T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MATHEMATICAL MODELING OF METABOLISM VIA TOP-DOWN AND BOTTOM-UP APPROACHES

MOHAMMAD JAFAR KHATIBIPOUR

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMICAL ENGINEERING

GEBZE

2020

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MATHEMATICAL MODELING OF METABOLISM VIA TOP-DOWN AND BOTTOM-UP APPROACHES

MOHAMMAD JAFAR KHATIBIPOUR

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMICAL ENGINEERING

THESIS SUPERVISOR

ASSOC. PROF. DR. TUNAHAN ÇAKIR

GEBZE

2020

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YUKARIDAN AŞAĞIYA VE AŞAĞIDAN YUKARIYA YAKLAŞIMLARLA METABOLİZMANIN MATEMATİKSEL MODELLENMESİ

MOHAMMAD JAFAR KHATIBIPOUR
DOKTORA TEZİ
KİMYA MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
DOÇ. DR. TUNAHAN ÇAKIR

GEBZE
2020

GEBZE TEKNİK ÜNİVERSİTESİ	**DOKTORA JÜRİ ONAY FORMU**

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 19/12/2019 tarih ve 2019/55 sayılı kararıyla oluşturulan jüri tarafından 25/12/2019 tarihinde tez savunma sınavı yapılan Mohammad Jafar KHATİBİPOUR'ın tez çalışması Kimya Mühendisliği Anabilim Dalında DOKTORA tezi olarak kabul edilmiştir.

JÜRİ

ÜYE (TEZ DANIŞMANI)	: Doç. Dr. Tunahan ÇAKIR
ÜYE	: Prof. Dr. Murat ÖZDEMİR
ÜYE	: Dr. Öğr. Üyesi Emrah NİKEREL
ÜYE	: Dr. Öğr. Üyesi Pınar PİR
ÜYE	: Dr. Öğr. Üyesi Enes Seyfullah KOTİL

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve ...../..... sayılı kararı.

SUMMARY

Biological cells are complex dynamic systems. Intracellular molecular interactions can be categorized into different interaction networks based on the nature of interactions and their function. Gene regulatory networks, protein-protein interaction networks and metabolic networks are examples of frequently studied and addressed networks in the community of systems and synthetic biology. It is necessary to study such complex systems in both experimental and theoretical playgrounds. In this research, the focus is on the analytical study of the metabolic networks and the construction of computational tools that can help to model, study and analyze cellular metabolism. Computational methods for determination of the active metabolic networks at different cellular conditions/contexts are broadly categorized into two groups. Bottom-up approach in the context of metabolism is referred to those methods that use a metabolic model of the organism as an input to estimate the active reactions at the specified condition/context through optimization, while constraining the solution space based on the available experimental data. On the other hand, top-down approach is referred to those methods that aim to infer the active metabolic network at a specified condition/context by analyzing the corresponding metabolome data. In this work, "JacLy" is introduced as a top-down Jacobian-based method that is developed to infer metabolic interactions of small networks from the covariance of steady-state metabolome data. Kinetic models of intracellular biochemical reactions are very suitable tools to study and analyze cellular metabolism. However, it is not an easy task to make reliable kinetic models of metabolic networks and there are many challenging obstacles in the procedure. As an effort to ease kinetic modeling of biochemical reaction networks and also to provide a suitable platform to reconcile top-down and bottom-up approaches, "Kinescope" is introduced as a computational tool developed in MATLAB for semi-automatic construction, simulation and analysis of kinetic models of metabolism. Applicability of both computational tools developed in this work, JacLy and Kinescope, is verified through *in silico* experiments.

Keywords: Computational Systems Biology, Dynamic Systems, Jacobian, Metabolic Network Inference, Elementary Reactions, Kinetic Modeling.

ÖZET

Biyolojik hücreler karmaşık dinamik sistemlerdir. Hücre içi moleküler etkileşimler, bu etkileşimlerin doğası ve işlevine bağlı olarak farklı etkileşim ağları şeklinde sınıflandırılabilirler. Gen regülasyon ağları, protein-protein etkileşim ağları ve metabolik ağlar, sistem biyolojisi ve sentetik biyoloji alanlarında sıklıkla ele alınan ve çalışılan ağlara örnek teşkil eder. Böylesi karmaşık sistemlerin hem teorik hem de deneysel alanda çalışılması gerekmektedir. Bu araştırmanın odağında, metabolik ağların analiz edilmesi ve hücresel metabolizmanın modellenmesinde, çalışılmasında ve incelenmesinde yardımcı olacak hesaplamalı araçların oluşturulması bulunmaktadır. Belirli bir koşulda aktif olan metabolik ağın belirlenmesi için kullanılabilecek hesaplamalı yöntemler iki ana grupta sınıflandırılabilir. Aşağıdan-yukarıya yöntemler, bir organizmanın metabolik modelini girdi olarak kullanarak ilgili koşulda aktif olan tepkimeleri optimizasyon yöntemiyle ve çözüm kümesini deneysel verilerle kısıtlayarak belirler. Yukarıdan-aşağıya yöntemler ise belli bir koşula ait metabolom verisini işleyerek o koşula ait aktif metabolik ağı belirlemeyi hedefler. Bu tez çalışması kapsamında, "JacLy" ismindeki yukarıdan-aşağıya yaklaşıma dayalı ve Jacobi temelli yöntem, kararlı-hal metabolom verisinin kovaryansından küçük ağların metabolik etkileşimleri tahmin etmek için geliştirildi. "Kinescope" isimli araç iseyarı-otomatik modelleme, simulasyon ve kinetik metabolik modellerin analizi için yeni bir hesaplamalı araç MATLAB'da geliştirildi. Bu tez çalışması kapsamında geliştirilen Jacly ve Kinescope araçlarının in silico deney verilerine uygulanarak validasyonu yapılarak uygulanabilirlikleri gösterildi.

Anahtar Kelimeler: Hesaplamalı Sistem Biyolojisi, Dinamik Sistemler, Jacobi, Metabolik Ağ Çıkarımı, Elementer Tepkimeler, Kinetik Modelleme.

ACKNOWLEDGEMENTS

Many people have helped me during my PhD studies, I am grateful to all of them. I would especially like to thank my advisor, Dr. Tunahan Çakır, whose help was not limited to advising me during my thesis research, but also, he has been continuously supporting me in different aspects of life, in and out of the academy. His patience and honesty in both teaching and research is admired. I am also thankful to the instructors and friends from the chemical engineering, bioengineering, molecular biology and genetics, and computer engineering departments for their kind helps. Especially, I would like to thank my friends in the bioengineering department offices 205 and 210 for providing such a warm and friendly atmosphere.

Finally, I am thankful to my parents, my sister and my brothers for their everlasting love and support. I dedicate this thesis to all the children, who are the owners of future and harbingers of peace.

TABLE of CONTENTS

	Page
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
TABLE of CONTENTS	viii
LIST of ABBREVIATIONS and ACRONYMS	xi
LIST of FIGURES	xii
LIST of TABLES	xiv
1. INTRODUCTION	1
2. LITERATURE REVIEW	3
2.1. Bottom-Up Approaches	3
2.1.1. Constraints Based on Transcriptome or Proteome Data	6
2.1.2. Constraints Based on Metabolome Data	8
2.2. Top-Down Approaches	10
2.2.1. Network Discovery Based on Time Series Data	10
2.2.2. Network Discovery Based on Steady-State Data	13
2.3. Kinetic Modeling of Metabolic Networks	15
2.3.1. Applications of Kinetic Models	18
2.3.2. Towards Large-Scale Kinetic Modeling of Metabolism	18
3. JACLY: A JACOBIAN BASED METHOD FOR INFERENCE OF METABOLIC INTERACTIONS FROM THE COVARIANCE OF STEADY STATE METABOLOME DATA	21
3.1. Methods	23
3.1.1. Problem Definition	23
3.1.2. Optimization Pipeline	25
3.1.3. Constraining the Solution Space by Generating Sparse Individuals	26
3.1.4. Scanning for The Scaling Factor	28
3.1.5. Fluctuation Vector	28
3.1.6. Using a Community of Estimated Jacobians Instead of Only One Elite Jacobian to Infer Structure of The Network	29

3.1.7. Quantification of Inference Performance 31
3.2. Results 32
3.2.1. Use of *in silico* Covariance Matrices for Metabolic Models of *S. cerevisiae* and *E. coli* 32
3.2.2. Use of *in silico* Metabolome Data for Metabolic Models of *S. cerevisiae* and *E. coli* 34
3.3. Discussion 38
4. KINESCOPE: A TOOL TO EASE KINETIC MODELING OF METABOLIC PATHWAYS 41
4.1. Method 44
4.1.1. Construction of the Ensemble 49
4.1.2. Automatic Break Down of the Enzymatic Reactions to their Elementary Steps 53
4.1.3. Manual Curation of the Elementary Steps 54
4.1.4. Automatic Construction of the Kinetic Model 56
4.1.5. Automatic Construction of the Symbolic Jacobian Matrix 60
4.1.6. Thermodynamic Constraint 62
4.1.7. Calculating the Elementary Reactions Rates 66
4.1.8. Sampling the Enzyme Fractions and Calculating the Rate Constants 69
4.1.9. Collecting Stable Models in the Ensemble 70
4.1.10. Screening the Ensemble 71
4.2. Results 73
4.2.1. Using Kinescope to Collect Stable Kinetic Models Satisfying Different Reaction Deletion Studies while Lumped-Kinetic Rate Expressions are Available 78
4.2.2. Using Kinescope for Automatic Construction and Collection of Stable Kinetic Models at the Elementary Reactions Level that Satisfy Different Reaction Deletion Studies 87
4.3. Discussion 89
4.3.1. A Note on Parameter Identifiability 90
4.3.2. A Note on the Required Computational Time and Curse of Dimensionality 92
5. USING KINESCOPE FOR KINETIC MODELING OF THE CENTRAL CARBON METABOLISM OF *E. COLI* AND FUTURE WORKS 94

5.1. Using Kinescope for Kinetic Modeling of the Central Carbon Metabolism of *E. coli* at the elementary reactions level 95
5.2. Future Works 99
6. CONCLUSIONS 101

REFERENCES 103
BIOGRAPHY 115

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		**Explanations**
Γ	:	Covariance Matrix
ΔG	:	Gibbs Energy Change
$\Delta G°$	:	Standard Gibbs Energy Change
CSTBR	:	Continuous Stirred Tank Bioreactor
D	:	Fluctuation Matrix
EM	:	Ensemble Modeling
EMEC	:	Elite Models Error Cutoff
FBA	:	Flux Balance Analysis
FIM	:	Fisher Information Matrix
FPR	:	False Positive Rate
FVA	:	Flux Variability Analysis
GA	:	Genetic Algorithm
GGM	:	Graphical Gaussian Modeling
GMA	:	Generalized Mass Action
GMEC	:	Good Models Error Cutoff
GPR	:	Gene Protein Reaction
J	:	Jacobian Matrix
MCA	:	Metabolic Control Analysis
MER	:	Main Elementary Reaction
MM	:	Michaelis-Menten
ODE	:	Ordinary Differential Equation
PC	:	Principal Component
SDE	:	Stochastic Differential Equation
SER	:	Side Elementary Reaction
TPR	:	True Positive Rate

LIST of FIGURES

Figure No.		Page
1.1:	Comparative demonstration of bottom-up and top-down approaches.	2
3.1:	Selecting a bounded area around the elite Jacobian.	30
3.2:	A schematic example of alignment and combination of Jacobian vectors in the selected community to come up with the final structure.	31
3.3:	The Spearman correlation between the predicted Jacobian matrix values by JacLy and the calculated values from the kinetic models.	37
4.1:	An example of a complex system with limited number of components.	42
4.2:	Modeling Cycle. Theory and experiment are complementary.	42
4.3:	MM-based kinetic modeling of a small metabolic network.	47
4.4:	Collecting stable models with different parameter sets that all converge back to the reference steady-state after a small perturbation.	50
4.5:	Screening the ensemble.	50
4.6:	Flowchart of the algorithm for construction of the ensemble.	52
4.7:	Automatic break down of enzymatic reactions to their elementary steps.	53
4.8:	Manual curation of the elementary steps.	55
4.9:	Representation of the automatically constructed kinetic model at the elementary reactions level.	59
4.10:	Flowchart of the algorithm for screening the ensemble.	72
4.11:	Reference Steady State.	75
4.12:	Kinescope, construction tab.	76
4.13:	Kinescope, curation tab.	76
4.14:	ToyModel response to a perturbation-observation experiment.	77
4.15:	Kinescope, simulation/screening tab.	78
4.16:	Schematic representation of the *in-silico* reaction deletions in the ToyModel.	79
4.17:	In-silico reaction deletion experiments.	79
4.18:	Unsupervised uniform sampling of the kinetic parameter space.	81
4.19:	Distribution of the good models based on the first experiment.	82
4.20:	Distribution of the good models based on the third experiment.	83
4.21:	Distribution of the elite models based on the first experiment.	84

4.22: Distribution of the elite models based on the second experiment. 84
4.23: Boxplots for the kinetic parameters based on their values in the collected thirteen models. 86
4.24: Simulation profiles of the fold changes in the concentration of metabolites based on the experiment 3. 87
4.25: Unsupervised sampling of the kinetic space for kinetic models at the elementary reactions level. 88
5.1: Schematic representation of a continuous stirred tank bioreactor. 95
5.2: Unstable models with an apparent steady-state may start to diverge when simulated for a longer time. 98
5.3: Response of one of the models with apparent stability to an impulse to the concentration of the extracellular glucose. 99

LIST of TABLES

Table No.		Page
2.1:	Different levels of information on metabolic networks.	11
3.1:	Inference results for the *in silico* metabolome data, comparison of JacLy and GGM.	35
3.2:	A summarized comparison of JacLy with GGM.	39
4.1:	Codes for different regulation mechanisms.	51
4.2:	Evaluation table after unsupervised sampling of the parameter space.	85
4.3:	Evaluation table after the supervised sampling of the parameter space.	86
4.4:	Evaluaiton table after the unsupervised sampling of the kinetic space for the models at the elementary reactions level.	88
4.5:	Computational time and number of successfully screened models versus the number of samples taken from the parameter space.	92
5.1:	Breaking down the biomass production reaction to its constituent reactions.	97

1. INTRODUCTION

Metabolic network is the outmost layer of cellular activity from the genome. The genome of a cell is a comprehensive and condensed information base, defining a boundary for the biochemical capacity of the cell. The processing of genetic information passes through several layers of fabrication and regulation before reaching their end products. This is from information to the function, from genotype to the phenotype. Metabolic enzymes count for a significant percentage of the end products of genes, and their activity sets the physiology of the cell. Since metabolic network activity is the major representative of cell functionality, it is of great importance to gain as much knowledge as possible on the active metabolic network at a specific cellular state.

Systems-based approach to molecular biology has contributed to an increased knowledge of metabolic pathways for an increasing number of organisms and led to almost complete metabolic networks for a number of major organisms, from yeast to human. Such static networks are available in a condition-independent manner through web-based databases such as KEGG or MetaCyc [1], or reconstructed in a format suitable for simulation by several researchers at genome scale [2, 3]. There are several mathematical approaches to process such networks to come up with condition-specific networks, the most common one being the Flux-Balance Analysis (FBA) framework [4]. This is a bottom-up direction toward the active network since already-known "parts," interactions, are used as inputs [5, 6].

In parallel to the developments on the knowledge of metabolic networks, techniques to measure metabolite levels at high throughput, termed metabolomics, have arisen [7, 8]. Quantitative or semi-quantitative metabolome data, although one of the most challenging compared to other omic sciences, have come a long way in a decade, from the detection and quantification of about 50 metabolites [9] to more than 1000 metabolites [10]. Metabolome data are a snapshot of the condition-specific status of the investigated organisms. Reverse-engineering metabolome data to discover the underlying network structure is the goal behind metabolic network inference approaches [11, 12]. The information content of metabolome data is revealed by processing it with correlation or optimization-based methods [13–15]. Such an approach to discover metabolic network structure is termed top-down approach since the parts, interactions, are not known a priori, and predicted from the whole set of

available biomolecules [5, 6]. Figure 1.1 illustrates the two alternative network discovery approaches.

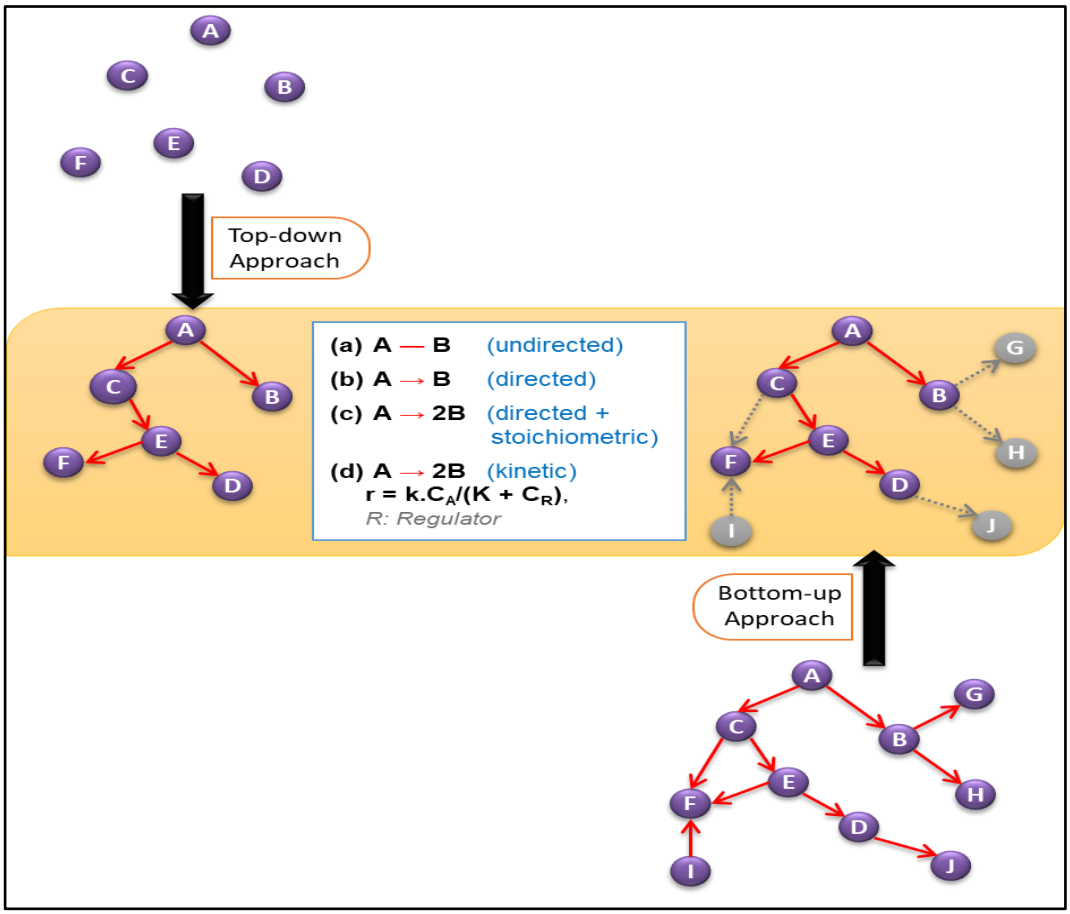


Figure 1.1: Comparative demonstration of bottom-up and top-down approaches to discover active metabolic network.

The research in this thesis can be mainly divided into two parts. The first part is about top-down and bottom-up approaches in discovery and analysis of metabolic networks. Chapter 3 presents "JacLy: a Jacobian-based method for the inference of metabolic interactions from the covariance of steady-state metabolome data" as a product of our research in this part. The second part focuses on design and production of a computational tool for semi-automatic construction, simulation and analysis of kinetic metabolic models, believing that kinetic models can provide an efficient platform to reconcile top-down and bottom-up methods in studying metabolic networks. Hence, the MATLAB-based computational tool "Kinescope" is presented in Chapter 4.

2. LITERATURE REVIEW

2.1. Bottom-Up Approaches

Different methods and algorithms have been used for the discovery and characterization of active metabolic networks at different states of cells and culture environments. In the bottom-up approach, everything starts from an already available network of biochemical transformations that covers all possible scenarios in the distribution of metabolic fluxes and sets an upper bound for the existence of reactions in the active metabolic network. Such a network is termed a static metabolic network. A static metabolic network can be provided either by a previously reconstructed genome-scale stoichiometric model or by a collection of all reactions whose existence in the organism of interest has been certified in literature and databases. Most popular among such databases are KEGG [16], MetaCyc [17], and Reactome [18]. Other efforts with more curated databases such as Rhea [19] and MetRxn [20] are also available. A genome-scale stoichiometric model is reconstructed based on the annotation of all genes in the genome of one organism to their end products and then to the corresponding reactions, leading to a list of Gene- Protein-Reaction (GPR) rules [21]. In this way, the minimum information content of a genome-scale model is (i) a list of reactions, and (ii) a list of gene-protein-reaction rules. The presence of gene-protein-reaction rules in stoichiometric models has enabled the opportunity for transcriptome and proteome data to be incorporated into the discovery methods of active metabolic networks [22].

Given a genome-scale reaction network, the aim is to find the active reaction network at a specific condition or for a specific cell type in a multicellular organism. The core of all such discovery approaches is a stoichiometric matrix. Each row of the stoichiometric matrix represents a metabolite and each column stands for a reaction, the corresponding element being the stoichiometric coefficient of that metabolite in that reaction. The relationship between the reaction rates in the network and the dynamic change in the concentration of metabolites is represented as given below:

$$\frac{dC}{dt} = S.v \tag{2.1}$$

where S is the stoichiometric matrix, C is the vector of intracellular metabolite concentrations, and v is a column vector of metabolic reaction rates (fluxes) to be determined. Under the assumption of steady state, the concentration of each intracellular metabolite is not going to change with time, meaning the sum of rate of reactions producing that metabolite is equivalent to the sum of rate of reactions consuming that metabolite (metabolic fluxes around each metabolite are balanced). This is represented mathematically as following:

$$S.v = 0 \tag{2.2}$$

This is an algebraic system of linear equations with all fluxes being zero as a trivial solution. In order to escape from the trivial solution, the value of at least one of the fluxes must be set to a non- zero value, that flux usually being an exchange flux between the intracellular and extracellular environment since the experimental measurement of exchange fluxes is relatively easier. The system is almost always underdetermined with a large solution space, mainly because of the existence of branch points in the metabolic network.

There are both experimental and computational approaches to estimate a condition-specific network for such a system. The experimental approach is based on stable-isotope (mostly ^{13}C) labeling of the major carbon source, and then tracing the propagation of the labeled carbon atoms down to protein-bound amino acids at isotopic steady state by using mass spectrometry or NMR spectroscopy [23–25]. The qualitative isotopic labeling information is then used as an input to two alternative methods. In one method, termed isotopomer modeling, a total flux distribution is estimated based on the experimental labeling results through a computationally demanding non-linear optimization formulation, which employs global iterative fitting and statistical analysis [23, 26]. The other ^{13}C- labeling assisted method is based on the estimation of the local ratios of fluxes emerging from a branch point [24, 27] rather than the absolute quantification of all fluxes. These experimental flux split ratios can be used to shrink the solution space of Equation 2.2 in a complementary flux calculation, leading to the discovery of a condition-specific network [28, 29]. Softwares are available for the rather sophisticated calculation of experimental fluxes (or flux ratios) from carbon labeling data for both methods [30–32]. A new trend in this area is to collect data at the non-stationary phase of isotopic labeling rather than

at the isotopic steady state, which was shown to be more informative in terms of predicting the flux-weighted active metabolic network structure [33–35]. Works on the tracing of intracellular metabolites rather than only 10–15 protein-bound amino acids have also appeared due to the higher coverage of metabolic pathways despite the inherent experimental difficulties in terms of higher turnover rates as well as stability issues [36–38].

The computational approach for the discovery of condition-specific metabolic network based on Equation 2.2 is known as constraint-based modeling. Constraint-based modeling methods aim to shrink the solution space of the equation as much as possible by putting relevant constraints on the system. The most common method, FBA, treats the problem in Equation 2.2 as an optimization problem and linear programing is applied to solve it. The stoichiometry of metabolic reactions (stoichiometric matrix), reaction directionality information, a physiologically relevant objective function, and the value of at least one of the exchange fluxes are all that are required for FBA to return a condition-specific flux distribution. The flux distribution returned by FBA is not necessarily unique, and there may be a variety of flux distributions all leading to the same optimum value of the objective function. Therefore, Flux Variability Analysis (FVA) must be used together with FBA, to determine the variability, if any, on each metabolic flux in regard to the condition of interest [39, 40]. The maximization of biomass production has been successfully applied as a reliable objective function for FBA to predict flux distributions in a variety of microorganisms [41, 42]. In some studies, it has been hypothesized that one objective function alone may not capture the metabolic behavior of the cell comprehensively. Therefore, multiobjective optimization platforms have been designed and utilized to come up with more specific flux distributions. Several modified versions of FBA including parsimonious FBA (pFBA) [43], and flexible-optimality FBA (flexoFBA) [29], have been developed in this manner. On the other hand, some research groups have developed methods based on the availability of additional omics data, which are discussed below. For a thorough review of a number of FBA-derived flux calculation methods, the readers are referred to another resource [44].

2.1.1. Constraints Based on Transcriptome or Proteome Data

The rate of an enzymatic reaction inside the cell is a function of several different factors, such as the concentration of substrates, products, and regulators of the enzyme and also the amount of available active enzyme for that reaction. Among these factors, the concentration of active enzymes can be related to the activity of genes through layers of transcription, translation, and post-translational modifications. Transcriptome data are much more accessible and comprehensive compared to the other omics data. Several different research groups have developed different strategies to incorporate transcriptome data into constraint-based models. The idea behind this is that the amount of mRNAs (gene activities) may be correlated with the concentration of active enzymes, and hence this can be utilized to provide additional constraints on metabolic fluxes. At the bottom line, if an enzyme coding gene is not transcribed at steady state, the corresponding reaction should be inactive at that steady state, if there is no other enzyme catalyzing that reaction. This idea was first used by Akesson *et al.* to set the flux values to zero for those reactions whose corresponding genes were expressed at low levels [45]. More sophisticated and structured versions of this approach appeared later, under the names of GIMME [46] and iMAT [47]. These approaches classify some reactions as inactive reactions based on the low expression levels of their associated genes. Then, they employ a computational framework which minimizes the contradiction between the classification and an active physiological flux distribution since some of these classifications may render the flux state unrealistic (such as zero growth rate). Several other alternative methods appeared recently to incorporate transcriptome data into the prediction of active metabolic network and flux distribution. In an interesting study for example, mRNA levels from transcriptome data were used as weights for the corresponding reactions to predict a flux distribution without using a conventional objective function such as the maximization of biomass growth [48]. A study [49] evaluated these methods systematically for the prediction of flux distributions, and the results were compared to that of parsimonious FBA as a reference method that does not consider the transcriptome data. In general, none of the methods could significantly improve the results of pFBA and none of them outperformed the others for the tested cases (*S. cerevisiae* and *E. coli*). Instead of the prediction of flux distributions, these methods, however, may significantly help in the discovery of active metabolic networks in context/tissue-specific cells and in the

conditions where a relevant objective function cannot be hypothesized. Transcriptome data are not necessarily correlated with the rate of corresponding reactions. Inconsistency between mRNA levels and reaction rates is a result of influence of several other factors in the regulation of enzymatic reactions. Therefore, if proteome data are available, it can be used instead of transcriptome data as a better representative for the concentration of active enzymes since proteome is hierarchically closer to the enzyme states than transcriptome data. The methods that are developed to integrate transcriptome data with the FBA method can all be used for the purpose of integrating proteome data. For example, GIMMEp [50] is the proteome equivalent version of GIMME. Some of such integrative methods were primarily tested with proteome data. INIT [51], for example, was developed by using proteome abundance data from Human Protein Atlas database. However, it was shown that utilizing proteome data instead of transcriptome data could not improve the prediction of flux distributions for the tested cases (*S. cerevisiae* and *E. coli*) [49]. In a study which used metabolome and proteome data in the flux calculation method, on the other hand, even the use of only proteome data was shown to improve the results compared to the traditional FBA [52].

Substrate concentrations, the concentration of enzyme regulators, the turn over number of the catalyzing enzyme, and the concentration of the active enzyme are all playing significant roles in the determination of reaction rates, and among them only the concentration of the active enzyme may be represented by the corresponding protein or mRNA concentration. Translated proteins are not necessarily active enzymes, and they may need to undergo post-translational modifications (e.g., phosphorylation/acetylation) to become capable of catalyzing the reactions. This is one of the main reasons behind inconsistency between protein levels and reaction rates. On the other hand, the turn over number (catalytic power) of one enzyme may differ by several orders of magnitude from the turn over number of another enzyme [53]. It means that although the concentration of one enzyme may be much less than the others in the network, the reaction catalyzed by that enzyme can proceed much faster than others. According to this fact, the use of the absolute concentrations of proteins or mRNAs to constrain reaction rates does not seem promising. However, the turnover number of one enzyme in an individual is an intrinsic parameter of the enzyme that does not change from one condition to another except by effective mutations that rarely occur. Because of this, the relative levels of proteins or mRNAs can be utilized to overcome the problem of big differences in turn over numbers. One steady-state

condition with available data on flux values and protein/mRNA levels can be taken as the reference state, and then the relative/differential levels of proteins/mRNAs to that reference state can be used to predict the flux distributions at the new conditions. Based on this approach, algorithms have been developed to incorporate relative/differential transcriptome data into metabolic-flux analysis, among which are MADE [54] and GX-FBA [55]. One other main reason for the inconsistency between protein levels and reaction rates is the distribution of flux control among different layers from genotype to phenotype. Metabolic fluxes can be regulated hierarchically (through gene expression levels) or metabolically (through metabolic interactions) [56–59]. Use of transcriptome or proteome data will not be helpful if the metabolic fluxes are controlled at the metabolic level.

2.1.2. Constraints Based on Metabolome Data

One approach to find more specific and physiologically relevant flux distributions is to provide additional constraints by specifying the directionality of reversible reactions. This can be done by taking Gibbs free energies of metabolites into consideration. The Gibbs free energy change of a reversible biochemical transformation (one reaction or a series of reactions) determines the direction of that transformation and its departure from reversibility. The earlier studies assumed standard conditions (all metabolite concentrations were assumed to be 1 M) and did not explicitly consider metabolite concentrations in the calculation of Gibbs energy changes of reactions due to the scarcity of metabolome data [60]. Recent studies, however, take the concentration of metabolites into account, when available, to perform thermodynamic-based metabolic flux analysis, leading to more reliable predictions [61–64].

Extracellular metabolome data can be used to constrain genome-scale metabolic models for the calculation of intracellular flux distributions by simply constraining the secretion and uptake rates of extracellular metabolites based on such data [65, 66]. In a different approach, Michaelis–Menten-based kinetics was used for the estimation of reaction rates for the reactions for which appropriate intracellular metabolome (and proteome) data are available [52]. The FBA framework was designed in such a way that the calculated fluxes are as consistent as possible with the kinetically derived reaction rates, if available. The simultaneous use of metabolome and proteome data

for this purpose significantly improved the results. The use of metabolome data alone also resulted in better predictions than the traditional FBA. In a recent study, a kinetic platform was established based on Michaelis–Menten equation to bridge gene expression levels, metabolite concentrations and metabolic fluxes without requiring the knowledge of kinetic parameters [67]. They could show that changes in metabolite concentrations relative to a reference steady state can be predicted by their formulation that includes information on network connectivity in addition to differential mRNA expression levels. All those works utilizing kinetic information demonstrate the necessity of dynamic models for a more comprehensive analysis of metabolic networks.

