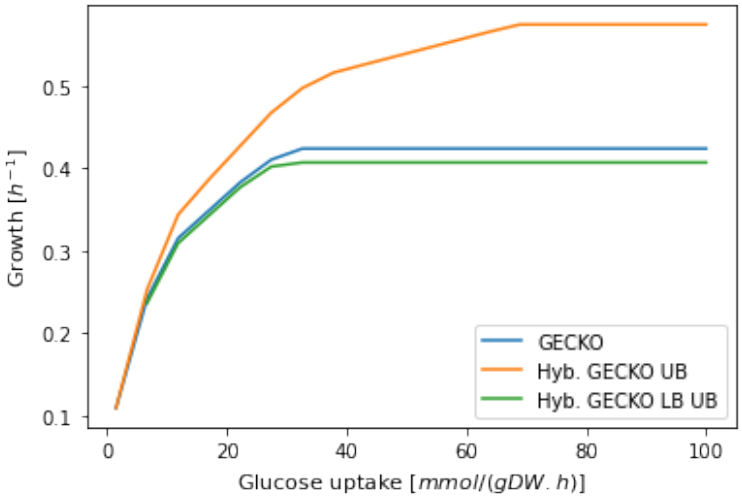


Figure 14: Predicted Growth by glucose uptake using GECKO.

In the GECKO formulation, growth is limited not only by the availability of metabolites in the medium but also by the enzymes concentrations. Consequently, growth on GECKO models has a maximum rate which, in this case, is 0.574 h^{-1}. The hybrid GECKO approaches exhibited lower maximum growths of 0.424 h^{-1} and 0.407 h^{-1} by only restraining the upper bound or both bounds of enzyme usage pseudo-reactions respectively.

4.3.3 Comparison of the Hybrid Approaches

A major difference emerges from restricting the solution space of the constraint-based model using sampling, and the hybrid GECKO approach. The second allows simulating different phenotypes resulting from single sets of modifications of kinetic parameters, thus achieving, theoretically, better predictions. The sampling approach could, however, be modified such that for each combination of V_{max} values, the resulting kinetic flux rates are imposed as constraints in the constraint-based model. This, however, would lead in (almost) all cases to infeasible linear programming problems. To overcome such a problem, the reactions upper and lower bounds can be set to a defined a range of allowable values, that is add a tolerance $(v - A/2, v + A/2)$ to flux rate values where A and v are respectively an amplitude and the flux value.

Alternatively, one may choose to use a partial MOMA or IMOMA to find the closest feasible solution in the constraint-based model space. The partial IMOMA only considers the kinetic/constraint-based mapped reactions in the objective and the biomass taking as reference, repectively, the flux rates derived from the kinetic formulation and the wild type growth. Thus, the IMOMA objective is to minimize $\sum |v_{kin} - v_{cb}| + |biomass_{wt} - biomass_{cb}|$. Both approaches are available in the hybrid implementation.

As illustrative examples, we considered changing all V_{max} values by a set of factors $f \in (0.8, 0.9, 1, 1.5, 2, 2.5)$. The amplitude tolerance was set to 0.1 and the maximum glucose uptake to 1.41 and 10 $mmol/(gDW.h)$. The results of such an experiment are presented in Table 15.

Table 15: Comparison of predicted growth by hybridization approach

Factor	GECKO		Amplitude		IMOMA	
0.8	0.101443	0.399110	0.098279	0.145845	0.000581	0.058393
0.9	0.101443	0.403039	0.098324	0.146887	0.000543	0.064866
1.0	0.101443	0.407070	0.098483	0.145005	0.000492	0.070649
1.5	0.101443	0.427869	0.098482	0.146477	0.000298	0.081164
2.0	0.101443	0.448821	0.097808	0.156880	0.000229	0.095574
2.5	0.101443	0.469858	0.097000	0.170730	0.000196	0.112384
Max. glucose uptake $mmol/(gDW.h)$	1.41	10	1.41	10	1.41	10

At a first glance, the growth predictions are very different between the approaches, where the GECKO approach has consistently higher growth prediction. Looking at the phenotypic results that consider a maximum glucose uptake rate of 1.41 $mmol/(gDW.h)$, in the case of the hybrid GECKO and amplitude approaches with a unit factor, these are more or less consistent with the constraint-based and kinetic models without hybridization. The IMOMA approach, on the other hand, falls very short and might not be a good hybridization method. Also, the hybrid GECKO simulations predict a same growth regardless of enzymes concentrations. This results from the phenotypes being solely limited by the concentration of metabolites in the medium. A similar conclusion may be inferred for the amplitude approach, although some variations on growth can be observed.