Kinetic models of biochemical reactions not only provide a rational platform for omics data –especially metabolomics – to be incorporated in the estimation of metabolic fluxes, but also, they enable the prediction and study of the dynamics of metabolic networks far beyond the steady state. Such models were only possible for small-scale metabolic networks until recently [68, 69], since they require detailed information on the enzyme kinetics of each individual reaction. Estimation of kinetic parameters is a major obstacle in the applicability of dynamic modeling of metabolic networks. New platforms and algorithms were established to circumvent this problem so that the estimation of explicit kinetic parameters is not a prerequisite to study the dynamic capacity and behavior of the system [70]. Approximative kinetic models (lin-log, power-law, mass action) on the other hand, try to fit a standard rate expression formula to all reactions of the network to increase the range of their applicability to larger networks [71, 72]. Thanks to approximative kinetics, attempts to reconstruct large-scale kinetic metabolic models with more than 100 reactions were recently presented [73–75], but their prediction power is limited to the conditions adequately close to the corresponding steady state.

As a better alternative to approximative kinetics, an approach was established and utilized based on the concept of parametric Jacobian, which covers the behavior of all possible kinetic models that are consistent with an experimentally observed operating point [76]. This approach provides an opportunity to detect and analyze bifurcation characteristics of the metabolic network without the need for explicit determination of kinetic parameters. Ensemble modeling of metabolic networks [77] is an elegant idea for large-scale kinetic modeling of biochemical reaction networks. In this method, each enzymatic reaction is broken down to its elementary reactions

that all follow mass-action kinetics. An ensemble of thermodynamically consistent kinetic models with different dynamic behavior that all converge to a reference steady state is collected with the help of intracellular metabolome data. This ensemble is then filtered by the results of perturbation experiments to filter out inconsistent models from the ensemble and to increase the predictability of remaining models. The approach was successfully applied, among others, to construct kinetic models of *E. coli* [78] and cancer metabolisms [79], leading to promising flux predictions. Table 2.1 summarizes different levels of information one might obtain about a metabolic network based on different modeling approaches and availability of experimental data.

2.2. Top-Down Approaches

Time series of metabolite concentrations in response to a perturbation, and also replicates of metabolome data at a specific steady state, both implicitly contain information on the structure of active metabolic network. Reverse engineering of these data to infer the condition-specific metabolic network without necessarily prior knowledge on the genome of the organism and its static metabolic network is an alternative to all bottom-up approaches that are based on the availability of a large-scale stoichiometric model of the organism. Although promising, less attention has been paid to these top-down approaches compared to bottom-ups mainly because of the technical obstacles in gathering reliable metabolome data in large scale. This limitation will be removed with future advancements in the detection and quantification of intracellular metabolites such as higher coverage and temporal resolution. At this stage, however, several research groups have established algorithms and methods for reverse engineering of metabolic networks by using either time series or steady-state replicates of metabolite concentrations [14, 80–82].

2.2.1. Network Discovery Based on Time Series Data

The use of time-series metabolite concentration data to predict the underlying network connectivity information first appeared in the literature about two decades ago. Time-lagged correlations combined with a projection technique called multidimensional scaling were shown to construct the structure of generic biochemical networks with few nodes [83].

Table 2.1: Different levels of information on metabolic networks.

1	At the lowest level of information, one wants to know what the structure of the network is, representing it with an undirected (or directed, if the reversibility information is available) graph in which each node stands for a metabolite and each edge stands for a biochemical transformation. Alternative to the retrieval from the metabolic reaction databases, the structure of the network – both directed and undirected – can also be estimated to some extent by analyzing and reverse engineering the metabolome data without the use of *a priori* database information on the reactions.
2	At a higher level, the information on the stoichiometry of reactions can be incorporated, leading to a directed stoichiometric biochemical network.
3	Having the stoichiometric structure of the network, one can characterize the metabolic state in more detail by quantifying the metabolic fluxes. In most cases, rather than a unique flux distribution, constraints are set on flux values to shrink the solution space. Such modeling approaches are known as Constraint-Based Modeling. This level of understanding the active metabolic network (structure + flux distribution) has been the area of focus in the research community for more than a decade. In most cases, the information provided at this level has been satisfactory for engineering research to design more efficient cell factories and also recently for medical research to distinguish significant differences between healthy and disease states.
4	There are however certain limitations at the above level although it provides a network activity structure weighted with fluxes. The dynamic behavior of the system cannot be captured, and the predictability power of such models is hampered mainly because they are not considering the role of regulatory mechanisms in controlling the rate of biochemical reactions. In some cases, the regulation of reaction rates plays such a dominant role that it would be hard to make any prediction by just considering the flux-based network activity structure. Here come the kinetic models into the picture, which take enzymatic regulations and metabolite concentrations into account for a dynamic and better prediction of network structure.

Correlation between time-series profiles of metabolites, with the consideration of the delay in the influence of one metabolite on the next, is the basis of the time-lagged correlation method for the inference of metabolic networks. The approach, called correlation metric construction, was later experimentally verified *in vitro* by inferring the first steps of glycolytic pathway in a 14-metabolite system [84]. Modified versions of the approach appeared later [85, 86]. In the latter, metabolic pathway of an

anticancer drug was deduced from the time-lagged correlations of corresponding metabolite concentration measurements. The modification introduced by the former work was recently improved by using mutual information similarity score rather than simple linear correlation [87]. The authors compared their method, called MIDER, with several other methods by applying it to different types of cellular networks, including *in vitro* glycolytic pathway data. The approach outperformed the other methods.

Another method to reconstitute a network using time-series data is based on perturbation experiments around steady state. The initial curve of concentration changes of metabolites in response to a pulse change on the concentration of a metabolite is processed with the method of zero initial slopes [88]. The method successfully inferred the structure of glycolysis based on *in vitro* experimental data [89]. Performance comparison of the method with the correlation metric construction approach was later provided based on *in silico* data of *S. cerevisiae* and *E. coli* central metabolic networks [14]. An approach based also on perturbation experiments, but with a different formulation aiming to calculate Jacobian matrix from time derivatives of concentration data, was first applied to gene networks [90]. A modified version of the approach recently used *in vivo* metabolite concentration measurements from tomato seedlings to reconstruct quercetin glycosylation pathway [91].

Apart from such model-free structure identification methods, model-based methods use time-series metabolite concentration data not only to identify network structure but also to estimate proper model parameters such as rate constants of kinetic expressions [81]. Majority of these approaches use power-law (also called S-system) formulation [92] to approximate reaction kinetics. An approach, for example, used S-system modeling with a multi-objective optimization by simultaneously minimizing the number of interactions and the error in the fitting [93]. They applied their method to major metabolites involved in ethanol fermentation. An earlier work analyzed a small three-metabolite network of phospholipid metabolism by combining S-system modeling and an evolutionary modeling method, genetic programing [94]. Later, a new representation of S-system approach, called S-trees, was combined with genetic programing to reverse-engineer yeast fermentation pathway in a more efficient manner by using *in silico* time-series concentration data of five metabolites [95]. In a sophisticated approach, others used symbolic regression based on genetic programing to infer both the structure and the model of yeast glycolytic oscillations from *in silico*

data [96]. Their use of acyclic graph encoding rather than tree-based encoding together with symbolic regression approach ensured the identification of parsimonious (sparse) models. Rather than S-system formulation, mass-action kinetics can also be used to infer pathway connectivity and reaction mechanism [11]. This minimizes the computational burden on the algorithm since only rate constants are to be estimated as parameters in the mass-action formulation. The authors tested their method with real time course experimental metabolome data of *Lactococcus lactis* glycolysis. A graphical user interface was later made available by the same group to ease the inference of kinetics and network architecture from dynamic data of biochemical pathways [97]. Genetic programing was also combined with mass-action kinetics in an algorithm, which ensures the estimation of biochemically more plausible models [98]. The small phospholipid network of [94] was inferred in a more compact way by this algorithm.

2.2.2. Network Discovery Based on Steady-State Data

The use of steady-state metabolome data to infer metabolic network structure has also drawn attention in the last decade. The biological variability in the metabolism of the organisms at around steady state is a known phenomenon due to slight variations in the enzyme levels or due to slight natural or environment-induced fluctuations within cellular processes. Slight variations in the steady-state measurements of metabolite levels can be informative on the network structure [12, 99, 100]. The most common approach here is to use the similarity measures such as Pearson correlation to assign edges between metabolites. One should note that such correlations are not necessarily strong among neighboring metabolites whereas there could be strong correlations among distant metabolites in the network [100]. In a comprehensive study, different alternative similarity measures (linear vs. nonlinear, and full vs. partial) were applied to *in silico* metabolome data belonging to two microorganisms to systematically analyze method performances [12]. The results revealed no clear superiority between linear (Pearson correlation) and nonlinear (mutual information) similarity measures. The best performing method was identified as nth order partial Pearson correlation, known also as Graphical Gaussian Modeling (GGM). Graphical Gaussian Modeling was also applied to metabolome data from blood serum samples to reconstruct human fatty acid metabolism [101]. Others [102] analyzed *in silico*

metabolome data of red blood cell metabolism by ARACNE approach [103], which is based on pruning mutual information scores. An elegant improvement on ARACNE based reverse engineering of metabolic profiling data was suggested later [104]. The approach puts a constraint on the possible metabolic transformations to satisfy the mass conservation between the connected metabolites. Synthetic data covering up to about 200 metabolites were generated to test the approach. One issue in such similarity-based approaches is that only pairwise interactions are aimed to be found. However, a metabolic reaction can involve more than two metabolites. Based on this reasoning, an attempt to also deduce triple interactions by using ternary mutual information was suggested [105]. Analysis of synthetic yeast glycolysis data and red blood cell data showed the success of this approach in capturing higher order interactions.

A different approach to discover active metabolic networks from steady-state data is based on Lyapunov equation. In Equation 2.1, the rate vector, v, is a complex non-linear function of concentrations, C. For systems around steady state, the equation can be expressed in terms of Jacobian matrix, J, by the help of linear approximation:

$$\frac{dX}{dt} \approx JX \tag{2.3}$$

With $X = C - C_s$, and C_s shows the steady-state metabolite concentrations. Jacobian matrix stores detailed information on the structure of the underlying network; such as the directionality of interaction, strength of interaction, and regulation type of interaction. For small fluctuations around steady-state, the right-hand side of Equation 2.3 becomes zero, and the left-hand side can be expressed in such a way that a link between the covariance matrix of metabolome data, Γ, and Jacobian matrix is provided. The details of the derivation are given elsewhere [99, 106].

$$J\Gamma + \Gamma J^T = -2D \tag{2.4}$$

D in the equation shows the extent of fluctuations. Equation 2.4, known as Lyapunov equation, can be used to infer metabolic network structure since it provides a link between the data-based covariance matrix and network connectivity stored in J. Reverse-engineering metabolome data by using the Lyapunov equation was first

discussed via a hypothetical three-metabolite system [99]. In Chapter 3, a novel computational tool, JacLy, is introduced, which was developed during this thesis study for the inference of small-scale metabolic pathways. JacLy takes the replicates of steady-state data from repeated measurements at the same steady-state condition and then utilizes the Lyapunov equation to estimate the Jacobian matrix of the system at that condition.

2.3. Kinetic Modeling of Metabolic Networks

Kinetic models of cellular metabolism are valuable tools for rational design of metabolic engineering strategies and to describe complex behavior of biological systems. Kinetic models have been used for a large variety of applications such as redesign of metabolic systems [71, 107–110], in synthetic biology [111] and for the estimation of optimal drug concentrations [112]. Although in theory kinetic models are better representatives of biochemical reaction networks compared to the stoichiometric models, their usage in practice has been hindered because of several issues among which the following two are the major ones: (i) lack of sufficient reliable data suitable for kinetic modeling [113, 114] (mainly due to the absence of a standard technology and protocol to monitor, measure and quantify dynamics of metabolism), and (ii) the fact that kinetic parameters measured *in-vitro* are not applicable to the modeling of intracellular biochemical reactions [115] (mainly due to the harsh differences between the intracellular micro-environment and the micro-environment in the test tubes where isolated enzymes are characterized). To circumvent this issue, some researchers utilized *in-vivo* data that usually includes fast measurement of intracellular metabolite concentrations at different time points after a stimulus experiment [69, 116, 117]. They used such data as input to parameter estimation algorithms to find appropriate kinetic parameters for their models. However, computational estimation of *in-vivo* kinetic parameters has also been proven to be a very difficult task with a very low success rate [118, 119]. Below, a brief review on the history of development of kinetic modeling of biochemical reactions will be provided and recent approaches towards kinetic modeling of large-scale networks along with their limitations and challenges will be discussed.

First publications on the study of rates of enzymatic reactions appeared in the scientific literature more than a century ago [120]. The work of Henry [121], Michaelis

and Menten [122] and others [120] had a significant impact on enzymology and biochemistry. The well-known rate expression of Michaelis and Menten and their experimental design based on the *in-vitro* measurement of the initial reaction rates at different substrate and/or inhibitor concentrations have been frequently used for kinetic characterization of many enzymes. Corresponding measured kinetic parameters are available in enzyme databases such as BRENDA [123]. These parameters and underlying Michaelis-Menten (MM) based rate expressions have proved useful in several fields such as Enzymatic Membrane Reactors [124], Enzyme Biosensors [125, 126] and others that deal with a limited number of isolated enzymes. However, the use of MM-based kinetics and corresponding parameters does not seem promising for the modeling of cellular metabolism [115, 118, 119]. Major reasons are given below for why the MM-based kinetic modeling approach might not be suitable for kinetic modeling of metabolic networks, especially the large-scale networks.

The first reason is that each of the MM-based rate expressions are deduced based on one or more simplifying assumptions that might not hold true at intracellular conditions. For instance, MM rate law itself assumes that among the elementary steps, "release of the product" is the rate limiting step. Also, it comes with the assumption that there is no difference between the enzyme-substrate and enzyme-product complexes. That is why both complexes are lumped into a single intermediate in the derivation of the MM rate law. Equations 2.5 to 2.7 show the complete set of elementary reactions for an irreversible mono-substrate enzymatic reaction that is not regulated by any means.

$$E + S \rightleftarrows ES \tag{2.5}$$

$$ES \rightarrow EP \tag{2.6}$$

$$EP \rightarrow E + P \tag{2.7}$$

E stands for the free enzyme, ES and EP are enzyme-substrate and enzyme-product complexes respectively and P stands for the product. The same reaction is described in the following elementary steps according to Michaelis and Menten:

$$E + S \rightleftarrows X \tag{2.8}$$

$$X \rightarrow E + P \tag{2.9}$$

X is the common intermediate during transformation of S to P, representing lumped form of ES and EP complexes. Although such simplifying assumptions might be appropriate for most of the enzymatic reactions, in cases where they are not valid for a few reactions, they can lead to very large deviations from the inherent dynamics of metabolism while simulating networks of biochemical reactions. This is mainly because the error in the calculation of the rate of even one reaction at each time point of the simulation can be propagated, influencing the dynamics of the whole network during the course of simulation. It must be mentioned that such assumptions made it possible for Michaelis-Menten and others to mathematically describe the rate of enzymatic reactions without any need for measurement of concentration of different enzyme forms (free enzyme and its different complexes), which otherwise would not be possible.

The second reason is that the MM-based rate expressions for multi-substrate enzymes that are also regulated by one or more regulators are highly nonlinear with many unknown parameters, leading to very complex and stiff sets of differential equations. Even if parametrically identifiable, parameter estimation takes a great deal of time and effort. Also, those kinetic parameters that were measured *in-vitro* cannot be directly used in the modeling of biochemical reactions in the intracellular conditions, as it was discussed at the beginning of this section.

Even after a successful parameter estimation, these models are not systematically open to investigate the uncertainties regarding the kinetic mechanisms (the elementary steps and their order) of one or more reactions. Considering the above-mentioned reasons and also outstanding advancements in both wet-lab technologies and computational platforms (both hardware and software), kinetic modeling of cellular metabolism asks for new modeling strategies and their corresponding experimental design. Regardless of the above-mentioned limitations, MM-based kinetic modeling approach has been commonly practiced for the modeling of different metabolic pathways in different organisms. Several MM-based models that have been successfully applied for metabolic engineering purposes are addressed in the next section.

2.3.1. Applications of Kinetic Models

Genetic manipulation of unicellular organisms such as yeast and bacteria to increase the production yield of specific metabolites is a common practice in metabolic engineering. The modified strains are called “cell factories” and they play a crucial role in the reduction of the total production cost in biotechnology industries. Design and creation of efficient cell factories have made it possible for biotechnology-based productions to compete with others (mainly chemical industry products) in the market. Metabolic models that are able to predict the desired phenotypes are crucial for rational design of cell factories. There has been successful applications of kinetic models for the production of organic compounds such as glycerol [127], ethanol [128], succinate [129] and 2,3-butanodiol [130] in food industry, as well as aromatic amino acids [131, 132]. The earlier kinetic models that have been used in metabolic engineering are based on mechanistic expression of some of central carbon metabolism reactions by using MM-based rate laws and are usually analyzed by Metabolic Control Analysis (MCA) framework. In an attempt to increase the production of acetoin and diacetyl in *L. lactis*, MCA was successfully applied on a small MM-based model (built using *in-vitro* kinetics) to identify suitable strain designs [109]. A detailed glycolytic model of *L. lactis* was constructed later [133] and MCA was used to identify the metabolites affecting fluxes of L-Lactate dehydrogenase, pyruvate dehydrogenase and hexose transporter. The same modeling approach coupled with MCA has also been used for the identification of strain designs that decrease the glycerol production in *S. cerevisiae* [127]. Kinetic models have also been used to increase the uptake rate of specific substrates. Nishio *et al*. [134] used a kinetic model to improve glucose uptake rate of *E. coli*. An unexpected combination of *ptsI* gene overexpression and *mlc* knockout was suggested by the model. However, when tried experimentally, it could greatly increase the specific rate of glucose uptake. Kinetic models were also used to increase the uptake rate of xylose in *S. cerevisiae* [108] and in *L. lactis* [107].

2.3.2. Towards Large-Scale Kinetic Modeling of Metabolism

Genome-scale stoichiometric models of metabolism are available for hundreds of organisms. These models do not contain any kinetic information and cannot simulate the dynamic behavior of metabolism in response to different signals, neither they provide any means for stability analysis or to understand the dynamic

characteristics of the system in general. Although they have been widely used in the prediction of active metabolic networks and/or corresponding metabolic flux distributions at steady-state conditions, their prediction power is limited because of not taking the enzyme kinetics into account. Kinetic models are better representatives of metabolic activity, but there are several big obstacles and challenges in constructing reliable kinetic models for significantly large metabolic networks. To alleviate the gap between stoichiometric constraint-based models and kinetic modeling, integration of kinetic expressions and/data has been proposed as one possible solution. To this aim, a few methods were presented by different researchers in recent years. Machado *et al.* [135] used randomly generated parameter samples to create a set of steady-state solutions for the central carbon metabolism of *E. coli*. The steady-state solution of the kinetic model could be mapped into the flux bounds of the constraint-based model, restricting the solution space. In another research, a simplified kinetic model was constructed by Cotten and Reed [136]. They integrated fluxomic, proteomic and metabolomics data to estimate kinetic parameters. Afterward, they incorporated the flux distributions as additional constraints into a constraint-based model to improve predictions over FBA.

In recent years, there has been an increase in the number of attempts to create alternative kinetic modeling approaches for large-scale networks. Some have tried to make kinetic models of relatively large networks by using a generalized kinetic expression for all the reactions in the model. Such kinetic expressions are not mechanistic, and they are also known as “approximative kinetics”. Different approximative kinetic formats such as logarithmic-linear, power law S-systems, Generalized Mass Action (GMA) and linear-logarithmic were evaluated in literature [137], and it was concluded that linear-logarithmic format combines all desired properties and seems the most appropriate kinetic format. Approximative kinetic formats require less kinetic information than mechanistic rate laws, and they facilitate the development and analysis of large-scale models but ignoring many real kinetic behaviors. One must pay attention that the prediction power of approximative kinetic models is limited to the conditions close to the corresponding steady state around which the model is constructed. Mass Action Stoichiometric Simulation (MASS) is another method that has been developed to incorporate kinetic information and experimental data into stoichiometric reconstructions [138, 139]. One of the major limitations of MASS is that it does not consider the uncertainties in the calculation of

kinetic parameters based on high-throughput omics data. Such a limitation however has been alleviated in the Ensemble Modeling approach [77], in which the parameter space is sampled while still constrained by the thermodynamics of the reactions and a reference steady-state flux distribution. Among all other approaches, the ensemble modeling approach has a very high potential for kinetic modeling of large-scale metabolic networks.

In Chapter 4 a computational tool, “Kinescope”, is introduced, which was developed within the scope of this thesis study for the semi-automatic construction of kinetic models of metabolic pathways. Kinescope is mainly constructed based on the idea of ensemble modeling, however the algorithm presented in the original article [77] was revised and several modifications were made to the algorithm. For example, one of the main flaws in the original algorithm is that it does not count for the changes in the transcription pattern between the two strains (e.g. wild type and gene deletion mutant) and so the relative changes in the total enzyme concentrations of different reactions are ignored, which can lead to large deviations in reaction rates. Also, the relative ratios of different forms of an enzyme (e.g. free enzyme, enzyme-substrate complex, enzyme-inhibitor complex, etc.) can change between two different steady-state conditions, however it is not considered in the original algorithm. Therefore, simulating the models with initial conditions based on the same ratios as the reference steady-state can lead to significant errors. In the revised algorithm presented in this thesis study, both of the above issues were considered. Moreover, an experimental setup is suggested that is appropriate for this modeling approach and can facilitate filtering of the ensemble until a handful of reliable models are found.

3. JACLY: A JACOBIAN BASED METHOD FOR INFERENCE OF METABOLIC INTERACTIONS FROM THE COVARIANCE OF STEADY STATE METABOLOME DATA

Reverse engineering metabolome data to infer metabolic interactions is a challenging research topic. Here, a Jacobian-based method (JacLy) to infer metabolic interactions of small networks (< 20 metabolites) from the covariance of steady-state metabolome data is introduced. The approach was applied to two different *in silico* small-scale metabolome datasets. The power of JacLy lies on the use of steady-state metabolome data to predict the Jacobian matrix of the system, which is a source of information on structure and dynamic characteristics of the system. Besides its advantage of inferring directed interactions, its superiority over correlation-based network inference methods was especially clear in terms of the required number of data replicates and the effect of the use of *a priori* knowledge in the inference. Additionally, Jacly uses the standard deviation of the replicate data as a suitable approximation for the magnitudes of metabolite fluctuations inherent in the system.

Inference of cellular interactions by processing biomolecular data is a widely used approach to investigate functional properties of cellular systems. Perturbations due to genetic/environmental alterations and diseases lead to changes in functionality due to change in cellular network structure, and network inference using the biomolecular data of the perturbation states uncovers the changes in network structure. When applied to the data of metabolite levels, the approach infers metabolic interaction [11, 12, 140, 141]. The general trend is to use dynamic data to infer directed metabolic networks, and steady-state data to infer undirected networks. On the other hand, there are approaches that infer directed metabolic interactions from steady-state metabolome data by utilizing inherent intrinsic variability in such data [15, 99].

While the principles of conservation of mass and conservation of energy set the boundaries for deterministic behavior of the metabolic network, the inherent randomness in this network, as it exists in other biological networks, leads to small variability in the steady-states of the system at equivalent macroscopic conditions [142, 143]. From a microscopic point of view, the inherent randomness is believed to be the result of existence of discrete particles in the system, and molecular fluctuations are inherent in the mechanism by which the system evolves [106]. Continuous change

in micro-environment as well as multilevel complex regulatory mechanisms in the metabolic network are also the causes of observed variability in the steady-states of the system [99]. Although this intrinsic randomness introduces a great obstacle and difficulty in modeling and simulation of metabolic networks, at the same time it provides an opportunity to infer and estimate the active metabolic network at a specific condition/context just by reverse engineering the corresponding replicates of metabolome data at steady state. Considering the fact that information on interactions between metabolites and hence the structure of the active metabolic network is implicit in these data, the main questions are (i) how much information on the structure of the network is hidden in the data, and (ii) how one may extract as much as possible of that information from the data.

One common approach that utilizes inherent variability in steady-state data is correlation-based inference methods, especially the Gaussian Graphical Model (GGM) approach. Correlation based approaches are capable of detecting strong interactions in the metabolic network to some extent. However, they infer interactions only in undirected manner, and they have limited power in the detection of weak interactions [12]. A directed network inference approach from steady-state metabolome data is also available in the literature [15, 99, 141]. The approach is based on the prediction of interaction strengths from the covariance of the data. The network structure information stored in the inherent variability of the data is reflected on the covariance of the data, and later used in the prediction of interaction strengths in the form of a Jacobian matrix. The Jacobian matrix of a cellular interaction system contains a significantly high amount of valuable information both on the structure and dynamic characteristics of the system. Numbers in this matrix easily provide us with detailed information on the underlying interactions in the network, such as direction of interactions, nature of interactions (positive or negative effects), and strengths of interactions [15, 99]. The Lyapunov equation provides a link between the Jacobian matrix of a cellular system and the covariance matrix of the replicates of steady-state data. This equation is the result of a Langevin type approach for description of stochastic processes at macroscopic level [144]. The Lyapunov equation was also used previously to infer differential changes in a Jacobian matrix rather than the inference of the network structure itself [145, 146]. In another work [147], a comparison of several least square and regularization methods in solving the Lyapunov equation for the Jacobian matrix is provided. However, in that work, the structure of the Jacobian

(zero and non-zero elements) is specified a priori by the stoichiometric matrix of the metabolic network. Therefore, the problem is reduced to the estimation of magnitudes for non-zero elements in the Jacobian matrix, which might not be considered as a network inference.

3.1. Methods

3.1.1. Problem Definition

Provided that replicates of metabolome data are available for an organism in a specific condition, and considering the fact that information on interactions between metabolites and hence the structure of the metabolic network is implicit in these data, the problem is to extract from the data as much knowledge as possible to infer the active metabolic network in that condition. A metabolic reaction network can be mathematically represented by writing mole-balance equations around its metabolites. This leads to a system of differential equations that can be summarized as in the following equation, where C is a vector of metabolite concentrations:

$$\frac{dC}{dt} = f(C) \tag{3.1}$$

For a system around steady-state, a linear approximation can be made to express the equations in terms of a Jacobian matrix, J [99, 138]:

$$\frac{dX}{dt} \approx JX \tag{3.2}$$

with $X=C-C_s$, and C shows concentrations fluctuating around steady-state values, Cs. Equation 3.2 can further be expressed as a Langevin-type equation to explicitly account for small fluctuations [99]:

$$\frac{dX_i}{dt} = \sum_j J_{ij}X_j + \sqrt{2D_i\delta_i(t)} \tag{3.3}$$

$$J_{ij} = \frac{\partial(\frac{dC_i}{dt})}{\partial C_j} \tag{3.4}$$

Where D_i shows the extent of fluctuation and δ_i is a random number from unit normal distribution. As demonstrated in the literature [106], Equation 3.3 can be written as follows at steady-state:

$$J\Gamma + \Gamma J^T = -2D \tag{3.5}$$

Equation 3.5 is known as the Lyapunov equation, and it provides a link between the Jacobian matrix of the network (J) and the covariance matrix of the replicate metabolome data (Γ) [15, 99].

The fluctuation matrix (D) accounts for the inherent randomness in the system. The diagonal elements of D reflect the magnitude of fluctuations observed on each metabolite, and the nondiagonal elements can be assumed as zero [99]. The equation is determined in terms of calculating the covariance matrix (Γ) while the Jacobian matrix (J) is provided, however, it is underdetermined in the case of calculating J while Γ is available. This is because there are n(n+1)/2 independent entries in Γ for an n-metabolite system due to the symmetric nature of the covariance matrix, whereas the Jacobian matrix has n^2 independent entries. This equation can be rearranged to a standard linear system of equations [15] and be represented as following:

$$Aj + 2d = 0 \tag{3.6}$$

In this equation, A is a square matrix of size $n^2 \times n^2$ derived from the covariance matrix, j is the vectorized form of the Jacobian matrix with size $n^2 \times 1$, and d is the vectorized form of fluctuation matrix with size $n^2 \times 1$, where n is the number of metabolites. Öksüz *et al.* used Equation 3.6 to solve for Jacobian matrix in an optimization platform using Genetic Algorithm (GA) [15]. Beside the minimization of the residual of Lyapunov equation, they used sparsity as a rational objective function to select Jacobians from the solution space that have a high number of zeros and satisfy Equation 3.6 as well. The multi-objective function that simultaneously maximizes the

number of zeros (sparsity) in the Jacobian matrix to be determined and minimizes the residual of Equation 3.6 can be represented as following:

$$f = (number\ of\ zeros) \times \lambda - \log_{10}(\|Aj + 2d\|) \tag{3.7}$$

The first term in the equation counts for the number of zeros in the Jacobian matrix that needs to be maximized, the second term counts for the residual of Lyapunov equation that needs to be minimized, and in total the objective function f is to be maximized. Lambda (λ) is a scaling factor discussed in detail in a section below. In order to balance between the two goals in the objective function, and also to refrain the solution from going to a very high number of zeros, the scaling factor was introduced in the objective function.

Using the exact covariance and predefined fluctuation vector as inputs to the algorithm, Öksüz *et al.* [15] validated the theoretical applicability of this approach. The same objective function was used in this study, however, after careful examination of the problem, an extensively modified algorithm was developed. The new algorithm is highly robust and could be applied to the replicates of *in-silico* metabolome data generated by simulating stochastic dynamic models of metabolism using stochastic differential equation (SDE) solvers.

Simulations and optimizations were performed in MATLAB (R2017a) on a desktop computer equipped with a 3.2GHz CPU and 4GB RAM. SDE simulations were performed using the SDE toolbox that is freely available as an external MATLAB toolbox [148]. Genetic Algorithm (GA) was implemented using the ga function in MATLAB's global optimization toolbox. A built-in parallelized version of ga was used with the help of MATLAB's parallel computing toolbox. Custom MATLAB functions were written for creation, crossover and mutation fields of GA. Maximum number of generations was set to 800 and a mutation rate of 5% was selected after careful examination of GA's behavior. The MATLAB codes for JacLy are available in appendices.

3.1.2. Optimization Pipeline

Genetic Algorithm (GA) was used to solve Equation 3.6 for j while A and d are settled. At each generation of GA, bit string vectors are generated for j as individuals.

With a candidate bit string for the structure of the Jacobian vector (zero and nonzero elements in j), Equation 3.6 can be reduced to a lower dimensional system of equations by removing the zeros in the j and also removing the corresponding columns in A.

$$A_r j_r + 2d = 0 \tag{3.8}$$

Since the Jacobian vector is sparse in structure (r<<n), this leads to considerable reduction in the number of unknowns to be determined, and increases the speed of the inference algorithm compared to the original algorithm in [15]. In Equation 3.8, j_r is the reduced form of Jacobian vector, obtained by removing those elements corresponding to zeros in the suggested individual, and A_r is formed by removing the corresponding columns in A. Equation 3.8 can be easily treated and solved as a line fitting problem, in which the elements of j_r are factors of the line equation and are estimated to make the best fit to the data. Minimizing the Euclidean norm of this fitting is one of the terms in the optimization objective function defined in Equation 3.7. The other objective is to maximize the number of zeros in the Jacobian matrix, considering the fact that metabolic networks are sparse networks (discussed in a section below).

3.1.3. Constraining the Solution Space by Generating Sparse Individuals

As the number of metabolites and hence size of the network increases, the solution space expands exponentially and the probability of finding the true candidate for Jacobian vector through a stochastic algorithm decreases significantly. Moreover, the computational time and effort increases dramatically with the size of the network [14]. In these situations, it is very important to constrain the solution space as much as possible to solve the problem (Equation 3.8) in a manageable time. One way to constrain the solution space meaningfully is to control the generation and reproduction of candidate individuals in GA such that those individuals with unwanted characteristics are not produced at all to be tested. Since metabolic networks are naturally sparse networks, setting a minimum sparsity parameter for the generated individuals can be used as a controlling parameter. This was another novelty in the algorithm compared to the original one [15].

It is known that metabolic networks are highly sparse, meaning that there are much less interactions (edges) in the network compared to the maximum possible

number of edges (fully connected network). The natural sparsity in several known metabolic networks was calculated, and it was observed that all the tested networks have a sparsity larger than 0.55. As a result, a minimum sparsity of 50% was selected as the default value in the algorithm. Just by definition of such a parameter, the solution space to search for non-zero values of Jacobian is greatly reduced. Sparsity parameter was defined as the following:

$$\text{Sparsity} \equiv \frac{\text{Total number of possible edges in the network} - \text{number of edges in the network}}{\text{Total number of possible edges in the network}} \quad (3.9)$$

$$\cong \frac{\text{Number of zeros in the Jacobian}}{\text{Total number of elements in the Jacobian}}$$

Based on this definition, a sparsity value of one will mean a network with not even a single edge whereas a value of zero will correspond to a complete digraph. It must be considered that the sparsity calculated from the Jacobian and the one calculated from the biochemical reaction network are not necessarily the same, since the Jacobian also counts for the regulatory interactions which are absent in the biochemical reaction network, but since the number of regulatory interactions is usually insignificant compared to the number of reactions, the two values are very close.

In order to minimize the computational effort and time, the minimum sparsity parameter was used as the control parameter in the creation of the initial population in GA, and then in the production of individuals at subsequent generations. To this aim, custom MATLAB functions were written and used for creation, crossover and mutation fields of GA in MATLAB. This was another novelty over the previous algorithm [15]. With this supervised control of individuals, bit-strings with unwanted characteristics had no chance to appear as the candidates for Jacobian vector, and it provided a significant contribution in constraining the solution space. The custom functions for GA were written in such a way that minimum sparsity is intrinsic in the generation of individuals and no time is consumed for control and filtering of the generated bit-strings.

3.1.4. Scanning for The Scaling Factor

The objective function (Equation 3.7) consists of two terms, one is the residual of Equation 3.8 to be minimized and the other is the number of zeros in the Jacobian vector to be maximized. In order to balance between these two values and also to prevent the optimization algorithm from converging to the too sparse solutions, a scaling factor (λ) is multiplied with the term for the number of zeros in the Jacobian vector. Since this parameter directly affects the magnitude of the objective function, it is important to find a range of lambda that leads to sensible solutions. Selecting very small values for the scaling factor leads to the conditions in which the optimization will not be very sensitive to the number of zeros in the Jacobian vector, and minimization of the residual of Equation 3.8 would be dominant in the objective function. On the other hand, large magnitudes of the scaling factor lead to the solutions with very high number of zeros in the Jacobian vector, with almost no sensitivity to the residual of Equation 3.8. There is always a narrow interval for the scaling factor, in which the optimization problem can find a Jacobian vector with optimum number of zeros that also leads to insignificant residual value for Equation 3.8. This interval for the scaling factor varies from problem to problem [15], depending on several factors among which are the size of the network and the accuracy and number of data replicates from which covariance matrix is calculated. Constant problem-specific λ values were used in the previous algorithm [15]. In order to circumvent the obstacles due to selection of a proper value for the scaling factor, it was decided to scan a range of values for the scaling factor in an unsupervised manner, instead of estimating a constant value for each specific problem. A range between 0.01-0.10 with increments of 0.005 was scanned. Since the algorithm is repeated 10 times due to the stochastic nature of genetic algorithm, it leads to a total of about 200 solutions per network inference problem. In this way, the optimization algorithm works repeatedly for each value of the scaling factor, and the optimum solution would have chance to appear among the candidate solutions. This was another improvement over the previous algorithm.

3.1.5. Fluctuation Vector

One of the major obstacles in utilizing the Lyapunov equation is introduced by the fluctuation matrix D since it may contain non-observable quantities [146]. The

fluctuation matrix plays a critical role in this equation, and the computed Jacobian matrices are highly sensitive to the values in the fluctuation matrix. After a reasonable fluctuation matrix is selected, the problem of solving the underdetermined equation to find the Jacobian can be formulated as an optimization problem.

The existence of a non-zero fluctuation vector is both physically and mathematically meaningful. Fluctuation vector represents the intrinsic noise in the molecular interactions that are the source of stochasticity in the replicates of data through which the information on the structure of the network is going to be extracted. A non-zero fluctuation vector also prevents the equation (6) to have null space, that otherwise would be problematic. On the other hand, it is not very clear how to find and how to set the values for the fluctuation vector in equation (6). In a previous work [15], a constant problem non-specific small value of 0.005 was used for all metabolites to mimic small fluctuations around metabolites. Here, it is hypothesized that standard deviation of the data replicates would be an acceptable resource to be used for estimation of the diagonal elements of the vectorized fluctuation matrix. The use of data-specific fluctuation vector elements in this manner rather than using a constant value for all problems is another improvement in this algorithm compared to the original algorithm [15].

3.1.6. Using a Community of Estimated Jacobians Instead of Only One Elite Jacobian to Infer Structure of The Network

JacLy scans a range of scaling factors (λ) in the objective function (Equation 3.7). For each scaling factor, optimization is performed 10 times, leading to hundreds of optimizations. The end result of each optimization is a Jacobian vector that has the maximum objective function value among thousands of other individuals. This Jacobian is called the best-found Jacobian for each optimization. Among all the best-found Jacobians, one can be selected as the elite Jacobian vector based on both sparsity and residual of Equation 3.8. In all the test studies, it was observed that if, instead of the elite Jacobian, a community of the best-found Jacobians in a bounded area around the elite Jacobian are combined and used to infer the structure of the final Jacobian, the accuracy of inference significantly increases. To do so, first a bounded area is defined around the elite Jacobian based on the number of zeros and the residual of Equation 3.8. For all of the tests in this study, -/+ 5% of the number of zeros in the elite Jacobian was selected to set the lower and upper vertical boundaries and -5% of

the residual of Equation 3.8 for the elite Jacobian was selected to set the lower horizontal boundary (see Figure 3.1). The Jacobian vectors in the bounded area are then binarized by setting their non-zero elements to one. The binarized vectors are then aligned on top of each other to form a binary matrix. Taking average over columns of this matrix leads to a new vector including fractional numbers between 0 and 1. The structure of the final Jacobian vector is then determined by setting a threshold of 0.5 to decide for zero and non-zero values. Those elements that are smaller than the threshold are set to zero and others to one. A looser threshold of 0.4 increases both TPR and FPR. It was also observed that selecting 0.4 as the threshold leads to better g-scores in general, however, there is the risk of sparsity being dropped to significantly lower values. As a result, while selecting a threshold for the combination of best-found Jacobians, the use of sparsity check as a caution is advised. Figure 3.2 provides a schematic of this procedure.

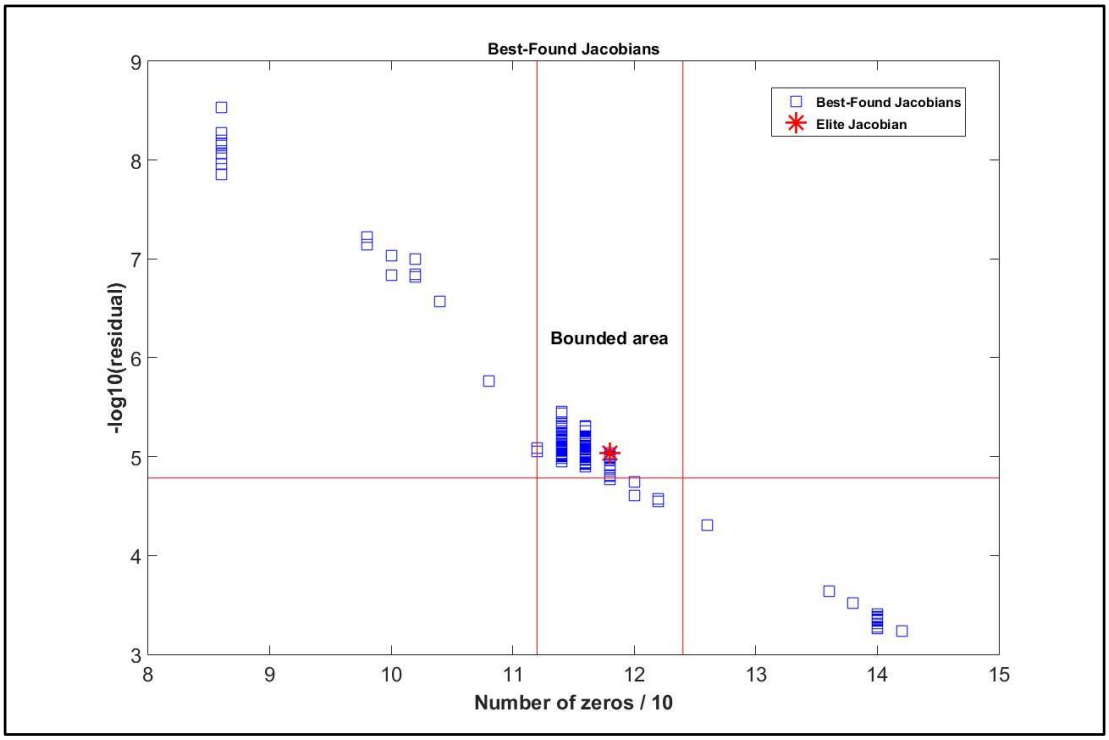


Figure 3.1: Selecting a bounded area around the elite Jacobian based on a percentage of the number of zeros in the elite Jacobian and the residual of Equation 3.8.

Jacobian vectors in the bounded area	Jacobian vector 1	1	1	0	1	0	0	...	1	0	1	0
	Jacobian vector 2	0	1	0	1	1	0	...	0	0	1	0
	Jacobian vector 3	1	1	1	0	0	0	...	0	0	0	1
	Jacobian vector 4	1	1	0	0	1	1	...	1	0	0	0
	Jacobian vector 5	0	1	0	0	1	0	...	1	0	1	1
	Jacobian vector 6 (elite)	1	1	1	0	0	0	...	1	0	0	1
	Jacobian vector 7	0	1	0	0	1	1	...	0	0	0	1
	Jacobian vector 8	0	1	1	1	0	0	...	1	0	1	1
	Jacobian vector 9	1	1	0	0	0	1	...	1	0	1	0
	Jacobian vector 10	1	1	0	0	1	0	...	0	0	0	1
		0.6	1	0.3	0.3	0.5	0.3	...	0.6	0	0.5	0.6
Structure of the final Jacobian		1	1	0	0	1	0	...	1	0	1	1

Figure 3.2: A schematic example of alignment and combination of Jacobian vectors in the selected community to come up with the final structure.

In this study, a mathematical proof is not provided to prove how using a community of best-found Jacobians around the elite Jacobian can improve the inference results. However, as far as tested with different *in-silico* datasets and noisy covariances, using such combinatorial approach not only leads to better inference results, but also it stabilizes the final output of the algorithm when applied to the same problem repeatedly. The use of a community of best-found Jacobians is another novelty over our previous work [15], which reported the results based on only the elite Jacobian.

3.1.7. Quantification of Inference Performance

While True Positive Rate (TPR) and False Positive Rate (FPR) are two quantities suitable for the evaluation and comparison of network inference results, g-score can be used as a single parameter to quantify performance of any inference method. g-score is calculated by the following equation:

$$g - score = \sqrt{(TPR \times (1 - FPR))} \tag{3.10}$$

3.2. Results

3.2.1. Use of *in silico* Covariance Matrices for Metabolic Models of *S. cerevisiae* and *E. coli*

The Lyapunov equation (Equation 3.5, and Equation 3.6 in the rearranged form) is underdetermined in terms of calculating the Jacobian matrix J given Γ and D as inputs, meaning that there is more than one Jacobian matrix satisfying the Lyapunov equation for each pair of Γ and D (see Methods section). To evaluate the applicability and performance of our method (JacLy) in predicting the network structure through the prediction of the Jacobian matrix, two kinetic models that are well known in the literature were used. The first model covers 13 metabolites of yeast glycolysis [68], and the second model covers 18 metabolites of central carbon metabolism in *E. coli* [69]. True Jacobian matrix was calculated for each kinetic model around its corresponding steady-state by using the detailed rate expressions and parameters in the models. Here, the same strategy as in the previous work [15] was followed, however, the highly improved genetic-algorithm-based dual objective formulation was tested by the *in silico* generated metabolome data, instead of using the exact covariance matrices. The purpose of this section is to demonstrate the improvements in the current version of the algorithm compared to the previously published one [15]. Having the true Jacobian matrix and predefined fluctuation matrices, the exact covariance matrix was calculated from equation (5). These covariance matrices are called "exact" covariances since they hold true to the Lyapunov equation. Exact covariances and corresponding fluctuation matrices were used as inputs to JacLy to evaluate its performance in finding J.

JacLy uses genetic algorithm for optimization, which is a stochastic optimization algorithm. Therefore, it is important to solve for the equation for enough number of times until a constant reproducibility parameter is achieved. For both models, a constant reproducibility is obtained after 20 runs. Out of 20 repetitive runs for each model, 19 and 18 of them could find Jacobian matrices that are in complete agreement (100% TPR and 0% FPR) with the true networks of yeast and *E. coli* models, respectively. These results show a great improvement over the previous work [15], which had a reproducibility parameter of 50% and 5% for yeast and *E. coli* models, respectively. On our desktop computer, each run takes around two minutes for the

yeast model and six minutes for *E. coli*, showing a 10 fold increase in computational speed over the previous work [15]. Such significant improvements in reproducibility and computational speed have been achieved solely by modifying the algorithm and corresponding functions (see Methods section). One should note that since JacLy incorporates a λ scan with 10 replicate solutions, the whole process of generating 200 solutions for *S. cerevisiae* took one hour while the time in the case of *E. coli* was two hours.

The performance of JacLy was also evaluated by using noisy covariances as input. To this aim, the same procedure as in the previous work [15] was used to add noise to the exact covariance of the yeast model. Random numbers were sampled from a normal distribution with a mean of 1 and a standard deviation of 0.005. This corresponds to a dataset with 50% noise [149]. The random numbers were then symmetrically multiplied with the elements of the exact covariance to generate a noisy covariance matrix. This was repeated to generate ten different noisy covariances and JacLy was applied on each. The average TPR and FPR are 74% and 5%, respectively. These numbers were 73% TPR and 11% FPR in the previous work [15]. These results show a considerable increase in the performance of JacLy compared to its ancestor in terms of the FPR value since exactly the same problem was solved with only improvements in the algorithm based on (i) the use of reduced form of the Lyapunov equation, (ii) the use of sparsity constraint, (iii) scanning the scaling factor and (iv) the use of a community of candidate Jacobian vectors, as discussed in detail in the Methods section. Additionally, note that a threshold of 0.4 in the combination of Jacobian vectors rather than 0.5 leads to a TPR of 84% and an FPR of 8%.

It was reported in the literature that statistical methods such as LASSO and Tikhonov regularization fail to solve Equation 3.6 whenever the condition number of matrix A is significantly large [147]. In order to evaluate the sensitivity of the method to the condition number of A, different fluctuation matrices along with the true Jacobians of yeast and *E. coli* models were used as inputs to Equation 3.5, and different exact covariances were calculated leading to different A matrices covering a range of condition numbers from 106 to 1025. JacLy was applied to each of those covariance matrices along with their corresponding fluctuation matrices. It was observed that the condition number of A doesn't have any influence on the performance of JacLy. Even for the largest condition numbers, JacLy was able to find the true Jacobian with similar computational time and reproducibility parameters. It should be kept in mind that not

being sensitive to the condition number of A in solving Equation 3.6 doesn't mean that the calculated Jacobian matrix is not sensitive to the changes in the fluctuation matrix. Indeed, Equation 3.6 is frequently ill-conditioned as it is also reported in other studies [147]. Small changes in the fluctuation matrix D lead to big changes in the calculated Jacobian matrix.

3.2.2. Use of *in silico* Metabolome Data for Metabolic Models of *S. cerevisiae* and *E. coli*

At this stage, JacLy was applied to the replicates of *in silico* metabolome data. Stochastic versions of yeast [68] and *E. coli* [69] models were used to generate 1000 replicates of steady-state metabolome data *in silico*. In this case, it was necessary to come up with a fluctuation matrix to be used as input to the method along with the covariance of metabolome data. As it was mentioned in the Methods section, it was hypothesized that standard deviation of data might be a reasonable source to be used for the construction of a fluctuation matrix. In a stochastic dynamic system, all or some of the sources of stochasticity are usually unknown. In the Lyapunov equation the fluctuation matrix D is the parameter counting for sources of stochasticity. Since standard deviation is a measure of variation in data, it was used as a reasonable source to construct the fluctuation matrix. Table 3.1 shows the inference results of JacLy applied to *in-silico* data for the yeast and *E. coli* with a comparison to GGM-based inference results. In GGM analysis a cut-off of 0.01 was used for p-values to decide on the significance of partial Pearson correlation values. The networks predicted by JacLy are directed while those estimated by GGM are undirected.

It must be considered that solving Equation 3.6 for the Jacobian vector is highly sensitive to the fluctuation vector d, and so it is of critical importance to come up with a fluctuation vector that is most reasonable for data replicates. Normalization of the data was thought as a way that might improve the correspondence between the covariance matrix Γ and the fluctuation matrix D in the Lyapunov equation. Data normalization doesn't have any effect on the results of similarity-based inference methods such as GGM. *In-silico* metabolome data for the yeast and *E. coli* were normalized to between 0 and 1 by dividing each value to the maximum value in the dataset. Normalized data was then used to make both covariance matrix Γ and fluctuation matrix D. When applied to the normalized data, JacLy showed a significant

improvement in inference results for the yeast data (0.95 TPR and 0.13 FPR, with a g-score of 0.91) while it had no effect on the inference results of *E. coli* data.

Table 3.1: Inference results for the *in silico* metabolome data, comparison of JacLy and GGM.

	In-silico data for Yeast			*In-silico* data for *E. coli*		
	TPR	FPR	g-score	TPR	FPR	g-score
JacLy	0.66	0.08	0.78	0.69	0.29	0.70
GGM	0.76	0.12	0.82	0.63	0.12	0.74

Another parameter influential on the result of network inference is the number of replicates in the data. Previous GGM-based analysis for the inference of metabolic interactions using *in silico* metabolome data for the same networks analyzed here showed a sharp decrease in the quality of the inference after the number of replicates decreased below 200 [12]. Here, the effect of number of datapoints on the inference results of JacLy was tested. Of 1000 replicates initially generated by stochastic differential equations, 100 randomly chosen replicates were used in the inference of the network for *S. cerevisiae*. Repeating this 10 times and taking the average, a TPR of 0.73 and an FPR of 0.16 was obtained by using JacLy, corresponding to a g-score of 0.78. On the other hand, GGM-based inference for the same randomly chosen 100 replicates resulted in average TPR and FPR values of 0.42 and 0.03, respectively, with a g-score of 0.63. Therefore, an advantage of JacLy over GGM is its considerable robustness in terms of the number of replicate datapoints used in the covariance/correlation calculation.

In the process of inferring a network for a set of metabolites, there are cases when existence (true positive) or non-existence (true negative) of an edge between two metabolites might be available as prior knowledge. Such information can be used as additional input to inference algorithms, resulting in a shrinkage of the solution space and so enhancing the computational speed and performance of the algorithm. The effect of using prior knowledge about non-existent edges on the performance of JacLy was tested. To this aim, 7 and 20 zeros of the true Jacobian matrices were selected as priorly known true negatives for yeast and *E. coli* models, respectively. This corresponds to 5% and 7% of the total number of elements in Jacobian matrices for yeast and *E. coli* models. This procedure was repeated 10 times for each model and

JacLy was applied to data each time. The average TPR and FPR over 10 repetitions for the yeast model are 0.75 and 0.08, respectively, leading to a g-score of 0.83. For *E. coli*, an average TPR of 0.79 and an average FPR of 0.31 was obtained, leading to a g-score of 0.74. These results show a significant improvement compared to the corresponding values in Table 3.1. Based on the results, JacLy performs considerably better when a very small portion of true negatives is introduced as prior knowledge. Specifying true negatives contributed to a better prediction of true positives. The correlation-based GGM approach, on the other hand, is not suitable for the use of prior knowledge.

In addition to the binary structure of estimated Jacobians, which is the main output in inferring the structure of an active metabolic pathway, the best-found Jacobians in the selected area around the elite Jacobian - before binarization and combination – were also compared with the true mechanistic Jacobians of the kinetic models, calculated by using detailed rate expressions and parameter values in those models. Since JacLy has a stochastic nature, the optimization was repeated three times on each SDE data. Afterward, the Spearman correlation was used to make the comparisons. The medians of correlations are around 0.45 and 0.25 for yeast and *E. coli* models, respectively while the medians of p-values are less than 0.0001 in all cases (See Figure 3.3).

The kinetic-model-based true Jacobian matrices, together with SDE-data-based covariance matrices were used as inputs to the Lyapunov equation to calculate an exact fluctuation matrix. The exact fluctuation matrix was then compared to the approximated one estimated by JacLy. It was observed that the exact D contained off-diagonal non-zero elements as opposed to the estimated one. Some of the elements had the same magnitude as the diagonal elements. On the other hand, our very simple method of estimating D led to quite acceptable TPR and FPR values in those case studies, and the standard deviation of data is logically related to the source of natural fluctuation in the system. Therefore, our estimation approach can be used because of its simplicity and applicability. However, research should be performed to develop a more accurate method of estimating D. On the other hand, one should note that SDE simulator algorithms and toolboxes, such as the one used in this study, have stability problems in terms of the generated noise when applied to highly nonlinear systems. This could also be another reason behind the inconsistency between the covariance of

SDE data and the true Jacobian matrix, which directly affects calculation of the fluctuation matrix from the Lyapunov equation.

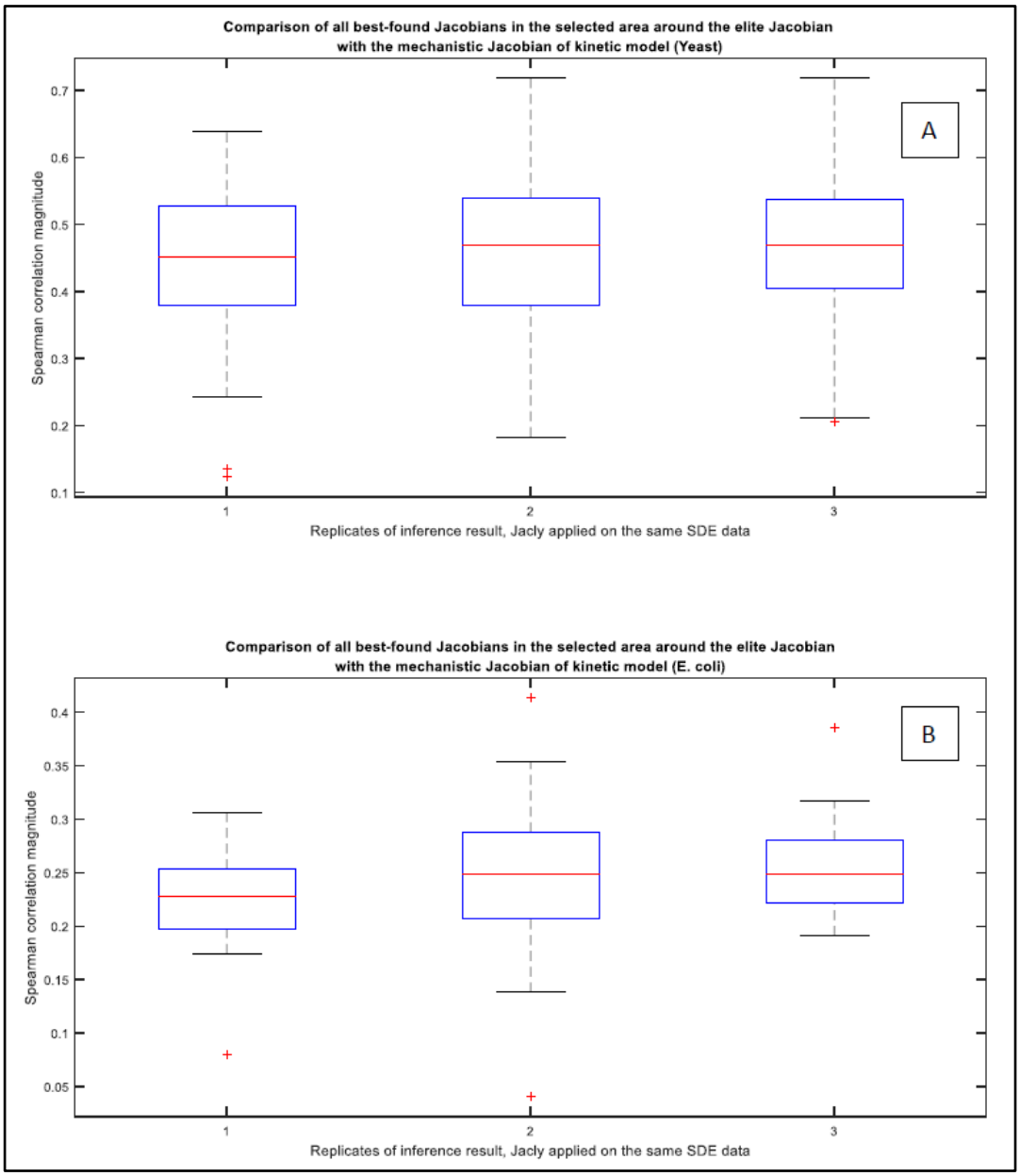


Figure 3.3: The Spearman correlation between the predicted Jacobian matrix values by JacLy and the calculated values from the kinetic models are shown here for both Yeast and E.coli. JacLy was applied three times (due to its stochastic nature) for each organism, and the correlation was calculated for each of the best-Jacobians determined around the elite Jacobians. The results are given below in the form of boxplots. (a) The results for the yeast. 1: 60 best-found Jacobians, 2: 100 best-found Jacobians, 3: 96 best-found Jacobians. (b) The results for E. coli. 1: 12 best-found Jacobians, 2: 100 best-found Jacobians, 3: 11 best-found Jacobians.

3.3. Discussion

JacLy is a network inference algorithm with specific focus on the inference of small-scale metabolic networks from steady-state data. It has significant advantages over its ancestor [15]. Here, algorithmic improvements which led to significant improvements in the runtime and prediction power of JacLy are reported. Major improvements were (i) Vectorizing all possible computations, significantly improving the runtime, (ii) the use of the reduced form of the Lyapunov equation by removing the columns corresponding to zero Jacobian vector entries, improving the runtime, (iii) the use of sparsity constraint in genetic algorithm to improve the runtime by eliminating the possibility of generating low-sparsity individuals, (iv) scanning the scaling factor rather than fixing it for each specific problem, making the algorithm more flexible and independent from the effect of chosen parameter value and improving the prediction power, and (v) the use of a community of candidate Jacobian vectors rather than using the elite Jacobian vector in the inference, improving the prediction power of the results. Inferring metabolic pathways via prediction of Jacobian matrices is also useful in estimating dynamic and mechanistic characteristics of the system under investigation.

One issue that is worth mentioning at this stage is the applicability of the approach in terms of the size of the network to be inferred. For example, each run for the *E. coli* model consumed about twice more computational time compared to that of yeast, while the *E. coli* model has only five more metabolites compared to the yeast model, an almost 40% increase in the number of network nodes. This dramatic increase in computational time with respect to network size – whenever the calculation of Jacobian matrix is involved in a network inference method – was observed and explained in previous studies [14], and it is indeed one of the major drawbacks of using such methods to infer larger networks. From this aspect, JacLy is more suitable as a small-scale (< 20 metabolites) network inference method. There are several network inference methods in the literature with a specific focus on small-scale networks [150, 151]. Since different cellular functions are biologically attached to smaller metabolic pathways or subnetworks, it still makes sense to be able to infer active subnetworks for a specific cellular condition rather than inferring the whole network. Table 3.2 summarizes some characteristics of JacLy through a comparison with GGM as one of the most common methods in inference of biological networks.

Table 3.2: A summarized comparison of JacLy with GGM.

	JacLy	***GGM***
Computational time versus network size	- NP-hard problem - Computational time increases exponentially by increasing the network size	No sensible change in the computational time from very small to very large networks
Accuracy versus number of data replicates	- Accuracy is a moderate function of number of data replicates - For lower number of data replicates, it outperforms correlation-based methods	- Accuracy is a very strong function of number of data replicates - Reduction in the number of data replicates has a very high negative impact on the quality of inferred network
Directionality of inferred network	Directed	Undirected
Meaningfulness of inferred edge's weights	- Mechanistically meaningful - Inferred values for the Jacobian elements are measures of interaction strengths and their sign (positive/negative) points into the nature of interaction	- Correlation values cannot be used as any physical or mechanistical parameter of the system

Currently, steady-state metabolome measurement data that are reported in the literature are limited in terms of the number of replicates. This limitation is not specific to our method; commonly used correlation-based inference methods are also suffering

from low number of data replicates and usually lead to significantly high number of false positives. Also, in case of real metabolome data, experimental measurement errors interfere with natural stochasticity of the system leading to lower quality in predicted networks. Moreover, since our method relies on the fluctuation matrix (D) as one of the inputs, this external noise is more troublesome. To test the applicability of our approach to the real metabolome data, random noises were introduced to the SDE data of the yeast model by following the approach presented by Fuente *et al.* [149]. For each metabolite, the random noise was sampled from a normal distribution with mean zero and a standard deviation equal to 10% of the variance in the steady state concentration of that metabolite. Ten sets of noisy data were generated by using this approach. The *in-silico* data already includes randomness due to natural stochasticity since it was generated using an SDE simulation toolbox. This random noise was still added to the data to count for the other sources of error such as measurement errors. Afterward, Jacly and GGM were both applied to the noisy data sets and the inference results were compared with those obtained from the noise-free SDE data. The average g-score dropped from 0.78 (noise-free data) to 0.71 (noisy datasets) for JacLy and from 0.82 to 0.79 for GGM. These results provide a theoretical base for applicability of our approach to real metabolome data that includes other sources of randomness in addition to the natural stochasticity of the system. Although these results show that GGM is less sensitive to the noisy data (as expected) but one should remember that these tests were performed using a 1000 replicate SDE data, that is a very high and unrealistic number to repeat same measurements to obtain metabolome data. As it was shown in section 3.2, JacLy clearly outperforms GGM for lower number of data replicates.

4. KINESCOPE: A TOOL TO EASE KINETIC MODELING OF METABOLIC PATHWAYS

Dynamism is undoubtedly a major aspect of life. All organisms dynamically interact with their environment. Whether it is a single-celled form of life such as bacteria or a complex multicellular organism such as human, they continuously receive signals from their environment and respond to those signals as needed. From a systems point of view, living things are “dynamical systems”. To study a system, both experimental and theoretical playgrounds are needed. Mathematical models are required for theoretical analysis of the system. Originally, to model something means to describe it in another language or format. Mathematics is the language of logic, formalism, intuition and quantity, hence mathematical modeling of a system makes it easier to analyze that system. All mathematical models come with “abstraction and simplifications” of the real system under study. As a result, there isn’t any model that can exactly and precisely describe a system and its behavior in different conditions. In other words, there is no “perfect” model. This is specially the case for more “complex systems” which include significantly higher number of interacting components, while both the components and the interactions among them may be of different natures and follow different mechanisms. In addition, the aggregate behavior of such a system, although being a function of the behavior of its individual components, may not be derivable from a linear combination of those individual behaviors. Figure 4.1 is a schematic representation of a complex system with a limited number of components. The more the number of simplifying assumptions in a model, the less would be its applicability to describe the real system in different conditions. Indeed, the models are not meant to be perfect. As it was mentioned, a mathematical model is a tool for theoretical analysis of a system. It leads to generation of hypotheses, through which new experiments can be designed, and based on the results of the experiments the model will be modified to be a better representative of the system. The modified model then can be used for the generation of a new hypothesis, and this cycle (Figure 4.2) can continue until the model is good enough to reliably represent the system at conditions of interest. (i) Not being able to experimentally verify a hypothesis, and (ii) not being able to simulate a model in a manageable time are two major barriers in the above-mentioned cycle and both can be circumvented by advancements in experimental and computational technologies.