Concerning the richer glucose medium, the amplitude approach fails to predict growth values consistent with the constraint-based modelings. This is mainly due to the defined allowed amplitude range of 0.1. Indeed, such a value narrows the solution space preventing higher growths. Alternative amplitude values might be considered, however the choice of such a value is challenging. In the extreme case, it would be preferable to limit fluxes to ranges derived from the sampling approach.

The hybrid GECKO approach seems to be the preferred choice. Besides delivering consistent results, it allows the integration of additional information: the enzymes k_{cat} and molecular weight, which may lead to better predictions given that an organism, in non stress conditions, will aim to use more efficient enzymes with a lower synthesis cost.

4.4 Hybrid Optimization of Succinate Production

In order to exploit the capabilities of our tool in strain design optimization, we decided to perform a hybrid optimization targeted at increasing succinate production. Succinate is a compound used widely in the pharmaceutical and food industries. This compound is often added to foods as it can be used as a food flavoring agent since it is a slightly acidic and astringent component to the umami taste (*e.g.* in shellfish). In addition to being used in shellfish, succinate is also widely available salt (mono or disodium) and is also added to flavor meats, soups, and many other foods (Tsukatani & Matsumoto, 2000). Already in the pharmaceutical industry, succinate is used as a counter ion in many pharmaceutical formulations.

To achieve a strain capable of increasing the production of succinate, we first needed to obtain a number of possible strain designs that may be capable of maximizing succinate production. We achieved this with the help of the integrated EAs and GOU problem function by constraining the model with the flux bounds

described in Table 14 for $\sigma = 4$ where we obtained several possible gene modifications (deletion or over/down gene regulation) capable of achieving this goal.

The optimizations were ran defining as objectives the maximization of the BCPY, with flux rates predictions obtained using IMOMA, and WYIELD ($\alpha = 0.3$) for 10 iterations of 500 generations using the algorithm GOU in a glucose medium. The NSGAII algorithm was used with populations of 100 individuals. The genes targeted for modifications were all metabolic genes, with the exception of genes associated to transport reactions and those associated with kinetic mapped reactions. The allowed folds for genes expression were 2^k for $k \in \{-5, -4, -3, -2, -1, 1, 2, 3, 4, 5\}$ and 0 (gene knockout). Folds greater than 1 correspond to up regulations, while folds lower than 1 are deletions (down regulations).

The solutions from all runs underwent a filtering process, removing all solutions with predict growth below $0.1h^{-1}$ and null minimum FVA. After analyzing the end results, the strain with better succinate production had the gene modifications presented in Table 16.

Table 16: Gene modifications capable of creating a new strain that maximizes succinate production.

Gene code	Gene	Enzyme/Polypeptide	BioCyc ID	Modification
b4388	serB	phosphoserine phosphatase	EG10945	0
b2508	guaB	inosine 5'-monophosphate dehydrogenase	EG10421	32
b3640	dut	dUTP diphosphatase	EG10251	32
b0722	sdhD	succinate:quinone oxidoreductase, membrane protein SdhD	EG10934	0
b1131	purB	adenylosuccinate lyase	EG11314	8
b1849	purT	phosphoribosylglycinamide formyltransferase 2	EG11809	16
b3941	metF	5,10-methylenetetrahydrofolate reductase	EG10585	2
b1380	ldhA	D-lactate dehydrogenase	G592	8

As we can see from the modification values in Table 16, two genes were deleted, and the remaining six were over expressed. The two deleted genes, b4388 and b0722, are connected to the L-serine biosynthesis pathway (Figure 15 and succinate dehydrogenase in the Krebs cycle. Biologically the deletion of these two genes will indirectly and directly influence succinate production since (1) the lack of the serB enzyme will directly prevent the production of L-serine, which is a direct precursor in the biosynthesis of L-cysteine, L-methionine, L-tryptophan, and glycine and which inhibits the initiation of this pathway (Sultan et al., 2021; Samsonov et al., 2022); this way the flux reaction pathway can relocate