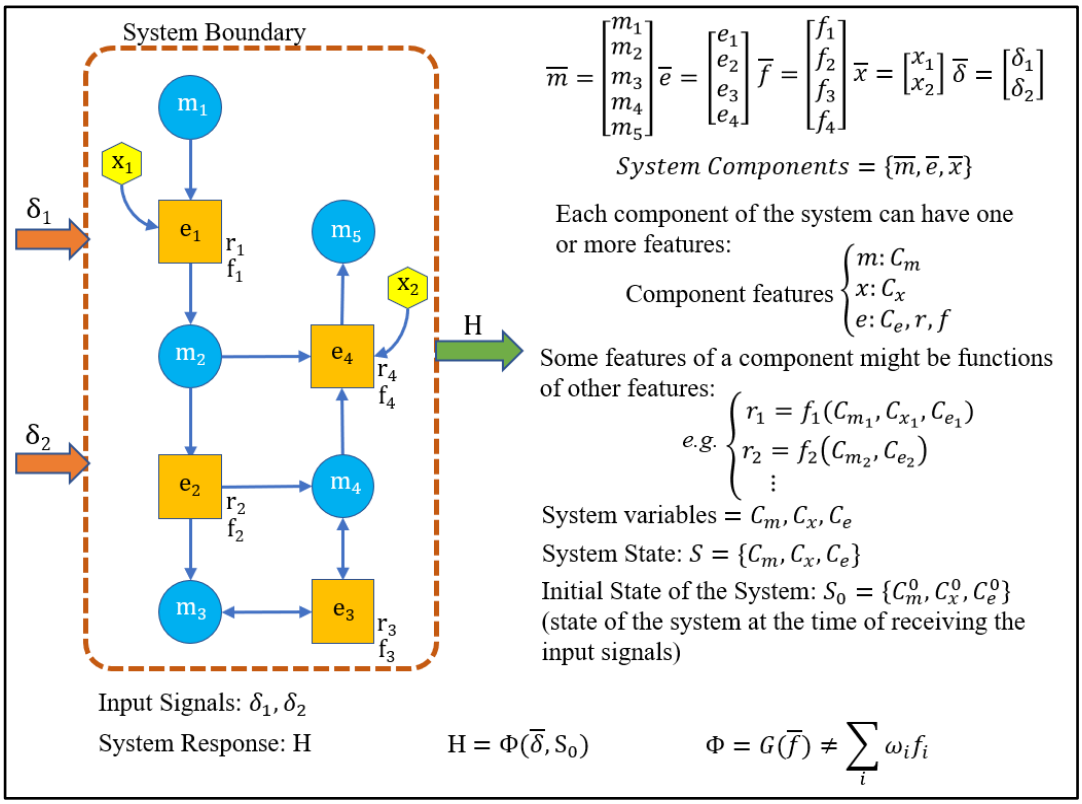


Figure 4.1: An example of a complex system with limited number of components. The response of the system (H) depends on the input signals and the state of the system at the time of receiving those signals. The aggregate behavior of the system (Φ), although can be a function of the behavior of individual components (f1, f2, ...), it may not be derivable by using a linear combination. The figure is intended to be a general representation of a complex system, but as an special example, m, x, e and r can be interpreted to be metabolites, cofactors, enzymes and reaction rates respectively.

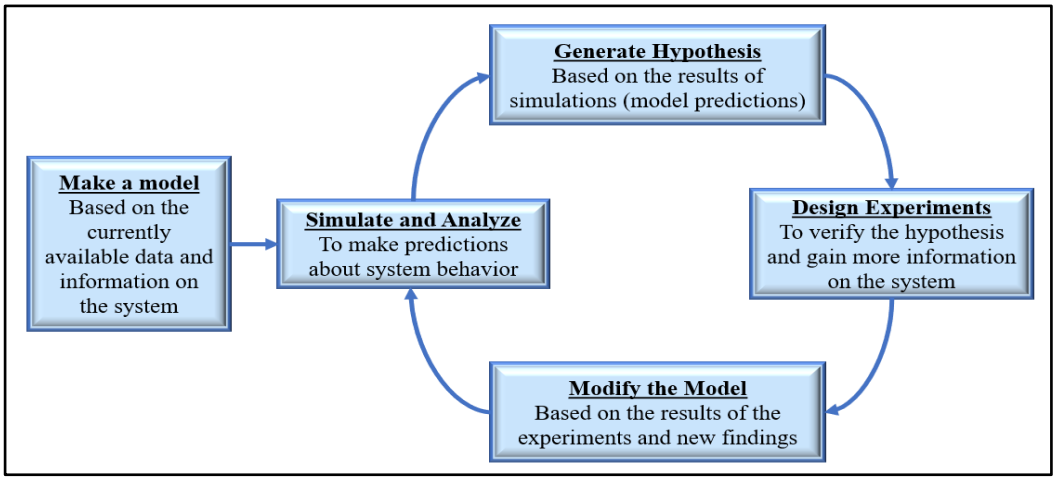


Figure 4.2: Modeling Cycle. Theory and experiment are complementary. Through the modeling cycle that makes a link between theoretical and experimental playgrounds, the knowledge about a system can be increased.

When presenting a model, it is moral to clearly state the conditions in which it can represent the behavior of the system fairly well, and also the conditions in which the model may behave completely different from the real system. Also, whenever using a model, it is critical to pay attention to those conditions. Using a model in conditions where it is not meant to represent the system can lead to generation of false hypotheses and subsequent confusion. Metabolism is a dynamical system. The network of biochemical transformations continuously responds to the concentration of metabolites and enzymes. Sensitivity and specificity of the enzymes to their substrates and regulators is an important factor in this process. However, since transcriptional regulation of the metabolism and the subsequent changes in total enzyme concentrations play a major role in the determination of the metabolic fluxes, a model that only includes kinetics of the enzymatic reactions cannot represent the dynamic behavior of the metabolism shifting from one steady-state to another, even if it is comprehensively covering the kinetics of all reactions. Hence, kinetic models of metabolic pathways are applicable to the situations where the dynamic response of the metabolism is the result of enzyme kinetics only, such as when the perturbation is not strong enough to change the transcription pattern. However, even in cases where changes in the transcription significantly contribute to the determination of the new state, the first few seconds of the dynamic response can be fairly represented by the kinetics of enzymatic reactions alone. It is mainly because the enzymes in the corresponding metabolic pathways respond (primary response) to the perturbations much faster than the transcriptional regulatory network (secondary response), and it takes a little time for the signal to be transduced to the changes in the transcriptional pattern and subsequently to the changes in the total enzyme concentrations.

A brief introduction on how modeling can help with theoretical analysis of a system and how it provides a link between theory and experiment was given above. The system under study in this research is a metabolic network (a network of intracellular biochemical reactions). Definitions and concepts that are required for mathematical modeling of the dynamics of a metabolic network are provided in the following section. Further, the algorithm and mathematics behind “Kinescope” are comprehensively covered. Kinescope is a computational tool with a graphical user interface (GUI) that was developed in MATLAB as another product of this research. The main idea behind design and development of Kinescope is to create a tool that can be used for semi-automatic construction and analysis of kinetic models of biochemical

reaction networks, and to take a further step towards large-scale kinetic modeling of metabolic networks. The algorithm behind Kinescope was designed mainly based on the idea of ensemble modeling [77]. Metabolic ensemble modeling was introduced in 2008 as a novel approach for kinetic modeling of biochemical reaction networks. The approach relies mainly on a steady-state metabolic flux distribution vector as a necessary input to the algorithm and benefits from mass action kinetics as a well-known rate expression by breaking the enzymatic reactions into their elementary reaction steps. Many different parameter vectors are calculated in a way that, when simulated, all converge to the same reference steady state flux distribution. Collected models are then screened based on the additional experimental data from different experiments. Major and minor additions and changes were made to the original work, as they will be discussed in next sections. A major modification for example, was automatic construction of a symbolic Jacobian matrix that can be used to mathematically verify stability of each kinetic model without any need for simulations. Such a modification had a great impact on the required time for collection of stable models in the ensemble and on the reliability of the collected models as well. Kinescope is made upon 16 computational functions for collection and screening of kinetic models, a core script with several functions for GUI objects creation and interactions, and in total, several thousand lines of coding in MATLAB. All the functions and scripts have been written from scratch. In the final sections of this chapter, two case studies will be provided as examples of applying Kinescope to small scale networks with different regulatory mechanisms, and results will be discussed.

4.1. Method

Before constructing a mathematical model, it is better to clearly determine the following objects for the system under study:

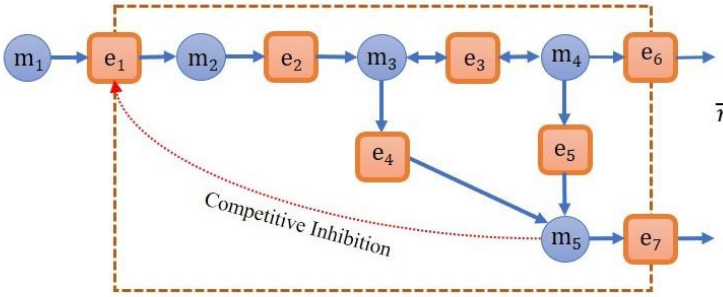
- System Components and their Features: Components of a system are those elements that can interact with each other, and the behavior of the system is determined through those interactions. In a graphical representation of a system, the components are represented as graph nodes, and the interactions among the components can be shown by connecting the corresponding nodes to each other with directed or undirected edges. Components of a system may be of different categories, and every single component in a category can still have its own

individual identity. Each category may have one or more features assigned to it. When modeling a system, one must determine "what the required features are for each category of the components". These two questions can help to find the required features: (i) What "data" is being used to make the model? and (ii) What are the "variables" of the system and how they relate the system inputs to the outputs? In the case of metabolic networks, metabolites and enzymes are two categories of the components. In the classical modeling approach that is based on Michaelis-Menten type rate expressions, intracellular abundance of metabolites (C_m) is usually the only feature assigned to the metabolites, while the enzyme abundance (C_e), rate of the reaction it catalyzes (v), and a kinetic rate expression (f) can be three suitable features for the enzymes. C_m, C_e and v are quantities while f is a mathematical expression. Figure 4.3 is an abstract representation of a Michaelis-Menten (MM)-based kinetic model of a small metabolic network.

- System Boundary: A real (physical) or imaginary boundary that separates the system under study and all of its components from the rest of the universe. In the case of metabolic networks, the real boundaries can be cell membrane or the membrane of the organelles such as mitochondria, but usually it is an imaginary boundary around the metabolites and enzymes of the pathway under study in kinetic modeling.
- Functional Units: Functional units of a system are those components that make an operation on what they receive as inputs, and their outputs depend on that operation. The operation can be represented in the form of a mathematical expression, and it is usually where the essence of mathematical modeling resides. In the case of the metabolic system modeled in Figure 4.3, the functional units are the enzymes.
- Model Variables: Model variables are those features that can change during simulation of the model. They may be independent or dependent variables. In the model represented in Figure 4.3, concentration of the metabolites (C_m) are independent variables while rate of the reactions (v) are dependent variables. Consider that model variables are not necessarily the same as the system variables. For example, enzyme abundance (C_e) is a variable of the system in a metabolic network, and it can change according to the changes in the

transcription, translation and post-translational modification processes. But since there is no mathematical expression to count for those changes, it is not considered as a variable of the model in MM-based model of Figure 4.3.

- Model Parameters: Model parameters are those quantitative values that do not change during the simulation of the model. There can be two types of parameters: (i) "Model-wide parameters" are those that belong to the model as a whole. They usually set a condition for all the components of the system and changing them can lead to changes in the aggregate behavior of the system. When coding a model into a computational language, model-wide parameters can be defined as "global parameters" in the code. (ii) Local parameters are usually the constants in mathematical expressions that represent the operation of the functional units in the system. In the case of the MM-based model of Figure 4.3, maximum reaction rates (v^{max}) and Michaelis-Menten constants (K_{MM}) are among local parameters of the model. Consider that, although the parameter values do not change during a simulation, they can be tuned before each simulation to study how the changes in those parameters (if practically possible) affect the behavior of the system.
- System Features: Any quality or quantity that can identify the system and its behavior as a whole, without focusing on its individual components, can be a feature of the system. Cell morphology, growth rate and transcription pattern are among other features that can be used in the study of unicellular organisms as systems. In the case of a metabolic system, a suitable feature is the metabolic flux distribution vector, which is a vector containing the rates of all the reactions in the corresponding network.
- System State: The state of a system can be represented as a collection of the variable values for one or more types of system variables. In the case of the metabolic system presented in Figure 4.3, the state of the system can be determined by the vector of metabolite concentrations ($\overline{C_m}$), or with a set that includes both metabolite concentrations and reaction rates ($\{\overline{C_m}, \overline{v}\}$). For a dynamic system, the concept of state has a tighter relation with the concept of time. The system can be either in a transition state or in a steady state. To determine the state of the system at a specific time point is similar to taking a snapshot of the system and looking at the values of its variables at that instant.



$$\overline{m} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{bmatrix} \quad \overline{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

m: metabolite
e: enzyme

System Components = $\{\overline{m}, \overline{e}\}$

Component Features:
m: Abundance (e.g. concentration) C_m
e: Abundance C_e, Reaction rate v, Rate expression f

$$\overline{C_m} = \begin{bmatrix} C_{m_1} \\ C_{m_2} \\ C_{m_3} \\ C_{m_4} \\ C_{m_5} \end{bmatrix} \quad \overline{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix}$$

Model Variables: $\overline{C_m}, \overline{v}$

System Feature = $\overline{v}$ System State = $\{\overline{C_m}, \overline{v}\}$

Mathematical Model

Model Reactions and Rate Expressions:

1: $1\,m_1 \rightarrow 1\,m_2$ $\quad v_1 = f_1(C_{m_1}, C_{m_5}) = \frac{v_1^{max} C_{m_1}}{C_{m_1} + K_{MM_1}(1 + \frac{C_{m_5}}{K_{1m_5}})}$

2: $2\,m_2 \rightarrow 1\,m_3$ $\quad v_2 = f_2(C_{m_2}) = \frac{v_2^{max} (C_{m_2})^2}{(C_{m_2})^2 + K_{MM_2} + 2K_{2m_2} C_{m_2}}$

3: $1\,m_3 \leftrightarrow 1\,m_4$ $\quad v_3 = f_3(C_{m_3}, C_{m_4}) = \frac{\frac{v_3^{max}}{K_{3m_3}}(C_{m_3} - \frac{C_{m_4}}{K_{3eq}})}{1 + \frac{C_{m_3}}{K_{3m_3}} + \frac{C_{m_4}}{K_{3m_4}}}$

4: $1\,m_3 \rightarrow 1\,m_5$ $\quad v_4 = f_4(C_{m_3}) = \frac{v_4^{max} C_{m_3}}{C_{m_3} + K_{MM_4}}$

5: $1\,m_4 \rightarrow 1\,m_5$ $\quad v_5 = f_5(C_{m_4}) = \frac{v_5^{max} C_{m_4}}{C_{m_4} + K_{MM_5}}$

6: $1\,m_4 \rightarrow$ $\quad v_6 = f_6(C_{m_4}) = \frac{v_6^{max} C_{m_4}}{C_{m_4} + K_{MM_6}}$

7: $1\,m_5 \rightarrow$ $\quad v_7 = f_7(C_{m_5}) = \frac{v_7^{max} C_{m_5}}{C_{m_5} + K_{MM_7}}$

Balance Equations around each Metabolite:

$$\left.\begin{aligned} \frac{dC_{m_1}}{dt} &= -v_1 + \delta_1 \\ \frac{dC_{m_2}}{dt} &= v_1 - 2v_2 + \delta_2 \\ \frac{dC_{m_3}}{dt} &= v_2 - v_3 - v_4 + \delta_3 \\ \frac{dC_{m_4}}{dt} &= v_3 - v_5 - v_6 + \delta_4 \\ \frac{dC_{m_5}}{dt} &= 2v_4 + 2v_5 - v_7 + \delta_5 \end{aligned}\right\} \quad \frac{d\overline{C_m}}{dt} = (S.\overline{v}) + \overline{\delta}$$

S: Stoichiometric Matrix
δ: Counts for any "external" sources or sinks that affect the abundance of the metabolites. For example, a constant amount of m_1 provided to a continuous reactor (a non-zero δ_1).

$$\overline{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix}$$

$S =$

v_1	v_2	v_3	v_4	v_5	v_6	v_7	
-1	0	0	0	0	0	0	m_1
1	-2	0	0	0	0	0	m_2
0	1	-1	-1	0	0	0	m_3
0	0	1	0	-1	-1	0	m_4
0	0	0	1	1	0	-1	m_5

Figure 4.3: MM-based kinetic modeling of a small metabolic network.

A literature review on the kinetic modeling of metabolic networks was given in Section 3 of Chapter 2. The limitations of classical MM-based modeling and the

obstacles in front of such modeling approach were highlighted. A few approaches towards large-scale kinetic modeling of metabolism were also reviewed, among which the Ensemble Modeling (EM) approach [77] seems very potent and was selected as the template algorithm to make Kinescope. As a part of this research, the original algorithm of EM went through several revisions and modifications. For example, a few simplifying assumptions are used to automatically break down each enzymatic reaction of a given stoichiometric metabolic model into a series of elementary (association/dissociation) reactions that their combination governs the catalytic mechanism of the bulk reaction. Such simplifying assumptions that are necessary for automatic construction of the structure of the kinetic model at the elementary reactions level out of a stoichiometric model, and the reasoning behind them, has never been clearly stated in previous works. Also, after automatic construction of the model at the elementary reactions level, Kinescope provides an interface for the user to manually edit and curate the elementary steps of those reactions whose catalytic mechanisms are well studied in the literature and the simplifying assumptions may not apply to them. Another novelty that greatly enhances the process of collecting stable parameter sets (parameter sets that lead to a stable steady state) is automatic construction of the symbolic Jacobian matrix of the kinetic model. Since all the elementary reactions naturally follow the mass action kinetics, it is possible to have a symbolic Jacobian matrix of the model through symbolic derivation of the balance equations around each molecule. This matrix is then used to evaluate the stability of the model for each parameter set, making it possible to reject those parameter sets that mathematically lead to unstable models, without any need for simulating them. This modification not only leads to collection of more reliable models, but also it decreases the required time for construction of the ensemble by several times. Another issue that is worth mentioning is the possible change in the relative abundance of different forms of an enzyme between two different steady states. For instance, after a gene deletion, not only the total enzyme concentrations may change, but also the distribution among its different forms (e.g. free enzyme, enzyme-substrate complex, …) may also change. Since each steady state has a finite threshold of stability within which the model simulations can converge back to that state, simulating the models for a new state (e.g. after gene deletion) with the same initial condition as the reference state may not work even for true parameter sets. In the following sections, the algorithm behind Kinescope is comprehensively explained.

4.1.1. Construction of the Ensemble

Based on the ensemble modeling approach, suitable dynamic models for a metabolic network are found through two main modules. The first module is to collect so many different models (or the same model structure but with different parameter sets) that, regardless of their dynamic behavior, all converge to a reference steady-state and are mathematically stable at that state (Figure 4.4). The collection is called the "Ensemble" and this first module is titled "Construction of the Ensemble" in this text. The second main module is to use the experimental data from different experiments to reduce the size of the ensemble. The conditions of each experiment are applied to all models in the ensemble, each model is simulated, and the corresponding model variables are compared to those observed/measured in the experiment. If the model output is not in agreement with the experimental observation, that model will be rejected from the ensemble. This procedure can continue until a handful of reliable models remain. In this manner, each experiment acts as a filter to screen the models in the ensemble (Figure 4.5). This module is titled "Screening the Ensemble".

The submodules for construction of the ensemble are explained in the following subsections. But before that, essential and optional inputs to the module are introduced. The optional inputs, whenever available, can efficiently constrain the sampling space for the parameters, hence effectively avoiding the unwanted models to be collected in the ensemble. Figure 4.6 is a schematic representation of how different submodules are organized to construct the ensemble.

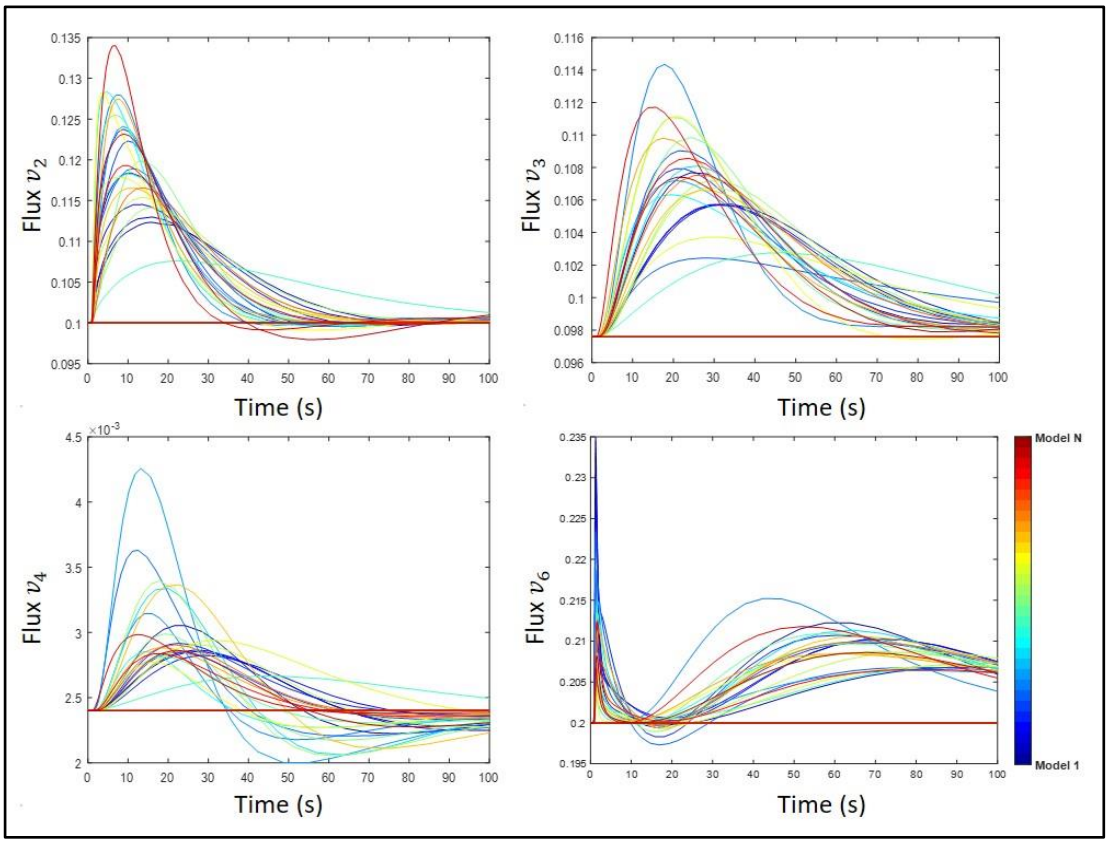


Figure 4.4: Collecting stable models with different parameter sets that all converge back to the reference steady state after a small perturbation.

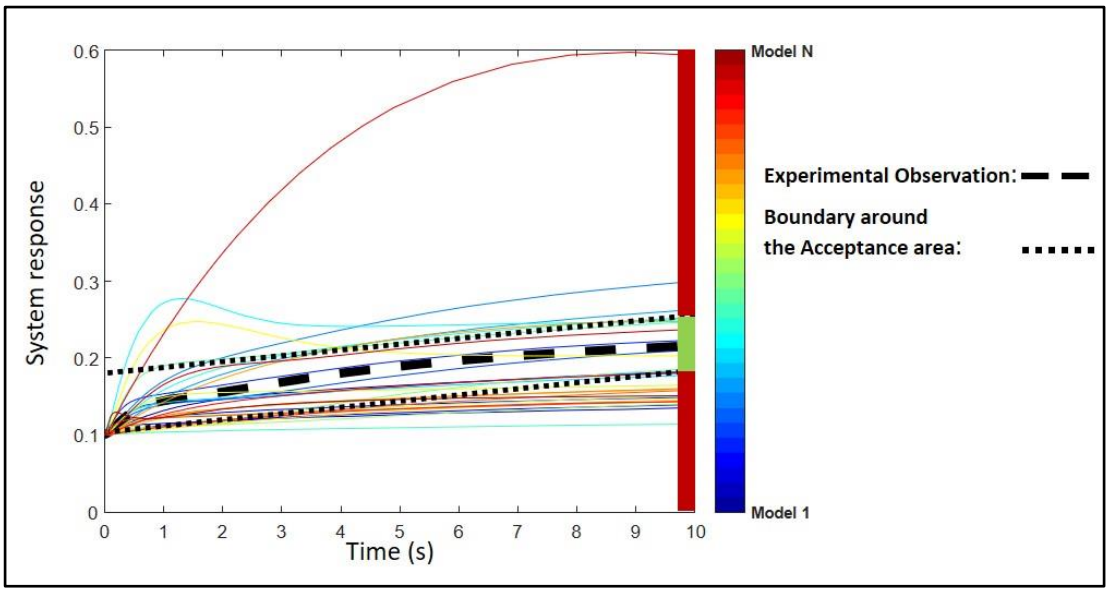


Figure 4.5: Screening the ensemble. The same condition/perturbation (e.g. change in the substrate concentration) as the experiment is introduced to all the models in the ensemble. Only those models that have a similar response (e.g. change in the secretion rate of a specific metabolite) to the experimental observation remain in the ensemble, others are rejected.

- Essential Inputs:
  i) Stoichiometric model of the metabolic network (including the stoichiometric matrix and corresponding list of metabolites and reactions).
  ii) A reliable metabolic flux distribution vector, obtained from a reference experiment at a stable steady-state condition.
- Optional Inputs:
  i) A regulatory matrix, containing information on which enzymatic reactions are regulated (inhibited/activated) by which molecules, and what the mechanism of the regulation is. This matrix can be the same size as the stoichiometric matrix with the same row and column labels for metabolites and reactions. However, instead of the stoichiometric coefficients, the corresponding numbers are some predefined codes (Table 4.1) to determine the type of regulation. By default, this matrix is considered as an empty matrix in Kinescope. Since it directly influences the structure of the model at the elementary reactions level, it is highly recommended to provide this matrix for known regulatory interactions in the network. For example, the regulatory matrix of the model shown in Figure 4.3 would be a 5×6 matrix in which the only non-zero element is the one corresponding to metabolite m_5 and reaction v_1 (element (5,1)), with a value of -1.

Table 4.1: Codes for different regulation mechanisms.

Regulation Mechanism	Code
Competitive Inhibition	-1
Uncompetitive Inhibition	-2
Mixed Inhibition	-3
Allosteric Inhibition	-4
Allosteric Activation	+4

  ii) Standard Gibbs free energy of the reactions, whenever available, can constrain the parameter sampling space through the expressions that are given in Section 4.1.6. This data does not need to be provided for all the reactions. By default, Kinescope assigns “NaN” values to the Gibbs energy feature of each reaction. This data can be partially provided only for those

reactions with a known Gibbs energy, to constrain the parameter sampling space as much as possible.

iii) Lower and Upper bounds for metabolite concentrations are needed to calculate lower and upper values of the Gibbs free energies of the reactions (Section 4.1.6). By default, the lower bound for the concentration of all metabolites is zero and the upper bound is 100 mM, which is much higher than the frequently reported concentrations for most of the metabolites in different metabolome datasets. Whenever experimental data is available on the intracellular concentration of some metabolites in several different cell conditions, such data can be used to come up with tighter boundaries for those metabolites.

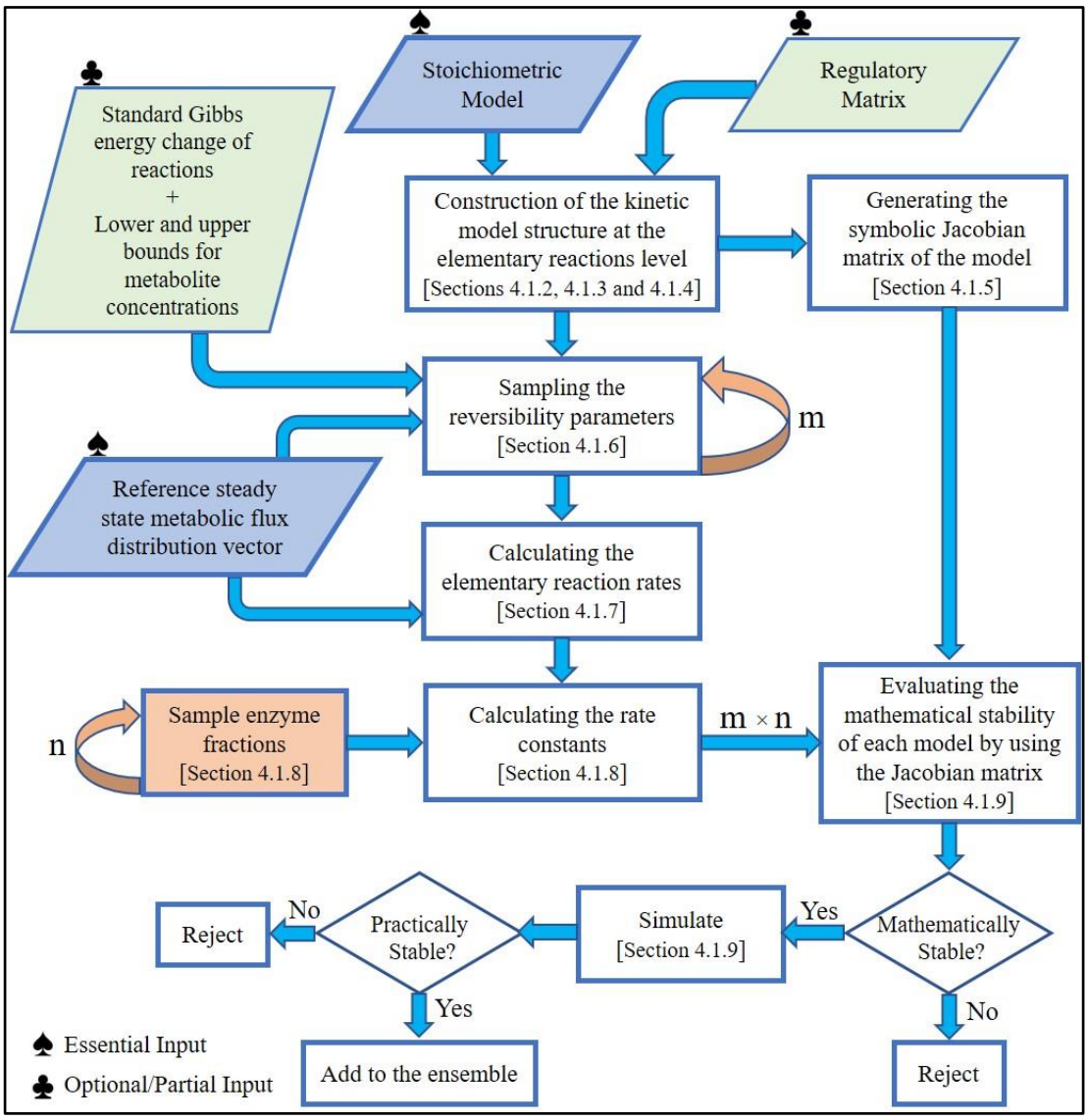


Figure 4.6: Flowchart of the algorithm for construction of the ensemble.

4.1.2. Automatic Break Down of the Enzymatic Reactions to their Elementary Steps

Kinescope automatically generates a set of elementary reactions for each enzymatic reaction of the stoichiometric model based on the following three assumptions:

Assumption 1: The “biochemical transformation step” can be ignored in many enzymatic reactions (Figure 4.7.a). After the attachment of all the required substrates and cofactors to their respective positions on the enzyme, reactants are transformed into the products of the reaction through manipulation of their chemical bonds. This is the transformation step which is in the middle of attachment of reactants and release of products. When dealing with the kinetics of the enzymatic reaction, the transformation step can be ignored in many cases, either because the quasi steady-state assumption of Briggs and Haldane [152] applies to the condition, or because the rapid equilibrium approximation of Michaelis and Menten [122] can be applied.