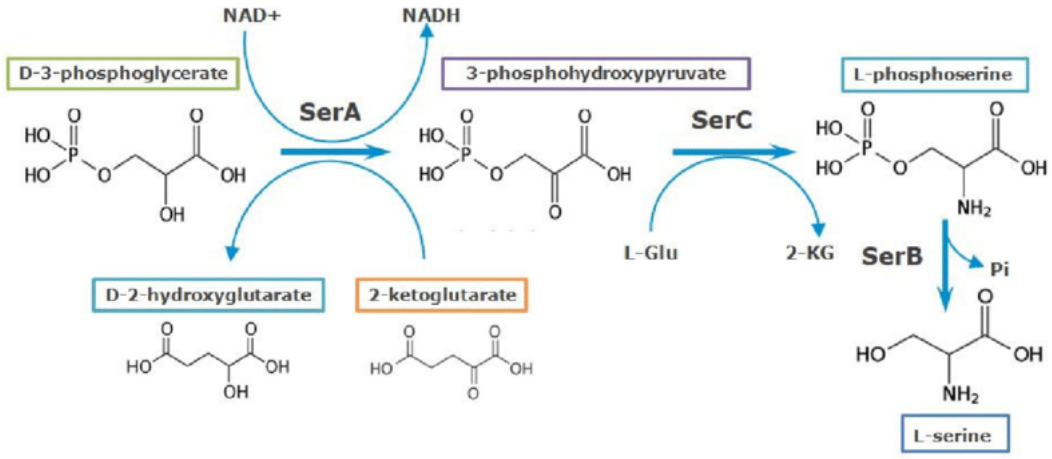


Figure 15: L-serine biosynthesis pathway. Image from (Samsonov et al., 2022)

energy into other reactions that promote the production of succinate and cell survival; (2) the lack of the polypeptide sdhD during the Krebs cycle, prevents the dehydrogenation of succinate into fumarate and by adding the fumarate reductase usage, which is used in anaerobic growth, the succinate production increases (Figure 7).

Of the remaining six genes, purB, purT, and guaB are involved in the histidine, purine, and pyrimidine biosynthesis superpathway, dut is involved in the pyrimidine deoxyribonucleotide dephosphorylation pathway, metF in the folate transformations III pathway, and ldhA in the lactic fermentation pathway (all pathways and genes can be seen in BioCyc website[2]). Looking at the literature, we could not find a single strain that, by over expressing one or more of these six genes, could maximize succinate production. This may imply that all of these genes are not indirectly involved in processes that maximize succinate production, but rather in processes relevant to cell stability and survival. Regarding the 2 deleted genes, b4388 and b0722, few strains were found to delete one of these genes but never together or in pair with the other six genes. These strains were found in SDDB database[3] and their comparison with our strain is shown in Table 17. Here we can see a great disparity between the succinate production values obtained in other strains and ours. According to our simulation results, our strain produces almost three times more succinate than the best succinate producing strain found. Despite this promising fact, our strain as one of the worst growth rate values among the strains described in Table 17, being only surpassed by the first strain where serB was deleted.

[2]https://biocyc.org/

[3]https://sddb.bio.di.uminho.pt/

Table 17: Strain values regarding glucose uptake, growth and succinate production and variation of our strain in comparison to four other mutant strains (two where serB was deleted and two where sdhD was deleted). Strains obtained in SDDB database.

	Glucose uptake ($mmol/(gDW.h)$)	Growth (h^{-1})	Succinate production ($mmol/(gDW.h)$)	Succinate variation ($mmol/(gDW.h)$)
Our simulation	10	0.109225	9.236243	[7.916763 - 9.437548]
serB deletion 1	10	0.08912	3.19504	[3.19504 - 14.29914]
serB deletion 2	10	0.4774	0.57423	[0.12037 - 1.20038]
sdhD deletion 1	10	0.70988	0.22732	[0.22505 - 0.65931]
sdhD deletion 2	10	0.70988	0.22732	[0.22505 - 0.7606]

The results obtained in our simulation are very promising since our tool was able to select a set of genes that boosted succinate production compared to other strains found. Despite this there are still many question that need to be answered, for example, knowing that the secretion of succinate changes the pH of the medium, will our strain be able to survive in this medium? As for its slow growth, is this a problem for its production in bioreactors? As future work it would be interesting to try to answer these questions with *in vivo* experiments.