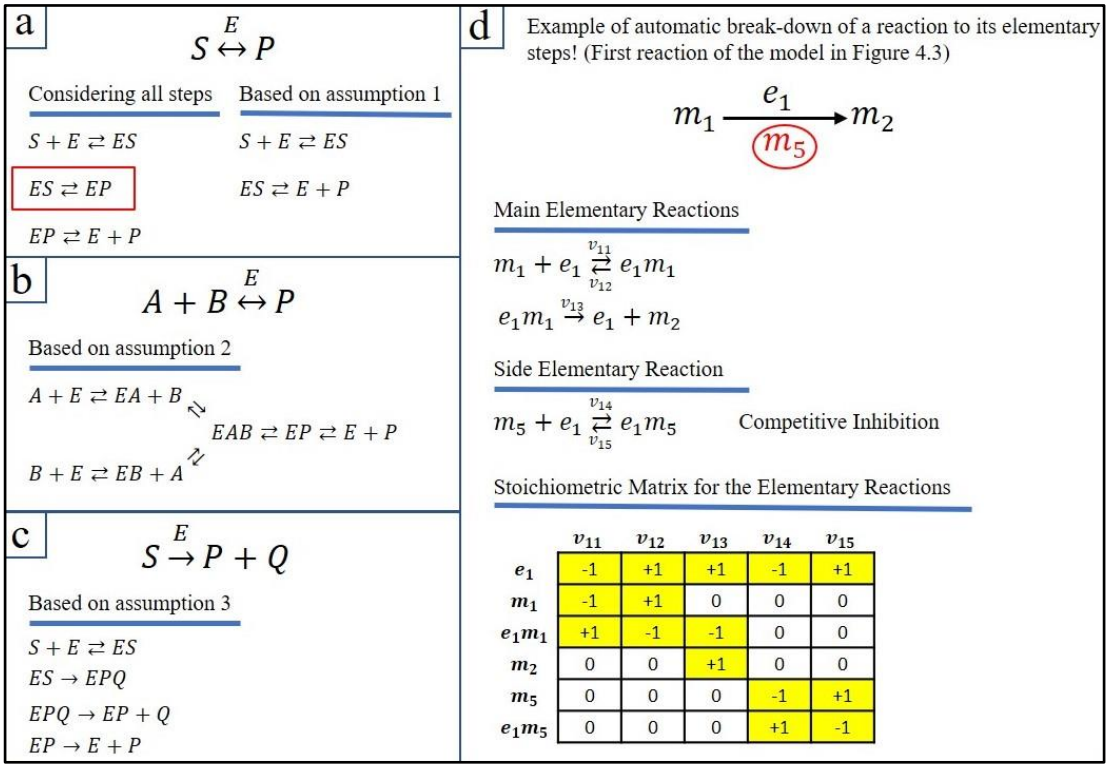


Figure 4.7: Automatic break down of enzymatic reactions to their elementary steps. a) Assumption 1, Biochemical transformation step is ignored. b) Assumption 2, Multi-substrate reactions are assumed to be random sequential. c) In the case of irreversible reactions, all the elementary steps before the transformation step are still reversible. d) Reforming the model to elementary reactions expands the stoichiometric matrix. Number of molecules and reactions increase significantly.

Assumption 2: Multi-substrate reactions are assumed to follow a random sequential mechanism. Unlike reactions with Ping-Pong mechanism, in a sequential mechanism, all reactants are first attached to the enzyme and the transformation happens afterward. Sequential reactions can be ordered or random. In ordered reactions, some reactants cannot attach to the enzyme unless the specific reactants attach first. In this case, the number of possible scenarios to reach to the final complex for the transformation is restricted. The random sequential is the more general case, in which different reactants can attach to the free enzyme, leading to different complexes that are able to become the final complex for the transformation (Figure 4.7.b).

Assumption 3: For irreversible reactions, the elementary steps before the transformation step are considered reversible, while all the other steps are irreversible (Figure 4.7.c). The idea behind this assumption is that the reactants can leave the enzyme before any transformation occurs.

Each column of the stoichiometric matrix in the original model becomes a matrix with several columns and rows in the elementary reactions model (Figure 4.7.d). It must be considered that, although the number of molecules and reactions increase significantly in the elementary reactions model, it is not an exponential increase and follows a linear manner. As the elementary reaction steps are constructed for each reaction of the stoichiometric model, they are categorized into two groups. Those elementary reactions that can be lined up in the form of a series of steps, starting from attachment of the first reactant to the enzyme and ending with release of the last product, are categorized as the “Main Elementary Reactions” (MER). Those elementary reactions that occur in parallel to the main steps, such as attachment of a competitive inhibitor to the free enzyme, are categorized as the “Side Elementary Reactions” (SER). This categorization helps with calculation of the rate constants (demonstrated in section 4.1.7).

4.1.3. Manual Curation of the Elementary Steps

Since the assumptions used for automatic construction of the elementary reactions do not apply to all enzymatic reactions, the possibility to manually curate and edit those elementary steps for reactions of choice is a necessity. Whenever information on the kinetic mechanism of a reaction is available in the literature, its elementary steps must be evaluated, and if the automatically generated ones are

different from the available information, they must be edited accordingly. Kinescope provides an interface to make it possible for the user to manually edit the elementary steps of such reactions. (Figure 4.8.b). An example of a reaction that would need manual curation is fermentation of pyruvate to lactic acid by the lactate dehydrogenase enzyme. This reaction follows an ordered sequential mechanism in which NADH is the first molecule to attach to the enzyme, after that it is possible for the pyruvate to associate. A comparison of the automatically generated elementary steps for this reaction with the one based on the literature is presented in Figure 4.8.a. The corresponding screen shots of the graphical interface of Kinescope for this task are also provided.

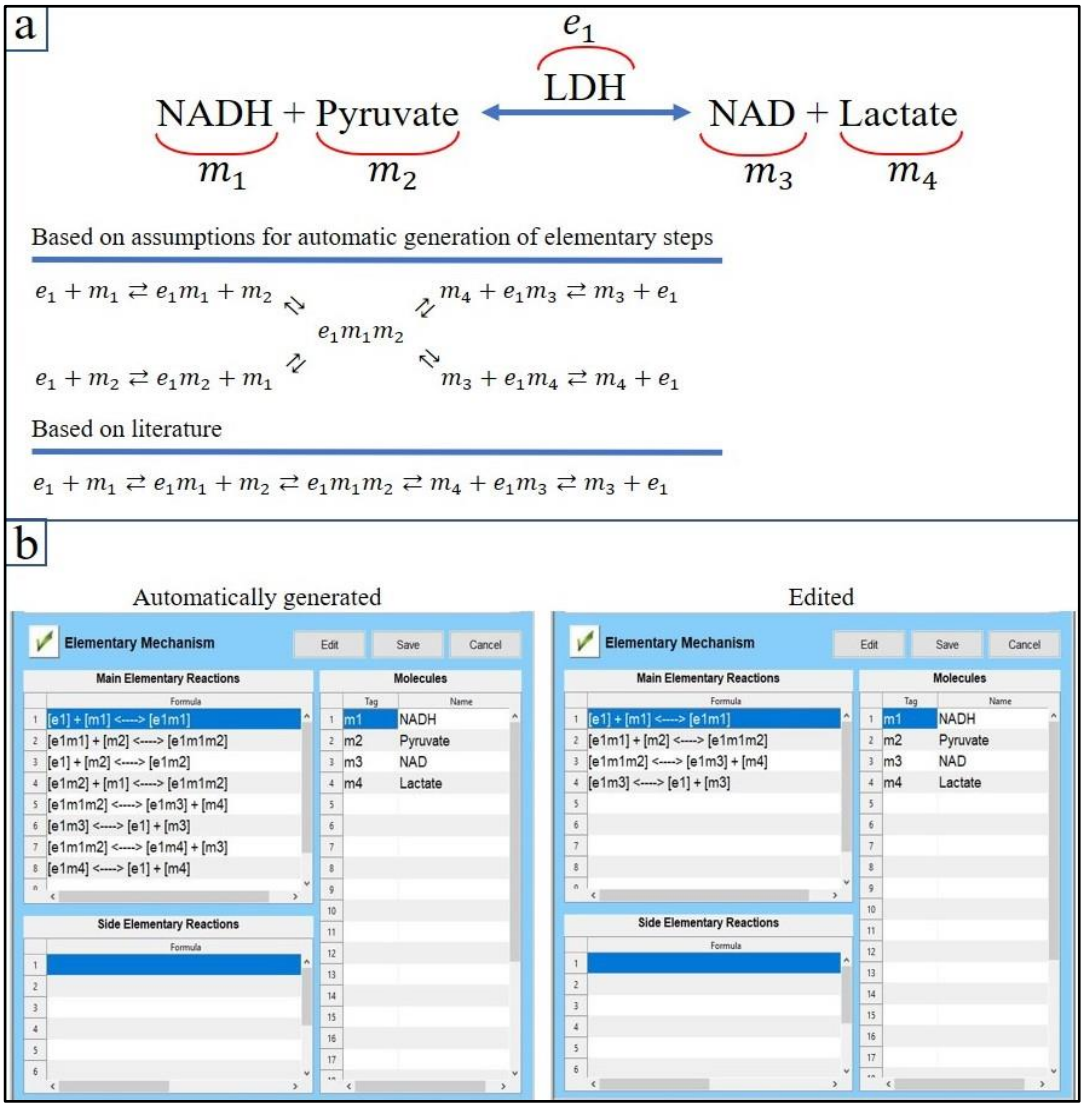


Figure 4.8: Manual curation of the elementary steps. a) Reaction catalysed by lactate dehydrogenase, example of an ordered sequential mechanism. b) Screen shots of the graphical interface of Kinescope to evaluate and edit the elementary steps.

4.1.4. Automatic Construction of the Kinetic Model

Direct attachment of two molecules to form a complex and direct detachment of a molecule from a complex are the most basic events in molecular interactions. An elementary reaction is a chemical reaction in which one or two chemical species react directly to form products in a single reaction step. Elementary reactions are hence the most fundamental kinetic events at the molecular level. The rate of any elementary reaction fundamentally follows the law of mass action. Based on the law of mass action, the rate of the reaction is directly proportional to the product of the concentrations of the reactants. A rate constant is then multiplied by the product of the concentrations of the reactants to make the proportionality relation into an equality relation.

After the model structure at the elementary reactions level is constructed through previous steps, a mass action rate expression is assigned to each elementary reaction in the model. Using these rate expressions and the stoichiometric matrix of the model (automatically constructed by scanning all the elementary reactions), mass balance equations are written around each molecule of the model, including both metabolites and enzyme fractions. However, all the metabolite concentrations are scaled by their corresponding concentrations at the reference steady-state, and those of free enzyme and enzyme complexes are scaled by the total concentration of the corresponding enzyme. Consider the following reaction being the i^{th}reaction in the stoichiometric model, in which the molecule S is isomerized to the molecule P by catalytic activity of enzyme E_i:

$$S \overset{E_i}{\leftrightarrow} P \tag{4.1}$$

The elementary mechanism for this reaction is:

$$S + E_i \underset{v_{i2}}{\overset{v_{i1}}{\rightleftarrows}} E_iS \underset{v_{i4}}{\overset{v_{i3}}{\rightleftarrows}} E_i + P \tag{4.2}$$

The mass action rate law assigned to the first elementary reaction would be:

$$v_{i1} = k_{i1}[E_i][S] \tag{4.3}$$

k_{i1} is the rate constant and brackets around a molecule indicate the concentration of that molecule. The right side of each mass action rate expression can be multiplied and divided by to the product of the total concentration of the corresponding enzyme (E_i in this case) and the concentration of the corresponding metabolite (S in this case) at the reference steady-state:

$$v_{i1} = k_{i1}[E_{i,T}][S]_{ref} \frac{[E_i]}{[E_{i,T}]} \frac{[S]}{[S]_{ref}} \tag{4.4}$$

$$v_{i1} = k'_{i1}[e_i][s] \tag{4.5}$$

$$k'_{i1} = k_{i1}[E_{i,T}][S]_{ref} \ \& \ [e_i] = \frac{[E_i]}{[E_{i,T}]} \ \& \ [s] = \frac{[S]}{[S]_{ref}} \tag{4.6}$$

Normalizing the variables or scaling them by a reference state to make them dimensionless is a common practice in physics and engineering and the original work of ensemble modeling [77] also made use of such practice. Concentration of the corresponding metabolite at the reference steady-state ($[S]_{ref}$ in this example) and the total concentration of the corresponding enzyme ($[E_{i,total}]$ in this example) are now intrinsic to the lumped rate constant (k'_{i1} in this example). Since concentrations of the metabolites at the reference state are constant values and do not change under any circumstance, they do not introduce any challenge in using the normalized equations for simulation of the kinetic model around steady-states other than the reference state. Simulating the model around the reference steady-state leads to the convergence of all the normalized metabolite concentrations (such as $[s]$) to 1. Convergence to other values around other steady-state conditions represents the fold change in the concentration of the metabolites with respect to the reference state. The total concentration of the enzyme ($[E_{i,total}]$) however, can change from one condition to another (e.g. as a result of change in transcription and translation). Although it does not introduce any challenge to simulate the model around the reference steady-state (it does not introduce any problem in the stage of collecting different models in the ensemble), using the same value to simulate conditions other than the reference state

(e.g. when filtering the collected models in the ensemble by using experiments that have different conditions than the reference state) is not rational. To take this fact into consideration, the fold change in the total enzyme concentration with respect to the reference state is multiplied by the lumped rate constants. For example, the rate expression of the Equation 4.5 is modified as follows:

$$v_{i1} = k'_{i1} f_{E_{i,T}} [e_i][s] \tag{4.7}$$

$$f_{E_{i,T}} = \frac{[E_{i,T}]}{[E_{i,T}]_{ref}} \tag{4.8}$$

$f_{E_{i,T}}$ equals 1 at the reference state. The format of Equation 4.7 is used as the universal format for the rate expressions in the construction of the kinetic models in Kinescope. Whenever proteomics data is available for the both states, it can be used for the calculation of $f_{E_{i,T}}$ parameters to simulate the models around a steady-state other than the reference state. In the absence of proteomics data, transcriptomics data can be used under the assumption that protein expression is roughly proportional to gene expression. As a result, having proteomics and/or transcriptomics data for the reference state is highly recommended although they are not necessary to construct the ensemble. Also, perturbation-observation experiments can be designed around the reference state to filter the collected models in the ensemble without any need for proteomics or transcriptomics data. For instance, the dynamic change in the concentration of one or more metabolites in the first few seconds/minutes after a perturbation can be used as an observation to filter the collected models. It is mainly because, even if there would be a change in the total enzyme concentration in response to a perturbation, it can be ignored for the first few seconds/minutes after the perturbation is introduced.

In Kinescope, all the rate expressions and the balance equations, along with other required commands, are automatically written into a text file and saved as an m-file of MATLAB, which can be directly used as the input model file to an ODE solver for simulation. A schematic of the automatically constructed kinetic model at the elementary reactions level is presented in the Figure 4.9, which can be compared to the Figure 4.3. Both kinetic models are for the same stoichiometric metabolic model,

however the one in Figure 4.3 is MM-based while the latter is at the elementary reactions level.

Assignment of mass action rate law to each elementary reaction

Reaction 1: $1\,m_1 \rightarrow 1\,m_2$ competitively inhibited by m_5

Main Elementary Reactions

$e_1 + m_1 \underset{v_{12}}{\overset{v_{11}}{\rightleftarrows}} e_1m_1 \quad v_{11} = k'_{11} f_{E_{1,T}}[e_1][m_1] \quad v_{12} = k'_{12} f_{E_{1,T}}[e_1m_1]$

$e_1m_1 \overset{v_{13}}{\rightarrow} e_1 + m_2 \quad v_{13} = k'_{13} f_{E_{1,T}}[e_1m_1]$

Side Elementary Reaction

$e_1 + m_5 \underset{v_{15}}{\overset{v_{14}}{\rightleftarrows}} e_1m_5 \quad v_{14} = k'_{14} f_{E_{1,T}}[e_1][m_5]$

$v_{15} = k'_{15} f_{E_{1,T}}[e_1m_5]$

Reaction 2: $2\,m_2 \rightarrow 1\,m_3$

$e_2 + m_2 \underset{v_{22}}{\overset{v_{21}}{\rightleftarrows}} e_2m_2 \quad v_{21} = k'_{21} f_{E_{2,T}}[e_2][m_2] \quad v_{22} = k'_{22} f_{E_{2,T}}[e_2m_2]$

$e_2m_2 + m_2 \underset{v_{24}}{\overset{v_{23}}{\rightleftarrows}} e_2m_2m_2 \quad v_{23} = k'_{23} f_{E_{2,T}}[e_2m_2][m_2] \quad v_{24} = k'_{24} f_{E_{2,T}}[e_2m_2m_2]$

$e_2m_2m_2 \overset{v_{25}}{\rightarrow} e_1 + m_2 \quad v_{25} = k'_{25} f_{E_{2,T}}[e_2m_2m_2]$

Reaction 3: $1\,m_3 \leftrightarrow 1\,m_4$

$e_3 + m_3 \underset{v_{32}}{\overset{v_{31}}{\rightleftarrows}} e_3m_3 \quad v_{31} = k'_{31} f_{E_{3,T}}[e_3][m_3] \quad v_{32} = k'_{32} f_{E_{3,T}}[e_3m_3]$

$e_3m_3 \underset{v_{34}}{\overset{v_{33}}{\rightleftarrows}} e_3 + m_4 \quad v_{33} = k'_{33} f_{E_{3,T}}[e_3m_3] \quad v_{34} = k'_{34} f_{E_{3,T}}[e_3][m_4]$

$\vdots$

$$\bar{y} = \begin{bmatrix} [e_1] \\ [e_1m_1] \\ [e_1m_5] \\ [e_2] \\ [e_2m_2] \\ [e_2m_2m_2] \\ [e_3] \\ \vdots \\ [e_6] \\ [e_6m_5] \\ [m_1] \\ [m_2] \\ [m_3] \\ [m_4] \\ [m_5] \end{bmatrix} \qquad \bar{v} = \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \\ v_{14} \\ v_{15} \\ v_{21} \\ v_{22} \\ \vdots \\ v_{61} \\ v_{62} \\ v_{63} \end{bmatrix}$$

Balance Equations

$$\left.\begin{aligned}
\frac{d[e_1]}{dt} &= -v_{11} + v_{12} + v_{13} - v_{14} + v_{15} + \delta_1 \\
\frac{d[e_1m_1]}{dt} &= +v_{11} - v_{12} - v_{13} + \delta_2 \\
\frac{d[e_1m_5]}{dt} &= +v_{14} - v_{15} + \delta_3 \\
\frac{d[e_2]}{dt} &= -v_{21} + v_{22} + v_{25} + \delta_4 \\
&\vdots \\
\frac{d[m_5]}{dt} &= -v_{14} + v_{15} + v_{43} + v_{44} + v_{53} + v_{54} - v_{61} + v_{62} + \delta_{21}
\end{aligned}\right\} \quad \frac{d\bar{y}}{dt} = (S.\bar{v}) + \bar{\delta}$$

$S =$

v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	v_{21}	v_{22}	v_{23}	v_{24}	v_{25}	…	v_{43}	v_{44}	…	v_{53}	v_{54}	v_{61}	v_{62}	v_{63}	
-1	+1	+1	-1	+1	0	0	0	0	0		0	0		0	0	0	0	0	e_1
+1	-1	-1	0	0	0	0	0	0	0		0	0		0	0	0	0	0	e_1m_1
0	0	0	+1	-1	0	0	0	0	0		0	0		0	0	0	0	0	e_1m_5
0	0	0	0	0	-1	+1	0	0	+1	…	0	0	…	0	0	0	0	0	e_2
0	0	0	0	0	+1	-1	-1	+1	0		0	0		0	0	0	0	0	e_2m_2
0	0	0	0	0	0	0	+1	-1	-1		0	0		0	0	0	0	0	$e_2m_2m_2$
⋮											⋮			⋮					⋮
0	0	0	-1	+1	0	0	0	0	0		+1	+1		+1	+1	-1	+1	0	m_5

Figure 4.9: Representation of the automatically constructed kinetic model at the elementary reactions level for the stoichiometric model of Figure 4.3.

In Kinescope, there is a dynamic relation between the model and the corresponding m-files so that the changes made in the model are directly translated into the m-files. Such characteristic of Kinescope, in addition to the semi-automatic construction of models at the elementary reactions level, makes it a valuable tool for studying the dynamism of biochemical reaction networks.

4.1.5. Automatic Construction of the Symbolic Jacobian Matrix

A Jacobian matrix contains valuable information on dynamic characteristics of a dynamic system. It is especially useful to evaluate the stability of a system at any of its steady states (equilibrium points). A dynamic system can be represented with the following equation:

$$\frac{d\overline{y}}{dt} = \boldsymbol{f}(\overline{y}) \tag{4.9}$$

In the case of a metabolic network being the system, $\overline{y}$ would be a vector containing the abundance data of the molecules (as in Figure 4.9). The steady states of the system can be found as solutions to the following equation:

$$\boldsymbol{f}(\overline{y}) = \mathbf{0} \tag{4.10}$$

A system can have more than one steady-state, as it is usually the case for the metabolic networks. However, not all steady states of a system are “stable”. Around a stable steady state, the system has the tendency to keep that state. In other words, it has the tendency to converge back to that state in response to perturbations. On the other hand, when a steady state is unstable, the system has no tendency to keep that state. That means it will diverge from that state in response to the slightest perturbations. Having a stable steady-state experimental condition as the reference point to collect the kinetic models in the ensemble, those parameter sets that lead to mathematically unstable models are of no interest and must be rejected. The stability of a model at a given steady state can be mathematically evaluated by calculating the eigenvalues of the Jacobian matrix of the model. If all the eigenvalues have a negative real part, the model is mathematically stable at that state. On the other hand, even if

only one of the eigenvalues has a positive real part, the model is unstable. If the largest real part of the eigenvalues is zero, the stability of the model at that state is unknown and further investigation is required. For a dynamic system represented by Equation 4.9, elements of the Jacobian matrix can be determined by the following equation:

$$J_{ij} = \frac{\partial(\frac{dy_i}{dt})}{\partial y_j} \tag{4.11}$$

J_{ij} is the element of the Jacobian matrix at row i and column j. For example, the first element of the symbolic Jacobian matrix of the model in Figure 4.9 is as the following:

$$\frac{\partial(\frac{d[e_1]}{dt})}{\partial[e_1]} = -f_{E_{1,T}}(\, k'_{11}[m_1] + \, k'_{14}[m_5]) \tag{4.12}$$

Since mass action kinetics is used as the rate expression for all the elementary reactions, a symbolic Jacobian matrix of the model can be automatically constructed by symbolic differentiation of the balance equations with respect to the concentration of each molecule. Having the symbolic Jacobian matrix, each parameter set sampled from the parameter space can then be used to come up with a numerical Jacobian matrix. In this way, the unstable parameter sets can be easily rejected without any need for simulations. In Kinescope, if the largest real part of the eigenvalues is zero, the corresponding parameter set is still added to the ensemble, since it still has the potential to make a stable model. But all the parameter sets that lead to eigenvalues with positive real parts are rejected. Based on our experience, only a small percentage of the sampled parameter sets lead to stable models. For example, out of 10000 sampling for the kinetic model of Figure 4.9, between 100 to 200 stable models can be collected, that is less than 2%. In addition, the required time to evaluate the stability of the models is incredibly reduced by using the symbolic Jacobian compared to collection of stable models through simulation. For example, evaluating the stability of 10000 parameter sets for the model of Figure 4.9 takes around 30 seconds by using the Jacobian matrix, while it takes more than 75 minutes to simulate those 10000 models on the same

computer. This equals to 150 times reduction in the required time for collection of stable models in the ensemble.

4.1.6. Thermodynamic Constraint

In biochemistry, the degree to which a substance tends to combine with another is defined as “affinity”. For each elementary reaction, the association and dissociation rate constants are directly related to the affinities between the involved molecules. Affinities, and hence the rate constants, cover a wide range of values in the real number system. In addition, the dynamic behavior of a biochemical reaction can be extremely sensitive to some rate constants. Considering these facts, it is not a wonder that the kinetic space is extremely huge, even for a small metabolic model, covering so many different dynamic behaviors. Rational constraints, however, can be exerted on the kinetic space to reduce its size, making it more feasible to find the appropriate rate constants. Anchoring the dynamic models to a reliable steady-state flux distribution, obtained from the reference experiment (explained in the next section), is a very effective method to constrain the kinetic space [129]. In addition, for the reversible reactions in the stoichiometric model, standard Gibbs energy change of the reactions along with the lower and upper bounds on the metabolite concentrations can be used as a further constraint in calculating the corresponding rate constants. Consider the reversible reaction of Equation 4.1 as an example, with standard Gibbs energy change of reaction ΔG_i° and reference steady-state flux value $V_{i,ref}$:

$$S \overset{E_i}{\leftrightarrow} P \qquad \Delta G_i^{\circ}, V_{i,ref} \tag{4.13}$$

The standard Gibbs energy change of a reaction is a measure of how far the standard-state (for substances in liquid solutions: 1 bar pressure and 1 Molar concentration) is from the equilibrium, and is related to the equilibrium constant of the reaction (K_{eq}) through the following equation in thermodynamics:

$$\frac{\Delta G^{\circ}}{RT} = -\ln\,(K_{eq}) \tag{4.14}$$

R being the universal gas constant and T stands for the temperature. The Gibbs energy change of reactions in liquid solutions is not a strong function of the pressure (even if it was, the pressure is not that different from the standard-state in almost all studies of biological cells). However, by changes in the concentrations of the involving molecules, the Gibbs energy change of a reaction (ΔG) can deviate significantly from its value at the standard-state (ΔG°). The following equation links the Gibbs energy change of the reaction to the concentration of the reactants and products:

$$\frac{\Delta G}{RT} = \ln(Q) - \ln\,(K_{eq}) \tag{4.15}$$

Q is the reaction quotient. The following equations can be written for the example reaction of Equation 4.13:

$$\frac{\Delta G_i}{RT} = \ln(Q_i) - \ln\,(K_{i,eq}) \tag{4.16}$$

$$Q_i = \frac{[P]}{[S]} \tag{4.17}$$

Using the lower and upper boundaries for the metabolite concentrations, two extreme cases can be determined for the reaction quotient. One is when the concentration of the products of the reaction are at the minimum ($[P]_{lb}$) and the concentration of the reactants are at the maximum ($[S]_{ub}$), and the other one is when the concentration of the products is at the maximum ($[P]_{ub}$) while the concentration of the reactants being at the minimum ($[S]_{lb}$). Using Equations 4.14, 4.16 and 4.17, the two extreme cases for the example reaction can be represented as the following:

$$\left(\frac{\Delta G_i}{RT}\right)_{lb} = ln\left(\frac{[P]_{lb}}{[S]_{ub}}\right) + \frac{\Delta G_i^{\circ}}{RT} \tag{4.18}$$

$$\left(\frac{\Delta G_i}{RT}\right)_{ub} = ln\left(\frac{[P]_{ub}}{[S]_{lb}}\right) + \frac{\Delta G_i^{\circ}}{RT} \tag{4.19}$$

On the other hand, the elementary reaction steps of the example reaction (Equation 4.13) can be written as the following:

$$S + E_i \underset{v_{i2}}{\overset{v_{i1}}{\rightleftarrows}} E_iS \qquad R_{i1} \tag{4.20}$$

$$E_iS \underset{v_{i4}}{\overset{v_{i3}}{\rightleftarrows}} E_i + P \qquad R_{i2} \tag{4.21}$$

R_{i1} and R_{i2} are the reversibility parameters for the 2 elementary steps of reaction i in the stoichiometric model. The reversibility parameter for each elementary reaction step is defined as the following:

$$R_{ij} = \frac{\min\left(v_{i(2j-1)}, v_{i(2j)}\right)}{\max\left(v_{i(2j-1)}, v_{i(2j)}\right)} \tag{4.22}$$

Based on the definition, the reversibility parameter is always greater than or equal to 0 and less than or equal to 1. It equals to 0 for an elementary step that is not reversible and equals to 1 when the corresponding elementary step is at the chemical equilibrium ($v_{forward} = v_{backward}$). At the reference steady-state, the sign of the corresponding metabolic flux ($V_{i,ref}$ in this example) can be used to reformulate Equation 4.22 to the following:

$$R_{ij,ref} = \left(\frac{v_{i(2j),ref}}{v_{i(2j-1),ref}}\right)^{sign(V_{i,ref})} \tag{4.23}$$

Taking the logarithm of both sides leads to:

$$\ln\left(R_{ij,ref}\right) = sign(V_{i,ref})\left(\ln\left(v_{i(2j),ref}\right) - \ln\left(v_{i(2j-1),ref}\right)\right) \tag{4.24}$$

Summing up Equation 4.24 over all the elementary steps of reaction i and substituting the elementary rates as defined by Equation 4.7 leads to:

$$\sum_{j=1}^{n_i} \ln\left(R_{ij,ref}\right) = sign\left(V_{i,ref}\right)\left(\sum_{j=1}^{n_i} \ln\left(v_{i(2j),ref}\right) - \sum_{j=1}^{n_i} \ln\left(v_{i(2j-1),ref}\right)\right)$$
$$= sign\left(V_{i,ref}\right)\left(\sum_{j=1}^{n_i} \ln\left(k'_{i(2j)}\right) - \sum_{j=1}^{n_i} \ln\left(k'_{i(2j-1)}\right) + \ln([p]) - \ln([s])\right) \quad (4.25)$$

n_i is the number of elementary reaction steps for reaction i, while $[p]$ and $[s]$ are the normalized concentrations (see Equation 4.6) and equal to 1 at the reference steady-state. So, Equation 4.25 becomes:

$$\sum_{j=1}^{n_i} \ln\left(R_{ij,ref}\right) = sign\left(V_{i,ref}\right)\left(\sum_{j=1}^{n_i} \ln\left(k'_{i(2j)}\right) - \sum_{j=1}^{n_i} \ln\left(k'_{i(2j-1)}\right)\right) \quad (4.26)$$

On the other hand, based on the elementary steps (Equations 4.20 and 4.21), the equilibrium constant of the example reaction i can be written as the following:

$$K_{i,eq} = \frac{k_{i1}k_{i3}}{k_{i2}k_{i4}} \quad (4.27)$$

Substituting for the rate constants by using Equation 4.6 leads to:

$$K_{i,eq} = \frac{k'_{i1}k'_{i3}}{k'_{i2}k'_{i4}} \times \frac{[P]_{ref}}{[S]_{ref}} = \frac{k'_{i1}k'_{i3}}{k'_{i2}k'_{i4}} \times Q_{i,ref} \quad (4.28)$$

Taking logarithm of both sides and rearranging the terms gives:

$$\sum_{j=1}^{n_i} \ln\left(k'_{i(2j)}\right) - \sum_{j=1}^{n_i} \ln\left(k'_{i(2j-1)}\right) = \ln\left(Q_{i,ref}\right) - \ln\left(K_{i,eq}\right) \quad (4.29)$$

Combining Equations 4.26 and 4.29 while considering Equation 4.16 leads to the following equation:

$$\sum_{j=1}^{n_i} \ln\left(R_{ij,ref}\right) = sign\left(V_{i,ref}\right)\frac{\Delta G_{i,ref}}{RT} \tag{4.30}$$

Since the exact values for the Gibbs energy change of reactions at the reference steady-state are not known, and to take the uncertainties into account, the equality constraint of Equation 4.30 can be transformed to the following inequality constraint by using Equations 4.18 and 4.19:

$$\left(\frac{\Delta G_i}{RT}\right)_{lb} \leq sign(V_{i,ref})\sum_{j=1}^{n_i} \ln\left(R_{ij,ref}\right) \leq \left(\frac{\Delta G_i}{RT}\right)_{ub} \tag{4.31}$$

Equation 4.31 can be used to verify if the direction of the net flux ($V_{i,ref}$) is thermodynamically allowable, and it can be reformulated to the following:

$$-\sigma_{i,lb} \leq \sum_{j=1}^{n_i} \ln\left(R_{ij,ref}\right) \leq -\sigma_{i,ub} \tag{4.32}$$

While $\sigma_{i,lb}$ and $\sigma_{i,ub}$ are defined as:

$$\sigma_{i,lb} = \min\left(\left|\left(\frac{\Delta G_i}{RT}\right)_{lb}\right|, \left|\left(\frac{\Delta G_i}{RT}\right)_{ub}\right|\right) \tag{4.33}$$

$$\sigma_{i,ub} = \max\left(\left|\left(\frac{\Delta G_i}{RT}\right)_{lb}\right|, \left|\left(\frac{\Delta G_i}{RT}\right)_{ub}\right|\right) \tag{4.34}$$

For each reversible reaction in the stoichiometric model, Equation 4.32 can be used to put a constraint on the reversibility parameters of its elementary reactions. How this thermodynamic constraint affects the kinetic space in which the rate constants are calculated is demonstrated in the next section.