5 Conclusion

In conclusion, with this work, we were able to create a tool for Mewpy that can perform kinetic/metabolic hybridization. In addition to this, it can use Gecko models in hybridizations. To perform this work, we chose the Chassagnole kinetic model of the central carbon metabolism, the iML1515 metabolic model, and the ec_iML1515_batch Gecko model since they pertain to one of the most studied organisms, the *Escherichia coli*.

In this work, we validated the previously built kinetic simulation by comparing the reaction rates obtained with the respective fluxes found in the literature of each model. We were also able to statistically demonstrate that different sigmas in a log-normal distribution impact sampling, even though this may not occur in all reactions. In the Chassagnole model sampling, we also detected that all seven reactions that constitute the pentose-phosphate pathway (Figure 12 blue) are highly correlated. A sequence of highly correlated reactions was also detected that leads the second phase of glycolysis to the Krebs cycle (Figure 12 green), even though this cycle is not present in Chassagnole model.

Furthermore, we evaluated our sampling approach by comparing the growth value obtained from constraining the metabolic model with the three samplings and without such restrictions. Here we verified no significant differences in growth values, implying that these constraints do not significantly affect growth. However, and contrary to intuition, higher σ values translate into lower maximum growth, despite the increasing volume defined by the constraints. Upon further investigation as to why this happened, we identified several lower and upper bounds of the sampling that contributed to such results. This permitted us to observe the impact of the sampling in the metabolic model and what the response of the microorganism might be in such conditions.

We also hybridized a GECKO model. Here, we verified that growth is not only limited by the availability of metabolites in the medium but also by the enzyme concentrations. We also observed that the hybrid GECKO approach exhibited lower maximum growths by only restraining the upper bound or both bounds of enzyme usage pseudoreactions. To see which one of the two developed approaches was preferred in hybridization, we compared the predicted growth of the GECKO and IMOMA approaches. Here we were able to identify that GECKO hybrid models are a better hybridization option than the IMOMA method since IMOMA has less consistent results when compared to GECKO and GECKO allows the integration of additional information, such as the enzymes k_{cat} and molecular weigh.

Finally, we performed a hybrid optimization of succinate production using this tool's Gene over/under (GOU) problem functions. Here we obtained a very promising strain where we observed a significant increase in succinate production by modifying a set of eight genes. When we compared our strain with others, we observed that our strain produced almost three times more succinate than the second-best strain. However, several questions regarding its survival in a different pH medium and its slow growth effect on bioreactor production remain.

Future work

In the future, some optimization strategies that were not performed due to lack of time will be done. Optimizations using Reaction over/under (ROU) problem functions, optimizations of hybrid kinetic/Gecko models with both ROU and GOU problems will be done. Also, further analysis of the simulation methods will need to be performed, in particular comparing phenotypes with experimental data in the laboratory.

Bibliografia

Anesiadis, N., Cluett, W. R., & Mahadevan, R. (2008). Dynamic metabolic engineering for increasing bioprocess productivity. *Metabolic engineering*, *10*(5), 255–266.

Antoniewicz, M. R. (2015). Methods and advances in metabolic flux analysis: a mini-review. *Journal of Industrial Microbiology and Biotechnology*, *42*(3), 317–325.

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., & Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of escherichia coli. *Biotechnology and bioengineering*, *79*(1), 53–73.

Chou, T.-C., & Talaly, P. (1977). A simple generalized equation for the analysis of multiple inhibitions of michaelis-menten kinetic systems. *Journal of Biological Chemistry*, *252*(18), 6438–6442.

Cornish-Bowden, A. (1974). A simple graphical method for determining the inhibition constants of mixed, uncompetitive and non-competitive inhibitors. *Biochemical Journal*, *137*(1), 143.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, *6*(2), 182–197.

Dromms, R. A., Lee, J. Y., & Styczynski, M. P. (2020). Lk-dfba: a linear programming-based modeling strategy for capturing dynamics and metabolite-dependent regulation in metabolism. *BMC bioinformatics*, *21*(1), 1–14.

Edwards, J. S., & Palsson, B. O. (2000). Metabolic flux balance analysis and the in silico analysis of escherichia coli k-12 gene deletions. *BMC bioinformatics*, *1*(1), 1–10.

Feng, J., Fu, W., & Sun, F. (2010). *Frontiers in computational and systems biology* (Vol. 15). Springer Science & Business Media.