4.1.7. Calculating the Elementary Reactions Rates

In Kinescope, the elementary steps for each reaction are categorized into two groups (as can be also seen in Figure 4.9). Those elementary reactions that can be lined

up in the form of a series of steps, starting from attachment of the first reactant to the enzyme and ending with release of the last product, are categorized as the "Main Elementary Reactions (MER)". Those elementary reactions that occur in parallel to the main steps, such as attachment of a competitive inhibitor to the free enzyme, are categorized as the "Side Elementary Reactions (SER)" (Not every reaction may have side elementary steps). At a steady-state condition, the net rate of each main elementary reaction step ($v_{forward} - v_{backward}$) equals the steady-state flux value of the corresponding bulk reaction in the stoichiometric model. So, the following equation holds true for the MERs at the reference steady-state:

$$v_{i(2j-1),MER} - v_{i(2j),MER} = V_{i,ref} \quad \forall\ (i \in \{1, \dots, M\}\ \&\ j \in \{1, \dots, N_i\}) \tag{4.35}$$

M is the number of reactions in the stoichiometric model and N_i is the number of the main elementary steps for the corresponding reaction i. Equation 4.35 is an algebraic system of linear equations with N_i independent equations and $2N_i$ unknowns. N_i more independent equations are required to solve for the elementary reaction rates. The reversibility parameters defined for each elementary reaction step in the previous section (Equations 4.22 and 4.23) can be used to provide the remaining required equations. By combining Equations 4.23 and 4.35, the following formulas can be obtained for the calculation of the MER rates at the reference state:

$$v_{i(2j-1),MER} = \frac{V_{i,ref}}{1 - R_{ij,ref}^{sign(V_{i,ref})}} \tag{4.36}$$

$$v_{i(2j),MER} = \frac{R_{ij,ref}^{sign(V_{i,ref})} V_{i,ref}}{1 - R_{ij,ref}^{sign(V_{i,ref})}} \tag{4.37}$$

For the reversible reactions in the stoichiometric model, the reversibility parameters can be sampled according to the inequality constraint of Equation 4.32. Otherwise, they are sampled randomly between 0 and 1. Using the sampled reversibility parameters in Equations 4.36 and 4.37, different elementary reactions rates are calculated in a way that the reference steady-state flux distribution is satisfied. Unlike MERs, the SER rates are not bound to the steady-state fluxes of bulk reactions,

and they cannot be calculated in the same manner as mentioned above. However, the net flux for each SER step must be equal to zero at any steady-state (e.g. the concentration of dead-end molecules such as enzyme-inhibitor complexes does not change at a steady-state). As a result, the following equation holds true for SERs at the reference steady-state:

$$v_{i(2j-1),SER} - v_{i(2j),SER} = 0 \tag{4.38}$$

In the current version of Kinescope, $v_{i(2j-1),SER}$ rates are randomly sampled between a lower and upper bound and Equation 4.38 is used to calculate $v_{i(2j),SER}$ rates. Bimolecular rate constants have an upper limit that is determined by how frequently molecules can collide, and the fastest such processes are limited by diffusion. In general, a bimolecular rate constant has an upper limit of $10^{10}\ M^{-1}s^{-1}$. As a result, $v_{i(2j-1),SER}$ rates are sampled according to the following inequality constraint:

$$0 < v_{i(2j-1),SER} < 10^{10}[X]_{ub}[E_i]_{ub} \tag{4.39}$$

$[X]_{ub}$ and $[E_i]_{ub}$ are the upper bound concentrations for the corresponding metabolite (e.g. a competitive inhibitor) and corresponding enzyme respectively. The default upper bound (if an upper bound vector is not provided as input) for all molecules in the system is $100\ mM$ (Section 4.1.1). The highest ever reported value for the total protein content of biological cells has been 4 million proteins per cubic micron ($1\ fL$) cell volume [153]. Using Avogadro number, this equals to a concentration of almost $6.6\ mM$. Consider that this concentration is for the total protein content of the cell and not for an individual protein. On the other hand, the most abundant protein found in a cell has been reported to be *RplL* in *E. coli* with an estimated value of 110,000 copies per cell [154], that is roughly equal to $167\ \mu M$. Considering these numbers, a concentration of $300\ \mu M$ was selected as the default upper bound for each enzyme. Using these default upper bound values for the metabolites ($100\ mM$) and enzymes ($300\ \mu M$), the inequality constraint of Equation 4.39 becomes:

$$0 < v_{i(2j-1),SER} < 3 \times 10^3\ Ms^{-1} \tag{4.40}$$

The above inequality is used as the default constraint to sample the SER rates. Whenever a specific vector can be provided as the upper bound for the metabolites (e.g. from the metabolomics datasets), and/or proteomics data is available at the reference steady-state condition, such data can be utilized through Equation 4.39 to come up with more specific constraints than Equation 4.40.

At this point, the rates of all the elementary reactions in the model can be determined for different samples of reversibility parameters and SER rates. Calculation of the corresponding rate constants is explained in the next section.

4.1.8. Sampling the Enzyme Fractions and Calculating the Rate Constants

At the reference steady-state, normalized metabolite concentrations (e.g. $[s]$) and the fold change in the total enzyme concentration ($f_{E_{i,T}}$) are equal to 1. As a result, Equation 4.7 can be used in the following way to calculate the elementary rate constants:

$$k'_{ij} = \frac{v_{ij}}{[e_{ij}]} \tag{4.41}$$

$[e_{ij}]$ is the fractional concentration of the corresponding form of enzyme i at the elementary reaction j. For example, $[e_{i1}]$ is the fractional concentration of the free form of enzyme i (represented as $[e_i]$ in Equation 4.7), and $[e_{i2}]$ is the fractional concentration of the "enzyme i-first substrate" complex (e.g. E_iS in the example reaction of Equation 4.1). While the elementary reaction rates are provided according to the previous section, fractional concentrations for each enzyme are sampled according to the following constraints:

$$\left(\forall\, i,j\colon 0 < [e_{ij}] < 1\right) \,\&\, \left(\sum_{j=1}^{n_i} [e_{ij}] = 1 \quad \forall\, (i \in \{1,\dots,M\} \,\&\, j \in \{1,\dots,n_i\})\right) \tag{4.42}$$

M being the number of total reactions in the stoichiometric model and n_i being the number of elementary steps for the corresponding reaction i. In the current version of Kinescope, fractional concentrations are sampled randomly from a uniform distribution while satisfying the constraints of Equation 4.42. More sophisticated sampling methods may improve the computational efficiency and may be applied in the future versions.

4.1.9. Collecting Stable Models in the Ensemble

After the structure of the kinetic model at the elementary reactions level is constructed, as demonstrated in sections 4.1.2 to 4.1.4, different sets of rate constants are calculated through the methodology described in sections 4.1.7 and 4.1.8, leading to generation of kinetic models with different dynamic characteristics. Although all the generated kinetic models satisfy the reference flux distribution as their steady-state solution (Equation 4.10), only those models that are stable at the reference state are meant to be collected in the ensemble. The stability of each model is first evaluated mathematically by using the symbolic Jacobian matrix (constructed according to the Section 4.1.5). For each generated kinetic model, the calculated rate constants and corresponding sampled enzyme fractions are substituted in the symbolic Jacobian to come up with a numerical Jacobian matrix at the reference steady-state. The mathematical stability of each model is then judged based on the eigenvalues of the numerical Jacobian, as explained in Section 4.1.5. A mathematically stable model means that it has the tendency to keep its state in response to the smallest possible (ε) perturbations. Since each stable condition has a practical threshold, and since the models that are not practically stable (based on the reference steady-state experiment) are of no interest, mathematically stable collected models are simulated in response to a small perturbation, and those models that do not converge back to the reference state are further rejected. In this way, an ensemble of different kinetic models that not only satisfy the reference steady-state experiment, but also are practically stable at that state, is constructed. The more the number of models that are collected in such an ensemble, the higher will be the probability of catching those kinetic models that behave as close as possible to the real metabolic system.

4.1.10. Screening the Ensemble

As already mentioned in Section 4.1.1, ensemble modeling of a metabolic network consists of two main modules. In the first module, different kinetic models that are all consistent with a reference steady-state are collected (Figure 4.6). Fundamentally, it is impossible for a single physical system to be at more than one state (e.g. metabolite concentrations and/or metabolic fluxes) at a specified time and space (Section 4.1). Hence, the dynamic behavior of a metabolic pathway in a single cell can be ideally represented with only a single kinetic model. However, there are uncertainties regarding the structure of the metabolic model (some undiscovered interactions that are present in the real metabolic system, especially regulatory interactions, may be missing in the models), and also regarding the elementary mechanism of some enzymatic reactions. Taking such uncertainties into account, all the models whose simulation results are in agreement with experimental observations are considered as potentially valid models. On the other hand, most of the experiments and studies in practice (such as in metabolic engineering applications) are not based on a single cell, but instead on a community of cells usually in a continuous culture. In such cases, an ensemble of different kinetic models that all satisfy the bulk experimental observations (e.g. biomass production, substrate uptake rates, product secretion rates, etc.) can be preferred over a single model, since it can cover for the heterogeneity of the cellular states in the culture (to be more specific, different combinations of enzyme fractions in this modeling approach).

After using the first module to construct the ensemble of kinetic models, available experimental observations can be used in the second module to screen the collected models. Figure 4.10 is a schematic representation of this process. At this stage, it is very important to use the experiments that are compatible with the criteria and conditions based on which the kinetic models were constructed. For example, an experiment in which the metabolism shifts from one steady state to another in response to a perturbation, with changing total enzyme concentrations during that shift (e.g. as a result of change in the transcription), cannot be simulated by the kinetic models constructed according to the procedure of the first module. In the current version of Kinescope, the fold change in the total enzyme concentrations with respect to the reference experiment can be provided as one of the inputs to the simulations, but they are treated as constant factors during simulations, and their temporal changes are not

mathematically represented in the models. However, the first few seconds/minutes of the metabolism dynamics in such experiments can still be simulated, based on the assumption that it takes some time for the total enzyme concentrations to change in response to a perturbation. According to the procedure presented in Figure 4.10 for the second module, different experiments can be used as filters to screen the collected models in the ensemble. For each filter, the corresponding experimental conditions are introduced as inputs and/or rules to the kinetic models in the ensemble. Each model is then simulated, and the simulation outputs are compared with the corresponding experimental observations. Those models that are not in acceptable agreement with the experiments are then rejected, reducing the number of potentially valid models in the ensemble and increasing the reliability of remaining models at the same time. Finally, the remaining models in the ensemble can be used for predictions. They can all be simulated in response to a specific perturbation, and their simulation results provide a statistical base for predictions. Probability values can be calculated for each different outcome based on the percentage of the models that lead to each outcome.

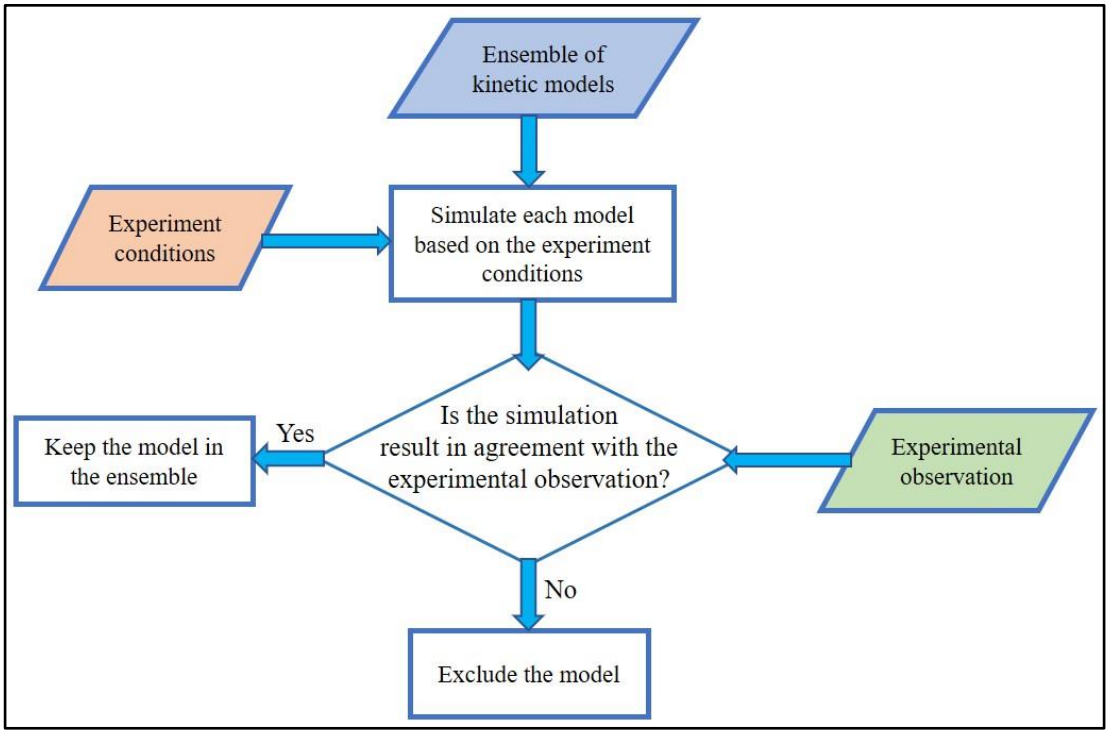


Figure 4.10: Flowchart of the algorithm for screening the ensemble.

At some point, it might happen that none of the models in the ensemble can satisfy an experiment. When this happens, any or a combination of the following three

issues can be the reason behind the inconsistency between the experimental results and simulations:

i) The kinetic space has not been sampled sufficiently, and models with suitable kinetic parameters have not been collected in the ensemble.
ii) Metabolic network structure is incomplete. One or more reactions and/or regulatory interactions are missing in the models.
iii) The elementary reaction steps or mechanistic rate equations are not valid for one or more reactions.

The first issue can be circumvented by increasing the number of samples from the kinetic space, collecting as much as possible different kinetic models in the ensemble. The bigger the size of the ensemble, and the higher the diversity of sampled kinetic parameter sets, the less will be the probability of missing suitable models in the ensemble. The other two issues are the key points in hypothesis generation for discovery of new interactions in the corresponding metabolic network, or for deduction of valid kinetic mechanisms for corresponding reactions in the network. In this manner, they provide a link between the theoretical and experimental playgrounds as demonstrated in Figure 4.2.

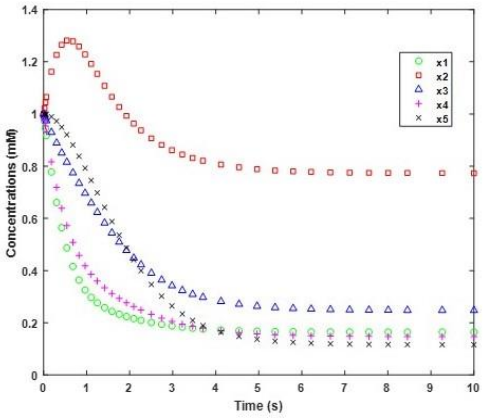
4.2. Results

To validate applicability of Kinescope in the construction of stable kinetic models for a desired reaction network, and to demonstrate the methodology, ToyModel of Figure 4.3 was used as a reference model. This ToyModel includes 5 metabolites and 7 reactions (1 input, 3 irreversible, 1 reversible, 2 outputs) with feedback inhibition of the uptake reaction by one of the secretion metabolites. The ToyModel was constructed manually by assigning MM-based rate expression to each reaction. An irreversible MM equation including competitive inhibition was assigned to the uptake reaction. The second reaction follows an irreversible MM-based mechanism with two identical molecules as the substrates. A reversible MM equation was assigned to the third reaction, and other reactions all follow simple irreversible MM rate expression. The structure and equations of the ToyModel can be seen in the Figure 4.3, and the corresponding MATLAB files including the ODE function of the model are provided in Appendix A (digital format). Biologically meaningful values

(reported in the literature and enzyme/reaction databases) were assigned to all the kinetic parameters in the MM-based rate expressions except for the V_{max} ones. An already balanced and biologically meaningful (in terms of reaction rates) metabolic flux distribution was then used to calculate the V_{max} parameters. This guaranteed the convergence of the ToyModel to a valid steady state.

First, ToyModel is simulated and a stable steady state is captured as the reference state. The steady-state flux distribution at the reference state along with the stoichiometric model of the ToyModel are then used as inputs to the first module of Kinescope to construct a kinetic model and collect as many stable models as possible. Originally, Kinescope would automatically make a kinetic model at the elementary reactions level for the given reaction network (Section 4.1.4), but it can also be used to scan the kinetic space and collect stable models if lumped-kinetic rate expressions are provided for each reaction in the network. This is a better alternative to the parameter estimation approach, mainly because it makes the parameter uncertainty issue (associated with estimated parameters) to fade away and prevents from overfitting of the model to a certain dataset. At the same time, samples taken uniformly from the parameter space provide a theoretical basis to study and analyze the biotransformation potential of a biochemical reaction network at the reaction kinetics level. As mentioned before, different steady states of a dynamic system could be determined as solutions to Equation 4.10. However, using a rational initial condition, simulation of the dynamic model over a large time span can also be used to find a steady state of the system. Simulation results and verification of the stability of the achieved steady state for the ToyModel are presented in Figure 4.11. Stoichiometric model of the ToyModel and the corresponding regulatory matrix are provided as the primary inputs to the Kinescope through the “Construction” tab (Figure 4.12). At this stage, “Prepare the Elementary Reactions Model” button can be clicked to automatically construct the kinetic model structure at the elementary reactions level.

a



b

	Steady State (mM)	Eigenvalue
x_1	0.1645	-4.8583
x_2	0.7731	-3.8412
x_3	0.2475	-0.9399
x_4	0.1466	-1.6076 + 0.4170i
x_5	0.1157	-1.6076 - 0.4170i

Steady State Fluxes ($\frac{mM}{s}$)

r_1	r_2	r_3	r_4	r_5	r_6	r_7
1	0.5	0.373	0.127	0.101	0.272	0.228

Figure 4.11: Reference Steady State. a) Simulation profiles. b) Steady state metabolite concentrations and reaction rates and corresponding eigenvalues.

Detailed stoichiometric information on the metabolites and the reactions are provided within the tables at the right side of the window under the construction tab. Under the “Curation” tab of Kinescope (Figure 4.13), it is possible to view the elementary reaction steps. Required changes on the reaction mechanisms can be made through this interface as described in Section 4.1.3. In the right panel in the curation window (ODE Set and Simulation File), a name is given for the corresponding m-files that are automatically created and used for the ODE simulations and Jacobian calculations. Required changes in the ODE simulation file, such as addition of the feed-stream influxes to the corresponding uptake molecules, can be made within this panel.

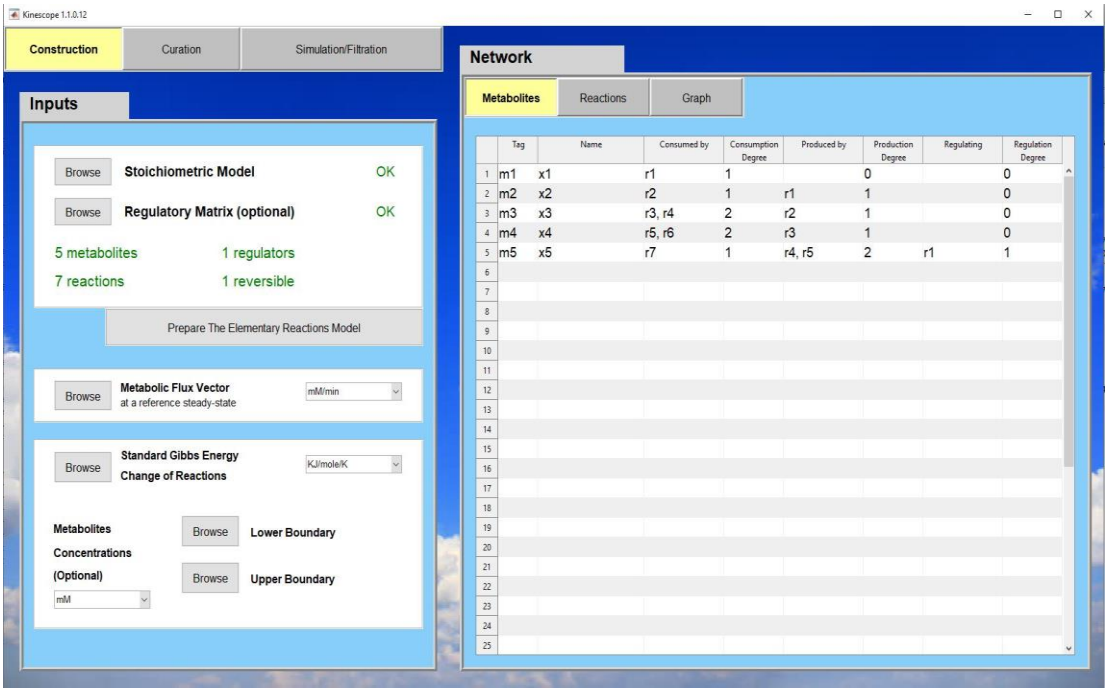


Figure 4.12: Kinescope, construction tab.

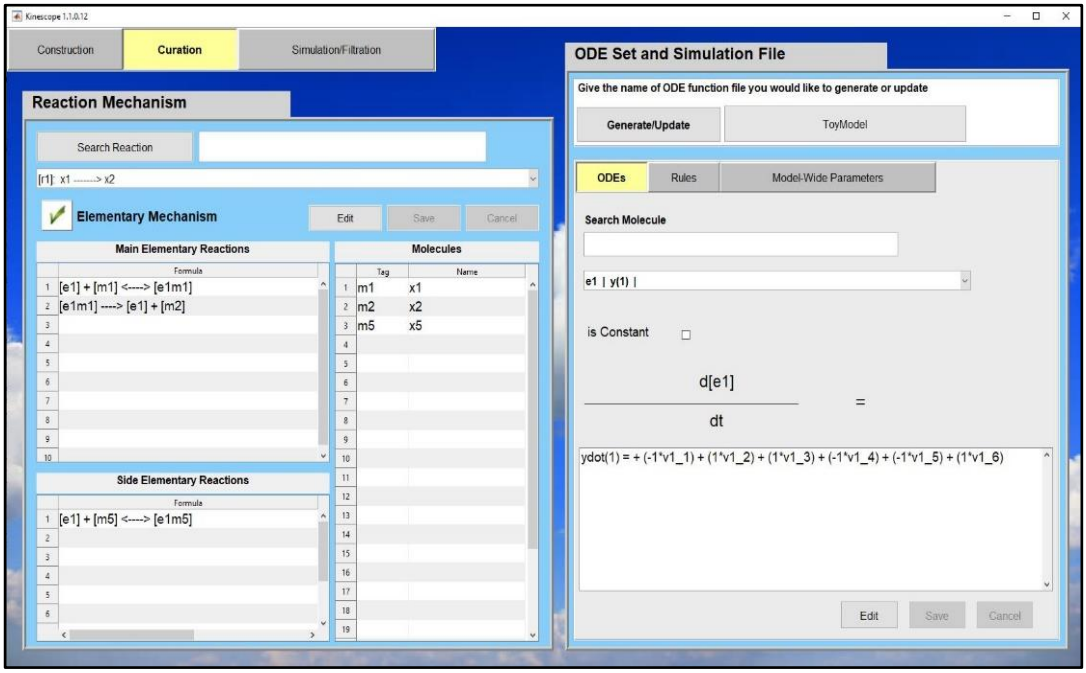


Figure 4.13: Kinescope, curation tab.

At this stage, the reference steady-state flux distribution vector along with other available data are provided through the “Inputs” panel in the construction window (Figure 4.12). It is now possible to construct an ensemble of kinetic models by sampling the kinetic space. In the “Parameter Set Collection” panel under the “Simulation/Filtration” tab, the maximum number of samples to be taken from the

kinetic space can be set by the user, and the stable models will be collected accordingly after pressing the "START" button (Figure 4.15). Afterwards, different experiments can be used to screen the collected models in the ensemble. For example, an impulse change (Figure 4.14.a) in the concentration of the uptake molecule (x_1) was introduced as a perturbation to the reference model ToyModel1. The simulation results were recorded as the system response (Figures 3.14.b and c). Any of the recorded responses may be selected as the experimental observation (Figure 4.14.d). The selected experimental observation is imported to the Kinescope by clicking the "Import Reference Observed Data" button under the "Simulation/Filtration" tab (Figure 4.15). Under the "Curation" tab, the "ODE Set and Simulation File" panel can be used to introduce the same perturbation to the kinetic models in the ensemble. In the end, the "SCREEN" button can be clicked to compare the simulation result of each model in the ensemble with the experimentally observed data (Figure 4.15). Those models with a similar response to the experimental data are kept in the ensemble, while the others are excluded.

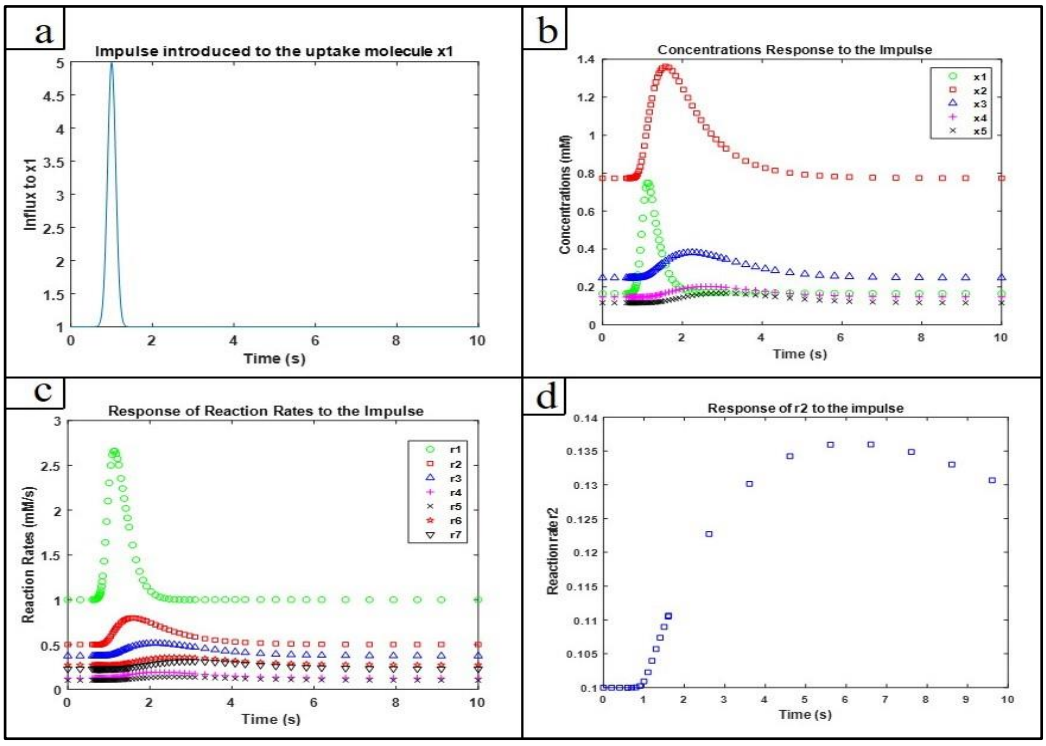


Figure 4.14: ToyModel response to a perturbation-observation experiment. a) The impulse function introduced to the uptake molecule. b) Metabolite concentration profiles. c) Reaction rate profiles. d) Selected reaction rate profile as experimental observation.

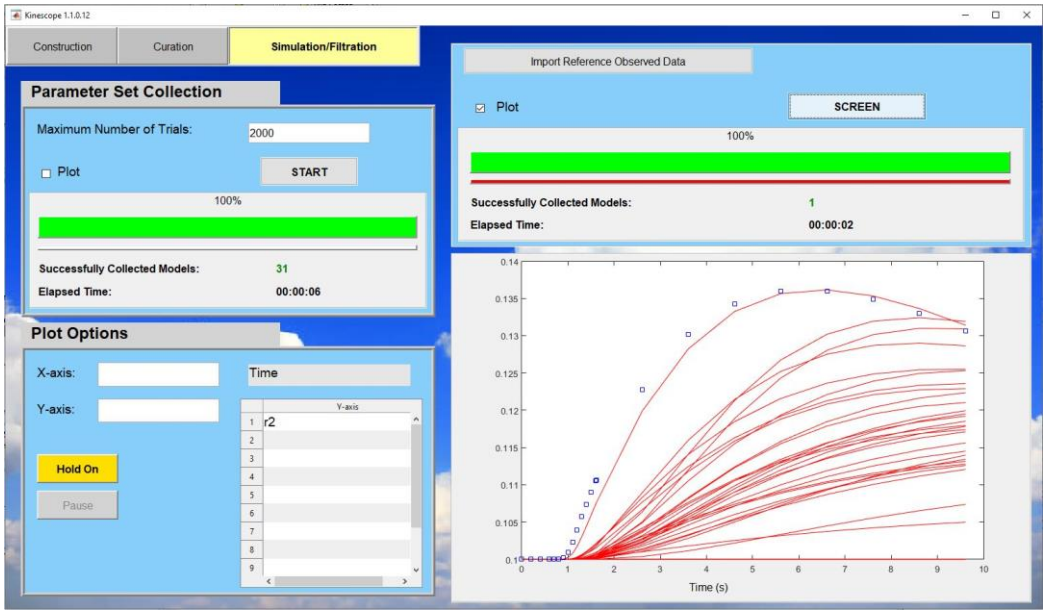


Figure 4.15: Kinescope, simulation/screening tab.

4.2.1. Using Kinescope to Collect Stable Kinetic Models Satisfying Different Reaction Deletion Studies while Lumped-Kinetic Rate Expressions are Available

Gene deletion studies play a central role in metabolic engineering. One of the major questions for a metabolic engineer is, “what is the optimum gene deletion strategy to increase the yield of a desired metabolite in a specified organism?”. Such organisms (usually unicellular such as bacteria or yeast) are known as cell factories in metabolic engineering. Traditionally, such studies were carried out by introducing random mutations to the cells and screening them based on their capacity in the production of the desired metabolite. As soon as genome-scale stoichiometric models of metabolism became available, many researchers started using them for metabolic engineering purposes, especially for the prediction of the optimum set of genes whose deletion would maximize the production yield. “How much the prediction results may be improved if the reaction kinetics are also taken into account?” has been a question in the research community for more than a decade. Because of the difficulties in kinetic modeling of large reaction networks, there has not been a clear answer to the above question. In this section, Kinescope is used to collect stable kinetic models that are in agreement with gene deletion experiments. The ToyModel from Figure 4.3 (also used in the previous section) is used as a reference model to generate *in-silico* data for

reaction deletion experiments. Figure 4.16 is a schematic representation of the *in-silico* reaction deletion experiments.