Fraenkel, D. (1987). Glycolysis, pentose phosphate pathway, and entner-doudoroff pathway. *Escherichia coli and Salmonella typhimurium: cellular and molecular biology. American Society for Microbiology, Washington, DC*, 142–150.

Gomez, J. A., Höffner, K., & Barton, P. I. (2014). Dfbalab: a fast and reliable matlab code for dynamic flux balance analysis. *BMC bioinformatics*, *15*(1), 1–10.

Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC bioinformatics*, *11*(1), 1–3.

Heinrich, R., & Schuster, S. (1998). The modelling of metabolic systems. structure, control and optimality. *Biosystems*, *47*(1-2), 61–77.

Höffner, K., Harwood, S. M., & Barton, P. I. (2013). A reliable simulator for dynamic flux balance analysis. *Biotechnology and bioengineering*, *110*(3), 792–802.

Ingalls, B. P. (2013). *Mathematical modeling in systems biology: an introduction*. MIT press.

Jahan, N., Maeda, K., Matsuoka, Y., Sugimoto, Y., & Kurata, H. (2016). Development of an accurate kinetic model for the central carbon metabolism of escherichia coli. *Microbial cell factories*, *15*(1), 1–19.

Jamshidi, N., & Palsson, B. Ø. (2010). Mass action stoichiometric simulation models: incorporating kinetics and

regulation into stoichiometric models. *Biophysical journal*, *98*(2), 175–185.

Jiang, P., Du, W., & Wu, M. (2014). Regulation of the pentose phosphate pathway in cancer. *Protein & cell*, *5*(8), 592–602.

Johnson, K. A., & Goody, R. S. (2011). The original michaelis constant: translation of the 1913 michaelis–menten paper. *Biochemistry*, *50*(39), 8264–8269.

Kanehisa, M., & Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27–30.

Keleti, T. (1986). Two rules of enzyme kinetics for reversible michaelis-menten mechanisms. *FEBS letters*, *208*(1), 109–112.

Kim, O. D., Rocha, M., & Maia, P. (2018). A review of dynamic modeling approaches and their application in computational strain optimization for metabolic engineering. *Frontiers in microbiology*, *9*, 1690.

Klipp, E., Herwig, R., Kowald, A., Wierling, C., & Lehrach, H. (2005). *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons.

Kurata, H., Maeda, K., & Matsuoka, Y. (2014). Dynamic modeling of metabolic and gene regulatory systems toward developing virtual microbes. *Journal of Chemical Engineering of Japan*, *47*(1), 1–9.

Lee, J. M., Gianchandani, E. P., Eddy, J. A., & Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol*, *4*(5), e1000086.

Leighty, R. W., & Antoniewicz, M. R. (2011). Dynamic metabolic flux analysis (dmfa): a framework for determining fluxes at metabolic non-steady state. *Metabolic engineering*, *13*(6), 745–755.

Lequeux, G., Beauprez, J., Maertens, J., Van Horen, E., Soetaert, W., Vandamme, E., & Vanrolleghem, P. A. (2010). Dynamic metabolic flux analysis demonstrated on cultures where the limiting substrate is changed from carbon to nitrogen and vice versa. *Journal of Biomedicine and Biotechnology*, *2010*.

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., ... others (2010). Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, *6*(1), 390.

Lin, E. (1987). Dissimilatory pathways for sugars, polyols, and carboxylates. *Escherichia coli and Salmonella typhimurium: cellular and molecular biology*, *1*, 244–284.

Llaneras, F., & Picó, J. (2008). Stoichiometric modelling of cell metabolism. *Journal of bioscience and bioengineering*, *105*(1), 1–11.

Luo, R., Wei, H., Ye, L., Wang, K., Chen, F., Luo, L., ... others (2009). Photosynthetic metabolism of c3 plants shows highly cooperative regulation under changing environments: a systems biological analysis. *Proceedings of the National Academy of Sciences*, *106*(3), 847–852.

Luo, R.-Y., Liao, S., Tao, G.-Y., Li, Y.-Y., Zeng, S., Li, Y.-X., & Luo, Q. (2006). Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions. *Molecular systems biology*, *2*(1),

2006–0031.

Machado, D., Costa, R. S., Ferreira, E. C., Rocha, I., & Tidor, B. (2012). Exploring the gap between dynamic and constraint-based models of metabolism. *Metabolic engineering*, *14*(2), 112–119.

Mahadevan, R., Edwards, J. S., & Doyle III, F. J. (2002). Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, *83*(3), 1331–1340.

Maia, P., Rocha, I., & Rocha, M. (2013). An integrated framework for strain optimization. In *2013 ieee congress on evolutionary computation* (pp. 198–205).

Maia, P., Rocha, M., & Rocha, I. (2016). In silico constraint-based strain optimization methods: the quest for optimal cell factories. *Microbiology and Molecular Biology Reviews*, *80*(1), 45–67.

Majewski, R., & Domach, M. (1990). Simple constrained-optimization view of acetate overflow in e. coli. *Biotechnology and bioengineering*, *35*(7), 732–738.

McCloskey, D., Palsson, B. Ø., & Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli. *Molecular systems biology*, *9*(1), 661.

Meadows, A. L., Karnik, R., Lam, H., Forestell, S., & Snedecor, B. (2010). Application of dynamic flux balance analysis to an industrial escherichia coli fermentation. *Metabolic engineering*, *12*(2), 150–160.

Mendoza, S. N., Olivier, B. G., Molenaar, D., & Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome biology*, *20*(1), 1–20.

Michaelis, L., & Menten, M. L. (1913). The kinetics of invertin action. *FEBS Letters*, *587*(17), 2712-2720.

Millard, P., Smallbone, K., & Mendes, P. (2017). Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in escherichia coli. *PLoS computational biology*, *13*(2), e1005396.

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., ... others (2017). iml1515, a knowledgebase that computes escherichia coli traits. *Nature biotechnology*, *35*(10), 904–908.

Moulin, C., Tournier, L., & Peres, S. (2021). Combining kinetic and constraint-based modelling to better understand metabolism dynamics. *Processes*, *9*(10), 1701.

Müller, A. C., & Bockmayr, A. (2013). Fast thermodynamically constrained flux variability analysis. *Bioinformatics*, *29*(7), 903–909.

Nelson, D. L., & Cox, M. M. (2012). Lehninger principles of biochemistry 6th edition. In (chap. 14–16). WH Freeman and Company.

Noor, E., Eden, E., Milo, R., & Alon, U. (2010). Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Molecular cell*, *39*(5), 809–820.

Oberhardt, M. A., Puchałka, J., Martins dos Santos, V. A., & Papin, J. A. (2011). Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS computational biology*, *7*(3), e1001116.

Oddone, G. M., Mills, D. A., & Block, D. E. (2009). A dynamic, genome-scale flux model of lactococcus lactis to

increase specific recombinant protein expression. *Metabolic engineering*, *11*(6), 367–381.

Ohno, S., Shimizu, H., & Furusawa, C. (2014). Fastpros: screening of reaction knockout strategies for metabolic engineering. *Bioinformatics*, *30*(7), 981–987.

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, *28*(3), 245–248.

Patil, K. R., Rocha, I., Förster, J., & Nielsen, J. (2005). Evolutionary programming as a platform for in silico metabolic engineering. *BMC bioinformatics*, *6*, 308.

Pereira, V., Cruz, F., & Rocha, M. (2021). Mewpy: a computational strain optimization workbench in python. *Bioinformatics*, *37*(16), 2494–2496.

Pettigrew, M. F., Resat, H., & Petzold, L. (2009). Kinetic modeling of biological systems. *Methods in molecular biology (Clifton, N.J.)*, *541*, 337–354.

Phair, R. D., & Misteli, T. (2001). Kinetic modelling approaches to in vivo imaging. *Nature Reviews Molecular Cell Biology*, *2*(12), 898–907.

Pizarro, F., Varela, C., Martabit, C., Bruno, C., Pérez-Correa, J. R., & Agosin, E. (2007). Coupling kinetic expressions and metabolic networks for predicting wine fermentations. *Biotechnology and bioengineering*, *98*(5), 986–998.

Reed, J. L., Vo, T. D., Schilling, C. H., & Palsson, B. O. (2003). An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome biology*, *4*(9), 1–12.

Rizzi, M., Baltes, M., Theobald, U., & Reuss, M. (1997). In vivo analysis of metabolic dynamics in saccharomyces cerevisiae: Ii. mathematical model. *Biotechnology and bioengineering*, *55*(4), 592–608.