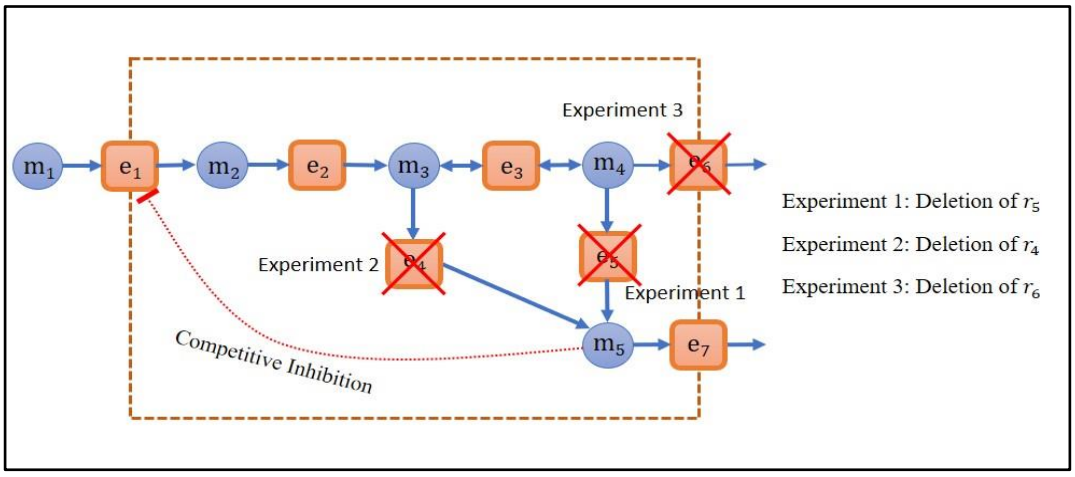


Figure 4.16: Schematic representation of the in-silico reaction deletions in the ToyModel.

Each *in-silico* reaction deletion experiment is carried out by setting the total concentration of the corresponding enzyme in the ToyModel to zero and simulating the model until a new steady-state is achieved. Figures 4.17.a and 4.17.b represent the steady-state values obtained in different experiments for the metabolite concentrations and reaction rates respectively. All the required MATLAB codes to reproduce the work (simulations, analyses, model collection and figures) presented in this section are provided in Appendix A (digital format).

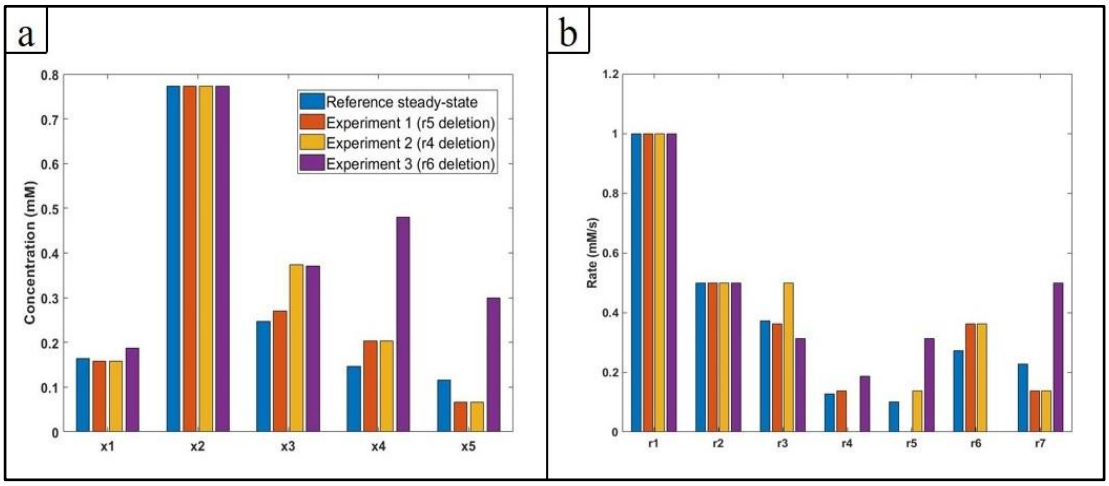


Figure 4.17: *In-silico* reaction deletion experiments. a) Final steady state concentration values. b) Final steady state reaction rate values.

In this section, lumped-kinetic rate expressions are provided for each reaction in the model, so the problem is reduced to the sampling of the kinetic parameter space and collecting those stable models that are in good agreement with reaction deletion experiments. The lower bound for all the kinetic parameters in the rate expressions was set to 0 and the upper bound to 20. Parameters are then sampled uniformly between the lower and upper boundaries. Ten thousand samples were taken in this study, out of which 9520 were stable according to the eigenvalues of their Jacobian matrices. This takes less than a minute on a workstation with the following characteristics:

- System Model: HP Z640 Workstation
- 2x Processors: Intel(R) Xeon(R) CPU E5-2620 v4 @2.1GHz, 8 Cores, 16 logical processors
- Physical Memory (RAM): 64 GB
- Operating System: Microsoft Windows 10 Pro
- MATLAB Version: R2019a Update 6 (academic version)

All the simulations and other computations reported in Chapters 3 and 4 were carried out on the same system with the above characteristics. Figure 4.18.a is a representation of the sampled parameter sets in the principal component space (first 3 principal components explaining 30% of variability in the 18 kinetic parameters of the model). One important question at this stage is, “In a relative comparison between the non-zero elements of the Jacobian matrix, how sensitive the stability of the kinetic model at the reference steady-state is to the changes in each Jacobian element?”. To answer this question, the calculated values for each non-zero Jacobian element (one value for each collected stable model from the uniform sampling, 9520 values in total) are scaled to an interval of 0 to 100 (minimum value to zero and maximum to 100). The number of stable models in each percentile is then counted and the standard deviation is calculated. Figure 4.18.b is a heatmap representation of such data. Based on this heatmap, the stability is most sensitive to the variables and parameters that appear in the mole balance equations around x_3 and x_4 (labeled as m_3 and m_4 during the model construction in Kinescope).

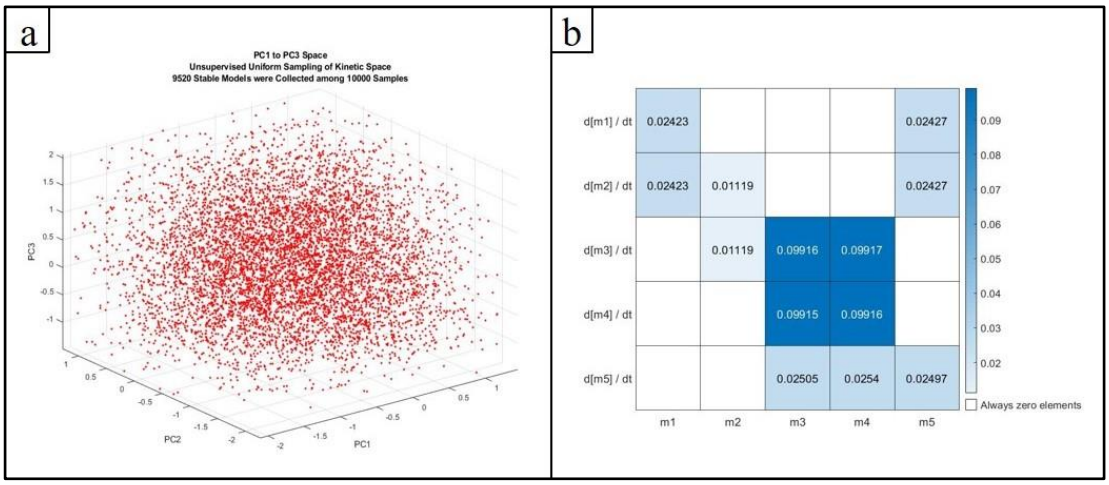


Figure 4.18: Unsupervised uniform sampling of the kinetic parameter space. a) Distribution of the sampled parameter vectors in the principal components (PC1 to PC3) space. b) Heatmap representation of the variance in the scaled values of Jacobian elements.

Next step is to analyze the sampled parameter space by comparing the simulation results of the collected stable models with their corresponding experimental values. For each reaction deletion experiment, total concentration of the corresponding enzyme is set to zero, and all the collected stable models are simulated over a large time span. For those models that successfully reach a new steady-state, fold changes in the concentrations of the metabolites are calculated (with respect to their steady-state values before the reaction deletion). The absolute relative errors between the calculated fold changes (from simulations) and their corresponding experimental values are determined. Collected stable models are then sorted based on their maximum relative error. This procedure is done independently for each reaction deletion experiment. The idea is to find those areas in the kinetic parameter space that lead to generation of models that their simulation results are in good agreement with all available experiments. A threshold is defined to distinguish good models with respect to each experiment. This threshold is called “Good Models Error Cutoff (GMEC)”. If the maximum absolute relative error between the simulation results of a model and the corresponding experimental values is less than the GMEC, that model is labeled as a good model for that experiment. GMEC is the same for all experiments, and the number of good models may differ from one experiment to another. A threshold is also defined to distinguish the elite models and it is called “Elite Models Error Cutoff (EMEC)”. In addition to the cutoff, there is also a limited quota available for the number of elite models that can be assigned to an experiment and there cannot

be more than 10 elite models for each experiment. If there are more than 10 models leading to a maximum absolute relative error less than the EMEC, only the best 10 will be selected as the elite ones. In this study, GMEC is set to 0.1 and EMEC to 0.01. Among the 9520 collected stable models, 5802 and 5235 models were detected as good models for experiment one and two respectively. But only 249 models were detected as good models with respect to the experiment three. Figure 4.19 shows the distribution of good models for the first experiment, based on the first two principal components (explaining 18.9% of variability in the kinetic parameters). Good models are shown with little yellow stars.

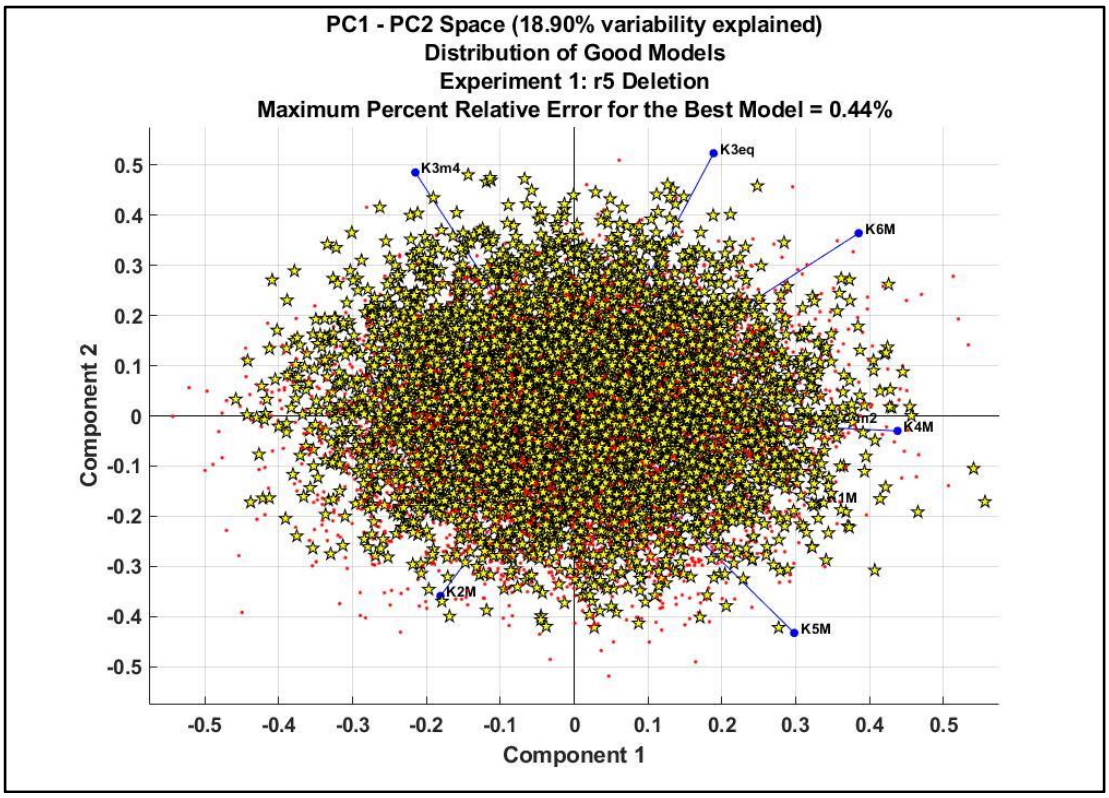


Figure 4.19: Distribution of the good models based on the first experiment.

As it can be seen, good models are dispersed evenly in the space. A similar graph was obtained for the second experiment. In comparison, good models collected based on the third experiment present a different distribution. As represented in Figure 4.20, although the first two principal components explain only 18.9% of the variability in the parameters, a clustering of good models based on the third experiment seems possible. They are more concentrated in the area with positive values for the first principal component and negative values for the second. Figures 4.21 and 4.22 present

the distribution of elite models for the first and second experiments respectively (there wasn't any elite model detected for the third experiment). Such analyses provide insight for development of algorithms that may detect suitable areas in the kinetic space and hence sampling that space more efficiently in a supervised manner. As mentioned earlier, the idea is not to fit the models to the experiments, but to find as many different patterns as possible in the parameters vector that lead to generation of reliable kinetic models. From this point of view, the problem is reformed to a pattern recognition problem, while each parameter is a feature of the system, and machine learning methods can be employed to find the interested patterns.

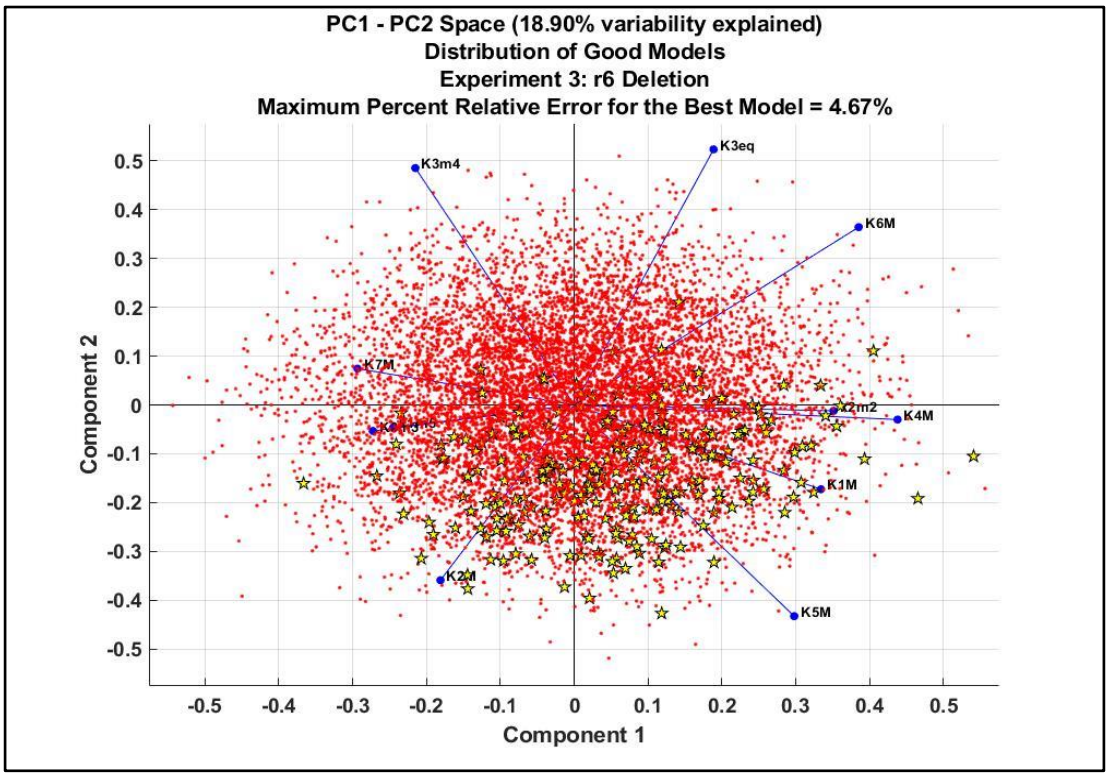


Figure 4.20: Distribution of the good models based on the third experiment.

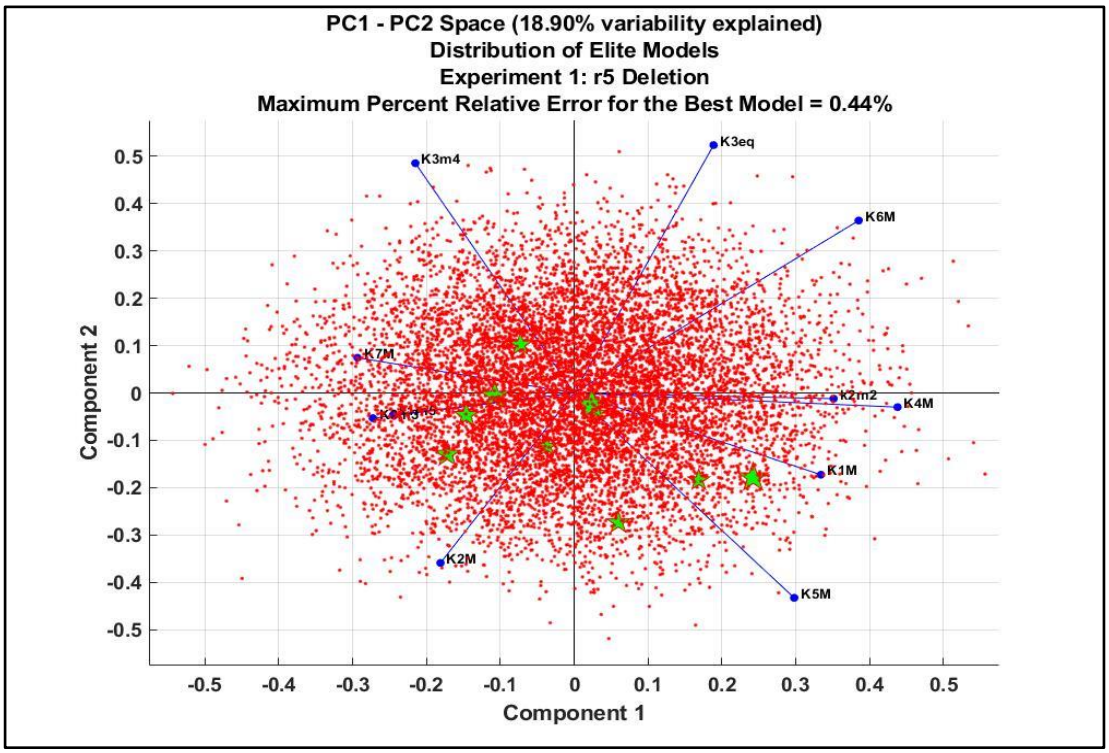


Figure 4.21: Distribution of the elite models based on the first experiment.

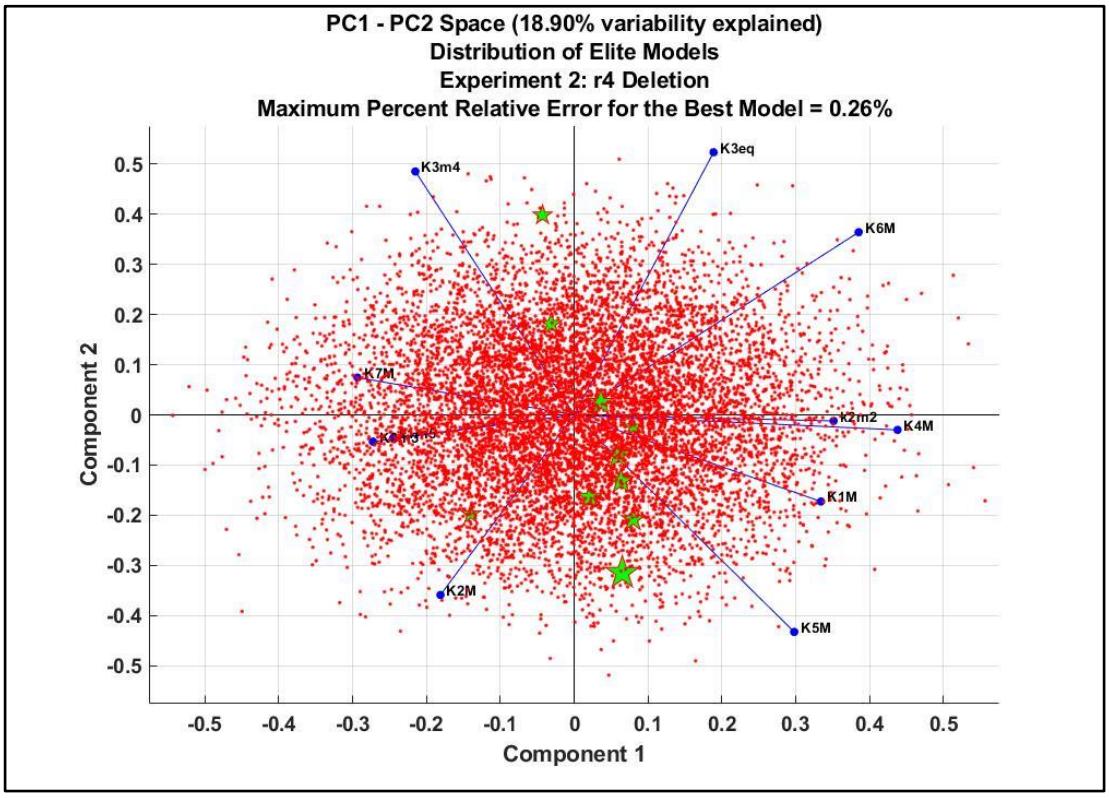


Figure 4.22: Distribution of the elite models based on the second experiment.

Utilization of machine learning methods for unsupervised classification of the sampled parameter vectors is not implemented in the Kinescope yet. For the time being, a different approach was practiced for supervised resampling of the kinetic space. An evaluation table is constructed based on the number of models that could satisfy different experiments (Table 4.2). Evaluation table helps in selection of a set of parameter vectors to be used in generation of seed for supervised sampling.

Table 4.2: Evaluation table after unsupervised sampling of the parameter space.

Experiment	Number of models with a maximum absolute relative error less than **20%**	**10%**	**5%**	**1%**
$\boldsymbol{r_5}$ **deletion**	7750	5802	3309	30
$\boldsymbol{r_4}$ **deletion**	7337	5235	1588	17
$\boldsymbol{r_6}$ **deletion**	1832	249	1	0
All experiments	1600	201	1	0

Selected parameter vectors from the evaluation table (yellow shaded cell with 201 models in this study) are aligned on top of each other (as row vectors) to form a matrix. Each column of this matrix represents one of the kinetic parameters. The mean and standard deviations are then calculated for each parameter. At this stage, the goal is to use the characteristics of those parameter vectors that lead to promising simulation results with respect to all the experiments (e.g. the yellow shaded cell is selected in this case), for a supervised resampling of the kinetic parameter space, to increase the chance of collecting more models that are in better agreement with all the experimental observations. Hence, it is a good idea to make use of machine learning methods to recognize the patterns in those promising parameter vectors and use those patterns for supervised resampling of the parameter space. However, it is not the case in the current version of Kinescope, and a simpler approach is followed. The calculated mean and standard deviations mentioned above are being used to resample the parameter space so that, instead of sampling the parameters from a uniform distribution between their lower and upper bounds, parameters are sampled from a normal distribution with the corresponding calculated mean and standard deviation values for each parameter. Following this procedure led to collection of more models that are in better agreement with all the experiments (Table 4.3).

Table 4.3: Evaluation table after the supervised sampling of the parameter space.

Experiment	Number of models with a maximum absolute relative error less than 20%	10%	5%	1%
r_5 deletion	9435	8594	6953	82
r_4 deletion	9276	8706	4112	98
r_6 deletion	7731	2213	19	0
All experiments	7401	2053	13	0

After supervised sampling, thirteen models were identified to have less than 5% relative error in their simulation results compared to the all three reaction deletion experiments (green shaded cell in Table 4.3). The variability in the magnitude of each kinetic parameter in these thirteen models is evaluated. As can be seen in Figure 4.23, all the parameters show a significantly high variance with more than 50% variation above and below their average value. This is an evidence that collected parameter vectors that satisfy the reaction deletion studies are not necessarily similar. In Figure 4.24, the simulation profiles of the fold change in the concentration of each metabolite based on the experiment 3 (deletion of r_6) are provided.

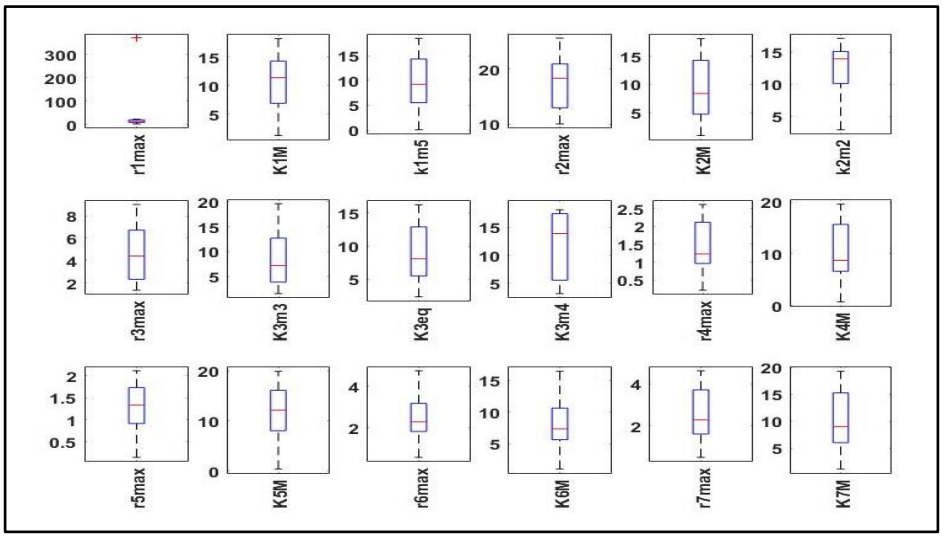


Figure 4.23: Boxplots for the kinetic parameters based on their values in the collected thirteen models.

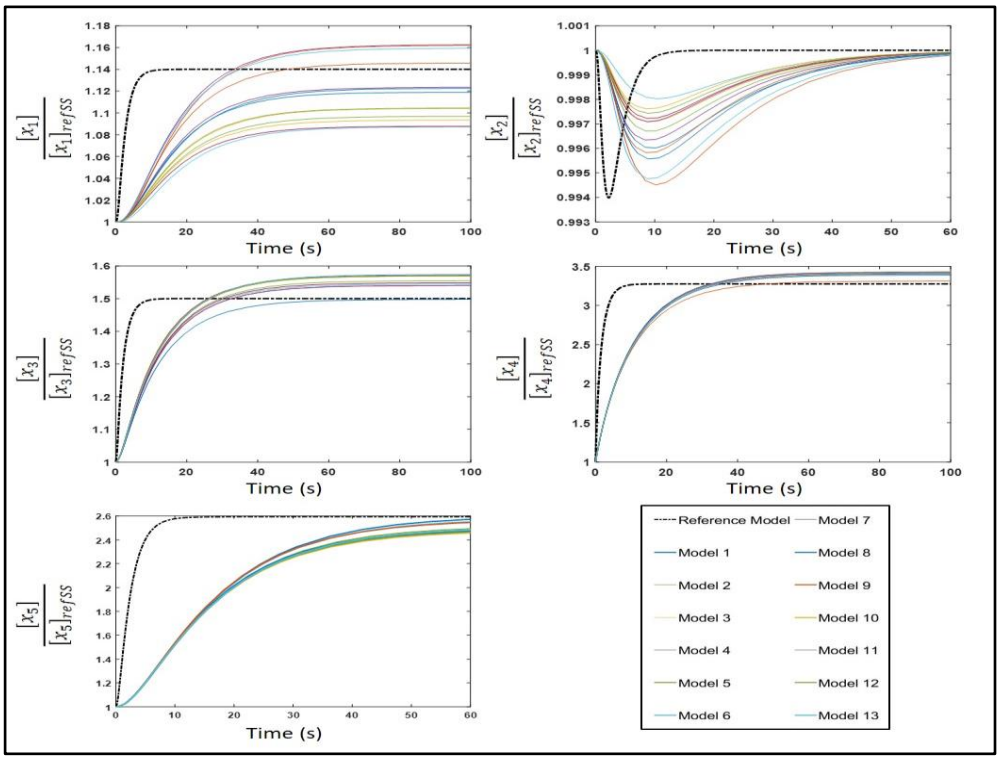


Figure 4.24: Simulation profiles of the fold changes in the concentration of metabolites based on the experiment 3. The simulation profiles for the collected 13 models with less than 5% relative error can be compared with eachother and with that of the reference model.

4.2.2. Using Kinescope for Automatic Construction and Collection of Stable Kinetic Models at the Elementary Reactions Level that Satisfy Different Reaction Deletion Studies

The same reference steady-state and reaction deletion experiments are used in this section as were used in the previous one. However, lumped-kinetic rate expressions are not provided, and the kinetic model structure is automatically created at the elementary reactions level. The elementary reaction sets are provided in excel format in Appendix A. All the required MATLAB codes to reproduce the results obtained in this section are provided in Appendix A. Figure 4.25.a represents the unsupervised sampling of the kinetic parameters in the principal component space. Its difference with Figure 4.18.a is obvious. It is mainly because the kinetic parameters for the elementary reactions are not directly sampled between some lower and upper bounds. Instead, they are being calculated based on the samples taken for the reversibility parameters (constrained by the reaction thermodynamics), enzyme fraction samples (constrained by a linear equality for each enzyme), and the reference steady-state fluxes.

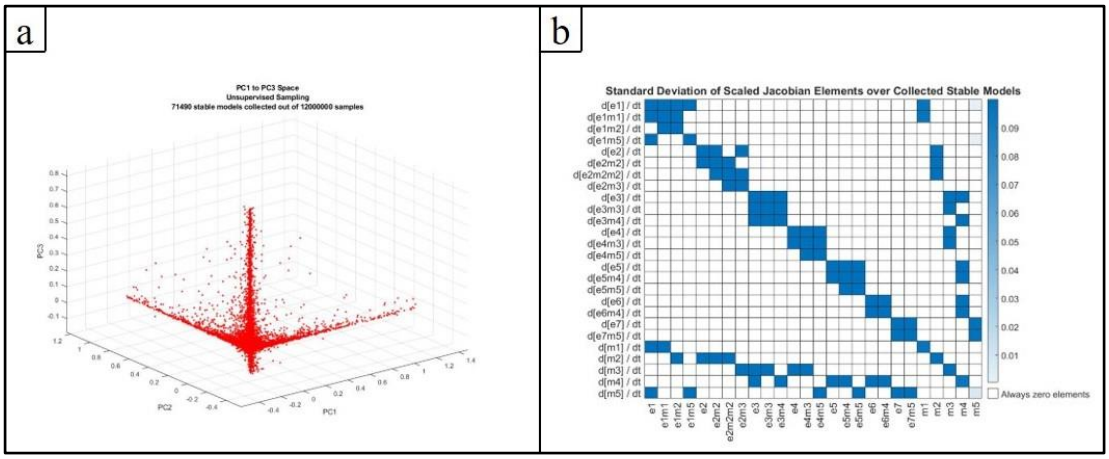


Figure 4.25: Unsupervised sampling of the kinetic space for kinetic models at the elementary reactions level. a) Distribution of the sampled parameter vectors in the principal components (PC1 to PC3) space. b) Heatmap representation of the variance in the scaled values of Jacobian elements.

More than seventy-one thousand stable models were collected from twelve million samples (3464 reversibility samples and 3464 enzyme fraction sampling for each reversibility set). The evaluation table after unsupervised sampling is presented below (Table 4.4).

Table 4.4: Evaluaiton table after the unsupervised sampling of the kinetic space for the models at the elementary reactions level.