Røe, K. (2006). *In vivo magnetic resonance spectroscopy and diffusion weighted magnetic resonance imaging for non-invasive monitoring of treatment response of subcutaneous ht29 xenografts in mice* (Unpublished master's thesis). Institutt for elektronikk og telekommunikasjon.

Sainz, J., Pizarro, F., Pérez-Correa, J. R., & Agosin, E. (2003). Modeling of yeast metabolism and process dynamics in batch fermentation. *Biotechnology and Bioengineering*, *81*(7), 818–828.

Samsonov, V. V., Kuznetsova, A. A., Rostova, J. G., Samsonova, S. A., Ziyatdinov, M. K., & Kiriukhin, M. Y. (2022). Revealing a new family of d-2-hydroxyglutarate dehydrogenases in escherichia coli and pantoea ananatis encoded by ydij. *Microorganisms*, *10*(9), 1766.

Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E. J., & Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology*, *13*(8), 935.

Schuster, S., & Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, *02*(02), 165-182. doi: 10.1142/S0218339094000131

Segre, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks.

Proceedings of the National Academy of Sciences, *99*(23), 15112–15117.

Shlomi, T., Berkman, O., & Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the national academy of sciences*, *102*(21), 7695–7700.

Smallbone, K., Messiha, H. L., Carroll, K. M., Winder, C. L., Malys, N., Dunn, W. B., ... others (2013). A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. *FEBS letters*, *587*(17), 2832–2841.

Smallbone, K., Simeonidis, E., Swainston, N., & Mendes, P. (2010). Towards a genome-scale kinetic model of cellular metabolism. *BMC systems biology*, *4*(1), 1–9.

Soh, K. C., Miskovic, L., & Hatzimanikatis, V. (2012). From network models to network responses: integration of thermodynamic and kinetic properties of yeast genome-scale metabolic networks. *FEMS yeast research*, *12*(2), 129–143.

Sousa, J. C. F., Faia, A. M., Ferreira, A. M., & Castro, L. T. (2010). Metabolismo microbiano produtor de energia. In *Microbiologia* (pp. 197–219). Lidel.

Sprenger, G. A. (1995). Genetics of pentose-phosphate pathway enzymes ofescherichia coli k-12. *Archives of microbiology*, *164*(5), 324–330.

Stephanopoulos, G., Aristidou, A. A., & Nielsen, J. (1998). Metabolic engineering: principles and methodologies.

Sultan, A., Jers, C., Ganief, T. A., Shi, L., Senissar, M., Køhler, J. B., ... Mijakovic, I. (2021). Phosphoproteome study of escherichia coli devoid of ser/thr kinase yeag during the metabolic shift from glucose to malate. *Frontiers in microbiology*, *12*, 657562.

Tsukatani, T., & Matsumoto, K. (2000). Flow-injection fluorometric quantification of succinate in foodstuffs based on the use of an immobilized enzyme reactor. *Analytica chimica acta*, *416*(2), 197–203.

Turányi, T., & Tomlin, A. S. (2014). Reaction kinetics basics. In *Analysis of kinetic reaction mechanisms* (pp. 5–37). Springer.

Valverde, J. R., Gullón, S., García-Herrero, C. A., Campoy, I., & Mellado, R. P. (2019). Dynamic metabolic modelling of overproduced protein secretion in streptomyces lividans using adaptive dfba. *BMC microbiology*, *19*(1), 1–13.

Vargas, F. A., Pizarro, F., Pérez, J. R., & Agosin, E. (2011). Expanding a dynamic flux balance model of yeast fermentation to genome-scale.

Varma, A., & Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and environmental microbiology*, *60*(10), 3724–3731.

von Kamp, A., & Klamt, S. (2014). Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS computational biology*, *10*(1), e1003378.

Waage, P., & Guldberg, C. M. (1986). Studies concerning affinity. *Journal of chemical education*, *63*(12), 1044.

Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., & Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, *26*(12), i255–i260.

Yugi, K., Nakayama, Y., Kinoshita, A., & Tomita, M. (2005). Hybrid dynamic/static method for large-scale simulation of metabolism. *Theoretical Biology and Medical Modelling*, *2*(1), 1–11.

Zimmerman, A. R., & Ahn, M.-Y. (2010). Organo-mineral-enzyme interaction and soil enzyme activity. In *Soil enzymology* (pp. 271–292). Springer.