Experiment	Number of models with a maximum absolute relative error less than **20%**	**10%**	**5%**	**1%**
r_5 deletion	1826	177	0	0
r_4 deletion	81	2	0	0
r_6 deletion	0	0	0	0
All experiments	0	0	0	0

These results are primarily interpreted as a failure in rational sampling of the kinetic space and suggest development of an algorithm for efficient sampling of reversibility parameters and enzyme fractions. Regardless of such failure, two stable kinetic models were collected (yellow shaded cell in Table 4.4), which not only satisfy the reference steady-state, but also are in good agreement with two different reaction deletion experiments (less than 10% relative error). Considering that these models were constructed automatically at the elementary reactions level (the only inputs to the

Kinescope was the stoichiometric model and the reference steady-state flux distribution), and also considering the hardships in manual construction of kinetic models that could satisfy reaction deletion experiments, this can be counted as a significant achievement. No algorithm has yet been developed for supervised sampling of the kinetic space for models constructed at the elementary reactions level.

4.3. Discussion

The importance of kinetic modeling of metabolic reactions is well recognized. However, the difficulty in developing kinetic models for metabolic systems due to lack of kinetic parameters has been repeatedly reported by many researchers. The need for an efficient and standard methodology for construction and analysis of kinetic models has been long recognized and is becoming more obvious. A methodology that has the potential to become a standard for construction of kinetic models of biochemical reaction networks was presented in this chapter. A computational tool (Kinescope) was developed in MATLAB based on this methodology. The idea of ensemble modeling was used as a template, modified and expanded significantly during the development of Kinescope. The algorithms behind different functions of Kinescope were explained in detail through sections 4.1.2 to 4.1.10. Applicability of Kinescope in construction of stable kinetic models for small-scale reaction networks that satisfy reaction deletion/overexpression studies was validated by *in-silico* experiments (Section 4.2). Whenever lumped-kinetic rate expressions are available (Section 4.2.1), Kinescope can be used to scan the parameter space and collect as many mathematically stable models that all converge to a reference steady-state. Dynamic characteristics of the collected models can be evaluated by analyzing their Jacobian matrices. Afterward, Different experiments (such as gene deletion/overexpression experiments) can be used to screen the collected models and identify different patterns in the parameter vector that lead to generation of promising models. Hence, a handful of kinetic models with different dynamic characteristics that all satisfy the experiments may be collected. Whenever rate expressions are not available (Section 4.2.2), Kinescope uses a few assumptions to automatically break down each reaction in the stoichiometric model into a series of elementary reactions. If information on the kinetic mechanism of an enzyme is available, the automatically generated elementary reaction series for the corresponding reaction can be manually edited through a graphical user interface.

Elementary reactions intrinsically follow mass action rate expressions. Unlike other simplified (e.g. Michaelis-Menten) or generalized (e.g. lin-log and S-systems) rate expressions, elementary reactions are not based on any simplifying assumption on the enzyme kinetics and are the closest mathematical representation of the molecular associations and dissociations. However, by breaking each reaction into its elementary steps, the number of model variables (in addition to the metabolites, each enzyme fraction would be a new variable) and kinetic parameters increases significantly (Not exponentially but linearly, 7-10 elementary steps for each reaction in average). As a result, development of algorithms that can efficiently sample the kinetic space to collect stable models with diverse patterns in their parameter vectors is critical (one of the future works), especially in case of constructing kinetic models for larger reaction networks. Another advantage of modeling at the elementary reactions level is that, in cases when none of the collected models can satisfy an experiment, it is easier to generate hypothesis for the reason behind the disagreement between the model simulations and the experimental observation (as explained in Section 4.1.10). It may be possible to develop an algorithm that can automatically generate, and rank order such hypotheses.

Successful utilization of the ensemble modeling approach, mainly for metabolic engineering purposes, has been reported in the literature [79, 131, 132]. However, there has not been any report of a computational tool for semi-automatic construction of kinetic models of biochemical reaction networks based on the idea of ensemble modeling. In addition, unlike the previously reported works that rely on qualitative screening of the models (increase, decrease or no change in the abundance of the observed molecule or rate of the observed reaction), successful screening and collection of models that are in quantitative agreement with a reference scenario is the first to be reported in this study.

4.3.1. A Note on Parameter Identifiability

Identifiability analysis considers the question of whether it is possible to determine the parameters of a model from data. Those parameters that cannot be learned from data are said to be unidentifiable. There are two types of unidentifiability, 1) Structural Unidentifiability and 2) Practical Unidentifiability. If parameters cannot be inferred from an infinite amount of perfect data, the model is structurally

unidentifiable. Structurally unidentifiable models tend to make the same predictions for different parameter values. On the other hand, a model may be structurally identifiable, that is, in theory it is possible to learn all of its parameters from data, but it may not be practical to do so. For example, it may require an unreasonable amount of data. Based on the following theorem, Fisher Information Matrix (FIM) can be used to evaluate the structural identifiability of a model:

"A model is locally structurally identifiable at θ_0 if and only if its fisher information matrix is non-singular at θ_0".

This method needs an estimate of the parameter values (θ_0) as an input and evaluates the structural identifiability of the model around the given parameter values, hence the identifiability evaluation being local and not global. If the Fisher Information Matrix has a zero eigenvalue, then the model is structurally unidentifiable. A possible extension for the evaluation of the practical identifiability can be as follows: "If FIM of a model has small eigenvalues, then the model is practically unidentifiable". This way of thinking about practical identifiability is useful but is not necessarily correct.

Currently, a major obstacle in classical kinetic modeling of biochemical reaction networks is the parameter estimation step. Such models are usually very non-linear and include many parameters. Some of those parameters are often practically unidentifiable, that is, their values cannot be "uniquely" determined from the available data. Possible causes can be lack of influence on the model outputs, interdependence among parameters, and poor data quality. From this perspective, uncorrelated parameters are the key tuning knobs of a predictive model. Therefore, before attempting to perform parameter estimation, it is important to characterize the subsets of identifiable parameters and their interplay. Once this is achieved, it is still necessary to perform parameter estimation, which poses additional challenges. One must consider that, even after solving the parameter identifiability and parameter estimation problems and finding a set of parameters that best fit available experimental data, it does not guarantee that the estimated parameters are the true parameters of the system as it is often observed that many models which have been calibrated based on a set of experimental data fail to represent (predict) a new experiment. This is usually due to the incompleteness of the models. That is, one or more elements that play a significant role in the new experiment do not exist in the model. However, even a complete model may fail to predict a new experiment, and this is closely related to the topic of parameter uncertainty. As it was mentioned earlier, an advantage of the methodology

presented in this chapter is minimization of the parameter uncertainty by unsupervised uniform sampling of the parameter space and collecting as many as diverse parameter sets that all satisfy a reference steady state experiment and are the potential candidates to be the true system parameters. This approach not only circumvents the problems that arise during parameter identifiability and parameter estimation, but also provides a suitable platform for hypothesis generation whenever the collected models fail to predict a new experiment.

4.3.2. A Note on the Required Computational Time and Curse of Dimensionality

To evaluate the required time for Kinescope to construct and screen an ensemble of kinetic models for a small-scale network, and also to evaluate its dependency on the number of samples taken from the parameter space, analyses of Section 4.2.1 were repeated for different number of samples. Table 4.5 provides a summary of the information collected from the repeated analyses.

Table 4.5: Computational time and number of successfully screened models versus the number of samples taken from the parameter space.

Number of samples	Elapsed Time (minutes)	Number of screened models with less than 5% error for all three experiments	
		Unsupervised	Supervised
1000	1	0	2
5000	4	1	8
10000	7	1	13
20000	14	3	32
30000	21	3	39
50000	34	3	65
80000	57	4	116
100000	78	5	144

As it can be seen, both the required computational time and the number of models with less than 5% error after supervised sampling are increased in a linear manner with respect to the number of samples. One must consider that although the

number of successfully screened models increases by increasing the number of samples, the diversity among the collected parameter sets may not necessarily increase. After some point, as the number of samples increases, it is expected to collect parameter vectors that are very similar to the already collected ones.

Another topic worth mentioning at this stage is “curse of dimensionality”. Curse of dimensionality refers to the difficulties that arise when dealing with data in high-dimensional space, for example when the data has too many features. Among several other domains such as combinatorics and optimization, sampling is one of the domains that frequently suffers from the curse of dimensionality. The common problem is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. There is an exponential increase in the volume associated with adding extra dimensions to a mathematical space. In the case of the methodology presented in this chapter, as the number of the parameters (length of the parameter vector) increases with the size of the biochemical reaction network, the number of samples that must be taken from the kinetic parameter space to provide a suitable initial ensemble of collected stable models increases exponentially. As a result, although the required computational time is a linear function of the number of samples (as can be seen in Table 4.5), since the number of required samples increases exponentially with the size of the reaction network, the unsupervised uniform sampling of the parameter space becomes exponentially time-intensive. One must consider that the curse of dimensionality and the data provided in Table 4.5 are two different things. While Table 4.5 shows the linear increase in the computational time with respect to the number of samples for the same small-scale biochemical reaction network, curse of dimensionality points into the exponential increase in the number of required samples as the size of the network increases.

5. USING KINESCOPE FOR KINETIC MODELING OF THE CENTRAL CARBON METABOLISM OF *E. COLI* AND FUTURE WORKS

Constraint-based stoichiometric models of metabolic networks do not deal with the concentrations of the metabolites. To make it feasible for these models to have a steady-state solution, pseudo reactions are added to them to supply the uptake molecules (such as extracellular glucose and oxygen) that otherwise would only be consumed by the uptake reactions without any source to replenish them. Such pseudo reactions are introduced to the stoichiometric models in the form of Equation 5.1 or 4.2, the standard usually being the latter with a negative flux value for the uptake molecules and a positive value for the secreted ones.

$$\rightarrow \text{"}extracellular\ molecule\text{"} \tag{5.1}$$

$$\text{"}extracellular\ molecule\text{"} \leftrightarrow \tag{5.2}$$

However, one can hardly find any organism operating at a steady state in natural ecosystems. They may be represented better as discreet fed-batch systems. Nevertheless, many processes in the biochemical and biotechnology industries are carried out at steady-state conditions by using continuous bioreactors. A significant amount of experimental protocols and techniques are also designed based on the steady-state conditions. Continuous stirred tank bioreactors (CSTBRs) are frequently used in laboratories to culture cells at controlled steady-state conditions (chemostat and turbidostat cultures). Figure 5.1 is a schematic representation of a CSTBR.

When making a kinetic model out of a stoichiometric model, the source reactions for the extracellular substrates (e.g. glucose) and the sink reactions for the secreted molecules (e.g. ethanol), and in general reactions in the form of Equation 5.1 or 4.2, must be removed from the model and replaced by the balance equations that consider the input/output of such molecules to/from the culture media through influxes and outfluxes. The general balance equation for such extracellular molecules can be written in the form of the following equation:

$$\frac{dC_i}{dt} = \frac{1}{V_{rxn}}\left(C_{i,feed}Q_{influx} - C_iQ_{outflux}\right) + \sum_j v_{ij}R_j \tag{5.3}$$

C_i is the extracellular concentration of the corresponding molecule in the reaction solution, $C_{i,feed}$ is the concentration of that molecule in the stream fed to the media with the volumetric flowrate Q_{influx}. V_{rxn} is the volume of the reaction solution, v_{ij} is the stoichiometric coefficient of the corresponding extracellular molecule in the reaction j, while R_j stands for any reaction that uptakes or exerts that molecule, and $Q_{outflux}$ is the volumetric flowrate of the stream leaving the media. Equation 5.3 is written based on the assumption of a homogenous reaction solution.

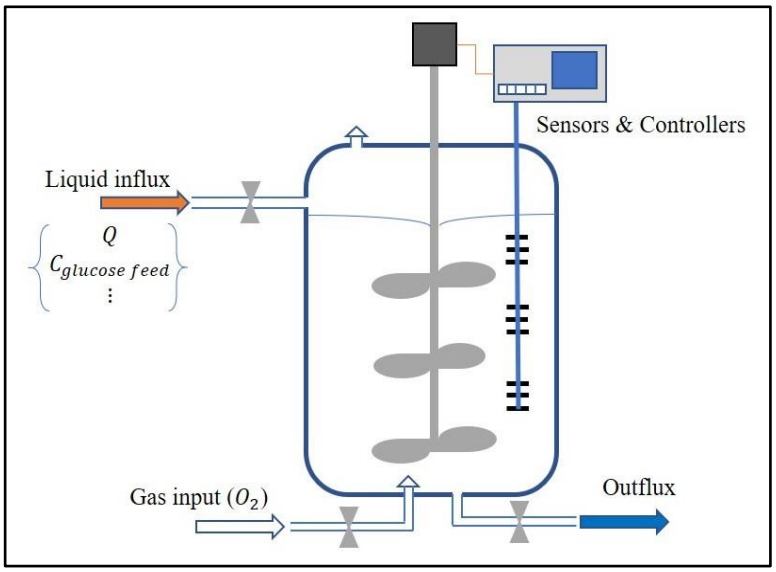


Figure 5.1: Schematic representation of a continuous stirred tank bioreactor.

5.1. Using Kinescope for Kinetic Modeling of the Central Carbon Metabolism of *E. coli* at the elementary reactions level

Kinescope was used to construct and collect kinetic models at the elementary reactions level for the central carbon metabolism of *E. coli*. The stoichiometric model was taken from the online metabolic model database of University of California San Diego [155], including 72 metabolites and 95 reactions. The reactions that could not carry any flux both at aerobic and anaerobic conditions were detected (using flux

balance and flux variability analysis) and removed from the model. Dead end molecules were also detected and removed from the model. The reduced model contains 61 reactions and 59 metabolites (provided in Appendix A). The measured metabolic flux values by ^{13}C labeling experiments for the wild type strain in a study [156] were used to constrain the corresponding reactions while using parsimonious flux balance analysis method to calculate the reference steady-state flux distribution for all reactions. Metabolic fluxes were calculated for a partially aerobic glucose limited condition (maximum possible rate of oxygen uptake and glucose uptake were constrained to 10 mmol/gDW/h) so that all the reactions in the model carry flux (the minimum being 0.1% of the glucose uptake rate). The biomass production reaction (Equation 5.4) was broken down to its constituent reactions according to Table 5.1.

$$1.496 * 3pg + 3.7478 * accoa + 59.81 * atp + 0.361 * e4p + 0.0709 * f6p + 0.129 * g3p + 0.205 * g6p + 0.2557 * gln + 4.9414 * glu + 59.81 * h_2o + 3.547 * nad + 13.028 * nadph + 1.7867\ oaa + 0.5191 * pep + 2.8328 * pyr + 0.8977 * r5p \rightarrow 59.81 * adp + 4.1182 * akg + 3.7478 * coa + 59.81 * h + 3.547 * nadh + 13.028 * nadp + 59.81 * pi \quad (5.4)$$

The reference steady-state fluxes for the biomass related reactions (e.g. depletion of biomass precursors such as $g6p$ into biomass synthesis) were calculated by using the corresponding stoichiometric coefficients multiplied by the reference steady-state flux value for the biomass reaction (calculated from flux balance analysis). This new stoichiometric model (provided in Appendix A), containing 71 reactions and 59 metabolites, and the corresponding flux distribution vector were used as inputs to the Kinescope for automatic construction of the kinetic models.

Thirty thousand different kinetic parameter vectors were sampled from the kinetic space, out of which none of them could lead to a mathematically stable model. Such an outcome was expected because of the previous experience with the ToyModel at the elementary reactions level (Section 4.2.2). The ToyModel at the elementary reactions level had 26 eigenvalues, and 71 thousand stable models were collected out of 12 million samples (less than 1% of sampled parameter vectors led to stable models). There are 390 eigenvalues for the kinetic model of the central carbon metabolism at the elementary reactions level (15-fold increase in the number of

eigenvalues), and the probability of sampling parameter vectors that lead to a Jacobian matrix with non-positive real parts in all of its 390 eigenvalues is very low.

Table 5.1: Breaking down the biomass production reaction to its constituent reactions.

Reaction tag	Formula	Steady state rate
ToBiomass_atp	$atp + h_2o \rightarrow adp + h + pi$	$59.81 \times r_{biomass}$
ToBiomass_accoa	$accoa \rightarrow coa$	$3.7478 \times r_{biomass}$
ToBiomass_nad	$nad \rightarrow nadh$	$3.547 \times r_{biomass}$
ToBiomass_nadph	$nadph \rightarrow nadp$	$13.028 \times r_{biomass}$
ToBiomass_3pg	$3pg \rightarrow$	$1.496 \times r_{biomass}$
ToBiomass_e4p	$e4p \rightarrow$	$0.361 \times r_{biomass}$
ToBiomass_f6p	$f6p \rightarrow$	$0.0709 \times r_{biomass}$
ToBiomass_g3p	$g3p \rightarrow$	$0.129 \times r_{biomass}$
ToBiomass_g6p	$g6p \rightarrow$	$0.205 \times r_{biomass}$
ToBiomass_gln	$gln \rightarrow$	$0.2557 \times r_{biomass}$
ToBiomass_glu	$glu \rightarrow$	$4.9414 \times r_{biomass}$
ToBiomass_oaa	$oaa \rightarrow$	$1.7867 \times r_{biomass}$
ToBiomass_pep	$pep \rightarrow$	$0.5191 \times r_{biomass}$
ToBiomass_pyr	$pyr \rightarrow$	$2.8328 \times r_{biomass}$
ToBiomass_r5p	$r5p \rightarrow$	$0.8977 \times r_{biomass}$
FromBiomass_akg	$\rightarrow akg$	$4.1182 \times r_{biomass}$

It is probable to collect one or a few stable models if the number of samples is increased significantly, however this won't be acceptable as a general solution to this issue. Constructing an initial ensemble of stable models with diverse kinetic parameter vectors that all satisfy a reference steady-state experiment is a fundamental step in this modeling approach. As it was also mentioned in the previous chapter, for efficient sampling of the kinetic parameter space, an algorithm is required to be developed that directly samples stable models only while maximizing the diversity in the sampled parameter vectors as well.

Although none of the sampled models were mathematically stable based on the eigenvalues of their Jacobian, it was observed that more than eight thousand models

could reach a steady-state based on the following condition defined to detect a steady-state convergence:

"After passage of 100 simulation time points, the concentration of each molecule at each time point is compared with its corresponding value at all previous 100 time points. The absolute relative changes are calculated and summed up for each molecule. If the maximum of summation of relative changes (among all molecules in the model) is less than 0.1%, it is considered that the corresponding model has reached a steady-state."

The models that could reach a steady state based on the above condition were further investigated by simulating them over a much larger time interval. It was observed that many of these models suddenly start to diverge at some point even though they display an apparent steady state for a significantly large interval (Figure 5.2). However, there are also models that keep their apparent steady state even when simulated over very large time intervals. To further investigate these models, an impulse was introduced to the concentration of the extracellular glucose (a sudden increase in the concentration of the glucose in the feed stream during a very short time interval, e.g. injection of a highly concentrated glucose solution to the feed stream in a second), and they were simulated in response to the glucose impulse. Interestingly, many of these models could converge back towards their apparent steady state (Figure 5.3), presenting an apparent stability regarding the changes in the extracellular glucose concentration. At this point, the reliability of such models and the reason behind their apparent stability even though their Jacobian tells different remain as open questions.

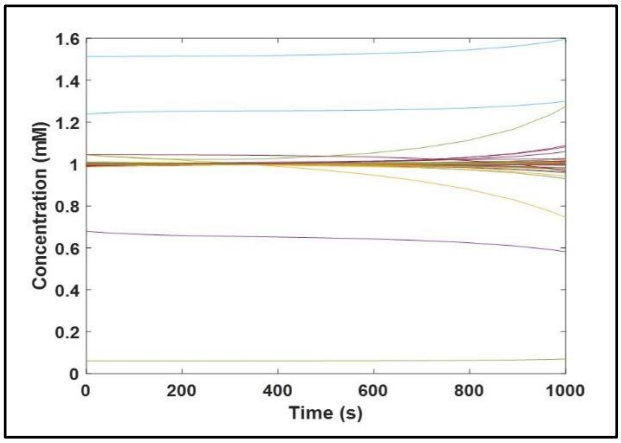


Figure 5.2: Unstable models with an apparent steady-state may start to diverge when simulated for a longer time.

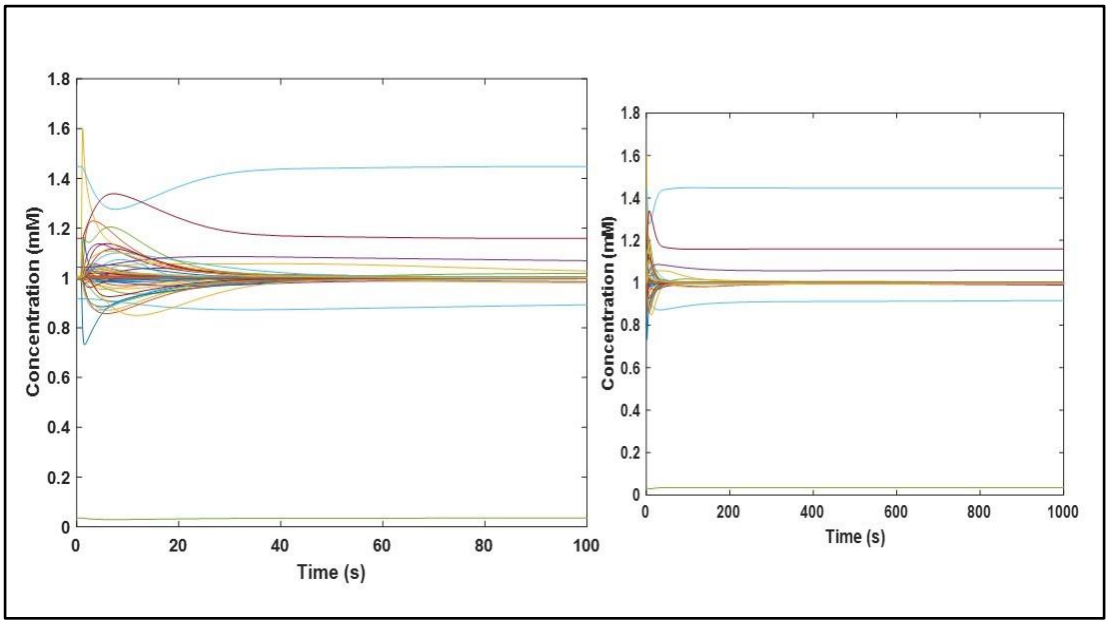


Figure 5.3: Response of one of the models with apparent stability to an impulse to the concentration of the extracellular glucose.

5.2. Future Works

i) Development of an algorithm for efficient unsupervised sampling of the kinetic parameter space, so that each sampled parameter vector is guaranteed to make the kinetic model being mathematically stable at the reference steady-state, while at the same time the diversity in the sampled parameter vectors is maximized. Such an algorithm can significantly improve the quality of the initial ensemble of the kinetic models that satisfy the reference steady-state experiment.

ii) Using machine learning methods for unsupervised clustering of the sampled kinetic parameter vectors based on the comparison of the simulation results with the corresponding experimental observations, hence recognizing those patterns in the parameter vectors that make the kinetic models being in acceptable agreement with different experimental observations. Such information/data can be used for supervised resampling of the kinetic parameter space, increasing the chance of collecting more kinetic models that are in better agreement with all experimental observations.

iii) Design and construction of a wet lab device for determination of the elementary reaction steps related to the function of a given enzyme. The input material to the

device would be the free enzyme, the known reactants and products of the reactions catalyzed by the corresponding enzyme, the known regulators of the corresponding enzyme, and any other molecule that is hypothesized to associate with the free enzyme or any of its complex forms. The output would be a table of all possible elementary reactions (associations and dissociations) related to the corresponding enzyme, leading to the generation of the elementary reaction series for both catalytic and regulatory interactions. Such information can be archived for each enzyme and used by Kinescope, hence minimizing the reliance of the Kinescope on the assumptions used for the automatic construction of the kinetic models at the elementary reactions level. The association and dissociation rate constants can be different even for the same enzyme in the same organism but at different physiological conditions. However, the elementary reaction steps and mechanism are usually the same even for the enzymes' homologues and orthologues among different organisms. As a result, construction of a database containing the above-mentioned information archived for each enzyme can be of a great value.

iv) Linking Kinescope to a database of the standard Gibbs free energy changes of reactions will make it much easier to benefit from the thermodynamic constraints while sampling the reversibility parameters (Section 4.1.6).

v) Development of an algorithm for automatic generation of hypotheses whenever the collected kinetic models are not in acceptable agreement with an experimental observation. Such an algorithm will recommend a set of rank ordered changes (e.g. addition of a reaction or regulatory interaction to the model, see Section 4.1.10) that might improve the agreement between the model simulations and all available experimental observations. Such changes can be made to the models, and, if there is a significant positive impact on the simulation results, they are considered as the potential *in-silico* discoveries, and one may try to validate them experimentally. Development of such an algorithm would make it much easier for a researcher to follow the modeling cycle represented in Figure 4.2.

6. CONCLUSIONS

Living things are complex systems. Although multicellular organisms and communities are more complex compared to the unicellular ones, the major complexity in studying the living things comes from the intracellular molecular interactions, which is something shared among all the organisms regardless of their taxonomy. A biological cell is the smallest unit of life. All the cells discovered on the earth until now use the same alphabet (A, C, G, T) to store the information in their library, which is DNA. They also use the same alphabet (22 amino acids) to construct the proteins. The basic machinery and mechanisms that they use to read the information from DNA until proteins are constructed are also very similar. However, there are too many molecules packed in a single cell, and the interaction possibilities are vast, leading to an extremely flexible system with too many possible outcomes. To study such complex systems, mathematics and computers are needed. In this research, the focus is on the analytical study of the metabolic networks and the construction of computational tools that can help to model, study and analyze cellular metabolism. Two independent computational tools, JacLy (Chapter 3) and Kinescope (Chapter 4), were developed in MATLAB during this research.

JacLy is a network inference algorithm with specific focus on the inference of small-scale metabolic networks from steady-state data. Thanks to the improvements introduced to the network inference algorithm, results reported in the previous work [15] could be obtained much faster, with much higher reproducibility, and with a higher prediction power. In addition, by applying the approach to the *in silico* metabolome data, it was shown that the use of standard deviation of replicates is a suitable approximation for the fluctuation matrix as one of the inputs to the algorithm. However, there might be more sophisticated ways of estimating a fluctuation matrix that better represent the nature of stochasticity in cellular metabolism. Finding more relevant fluctuation matrices for different biological networks can be an altogether separate research topic and can lead to an increase in accuracy and applicability of Lyapunov based inference methods such as JacLy. Also, the power of JacLy was especially obvious when a considerably lower number of replicates were used, or when a small portion of non-existent edges were introduced as prior knowledge. Prediction of the Jacobian matrix from steady-state data is another power of JacLy over GGM since Jacobian matrix is much more informative and biologically relevant in terms of

the network structure. In addition, albeit its remarkably better performance for lower number of replicates compared to a correlation-based inference as shown in this work, the use of JacLy for datasets with lower than 100 replicates should be cautioned.

Kinescope was introduced in Chapter 4 as a tool to ease the construction of kinetic models of biochemical reaction networks by following a standard and clear methodology. The main driving force behind development of Kinescope is to make a rational link between the theoretical and experimental playgrounds in the area of kinetic modeling of metabolism so that any researcher can unambiguously follow the modeling cycle of Figure 4.2. However, a major obstacle in this methodology is the uncertainty in the mechanism of action of the less studied enzymes at the elementary reactions level. Design and construction of a device for experimental determination of the elementary reaction steps of enzymatic reactions will help to overcome this obstacle to a great extent and is a major step that must be taken in the experimental playground, towards development of the standard protocol for kinetic modeling of metabolic pathways. The experimentally determined elementary steps for each enzymatic reaction can be archived in the form of a database. As it was mentioned in Section 5.2, development of such a database would be of a great value. Another step that would greatly contribute to the aforementioned standard protocol is development of an algorithm for automatic generation of hypotheses in cases when none of the model simulations is in agreement with experimental observations. A reasonable initial point for the development of such an algorithm can be careful examination of the Jacobian matrices, both the structure (zero and nonzero elements) and the interaction magnitudes, and comparing the Jacobian matrices of the relatively more successful models in simulating the system behavior with the rest of the Jacobian matrices.

All in all, kinetic modeling of the intracellular reactions must be looked from a new perspective, both at the experimental and theoretical playgrounds, and development of Kinescope has been initiated by such motivation to provide a suitable computational and analytical platform. It needs constant improvement until the modeling cycle of Figure 4.2 can be unambiguously followed.

REFERENCES

[1] Altman T., Travers M., Kothari A., Caspi R., Karp P. D., (2013), “A systematic comparison of the MetaCyc and KEGG pathway databases”, BMC Bioinformatics, 14, 112.

[2] Oberhardt M. A., Palsson B. Ø., Papin J. A., (2009), “Applications of genome-scale metabolic reconstructions”, Mol Syst Biol, 5, 320.

[3] Kim T. Y., Sohn S. B., Kim Y. B., Kim W. J., Lee S. Y., (2012), “Recent advances in reconstruction and applications of genome-scale metabolic models”, Curr Opin Biotechnol, 23(4), 617–623.

[4] Orth J. D., Thiele I., Palsson B. Ø., (2010), “What is flux balance analysis?”, Nat Biotechnol, 28(3), 245–248.

[5] Bruggeman F. J., Westerhoff H. V., (2007), “The nature of systems biology”, Trends Microbiol, 15(1), 45–50.

[6] Petranovic D., Nielsen J., (2008), “Can yeast systems biology contribute to the understanding of human disease?”, Trends Biotechnol, 26(11), 584–590.

[7] Kell D. B., (2004), “Metabolomics and systems biology: making sense of the soup”, Curr Opin Microbiol, 7(3), 296–307.

[8] Dunn W. B., Bailey N. J. C., Johnson H. E., (2005), “Measuring the metabolome: current analytical technologies”, The Analyst, 130(5), 606–625.

[9] Devantier R., Scheithauer B., Villas-Bôas S. G., Pedersen S., Olsson L., (2005), “Metabolite profiling for analysis of yeast stress response during very high gravity ethanol fermentations”, Biotechnol Bioeng, 90(6), 703–714.

[10] Psychogios N., Hau D. D., Peng J., Guo A. C., Mandal R., Bouatra S., Sinelnikov I., Krishnamurthy R., Eisner R., Gautam B., Young N., Xia J., Knox C., Dong E., Huang P., Hollander Z., Pedersen T. L., Smith S. R., Bamforth F., Greiner R., McManus B., Newman J. W., Goodfriend T., Wishart D. S., (2011), “The human serum metabolome”, PloS One, 6(2), e16957.

[11] Srividhya J., Crampin E. J., McSharry P. E., Schnell S., (2007), “Reconstructing biochemical pathways from time course data”, Proteomics, 7(6), 828–838.

[12] Çakır T., Hendriks M. M. W. B., Westerhuis J. A., Smilde A. K., (2009), “Metabolic network discovery through reverse engineering of metabolome data”, Metabolomics, 5(3), 318–329.

[13] Weckwerth W., Loureiro M. E., Wenzel K., Fiehn O., (2004), “Differential metabolic networks unravel the effects of silent plant phenotypes”, Proc Natl Acad Sci, 101(20), 7809–7814.

[14] Hendrickx D. M., Hendriks M. M. W. B., Eilers P. H. C., Smilde A. K., Hoefsloot H. C. J., (2011), “Reverse engineering of metabolic networks, a critical assessment”, Mol Biosyst, 7(2), 511–520.

[15] Öksüz M., Sadıkoğlu H., Çakır T., (2013), “Sparsity as Cellular Objective to Infer Directed Metabolic Networks from Steady-State Metabolome Data: A Theoretical Analysis”, PLOS ONE, 8(12), e84505.

[16] Kanehisa M., Goto S., Sato Y., Kawashima M., Furumichi M., Tanabe M., (2014), “Data, information, knowledge and principle: back to metabolism in KEGG”, Nucleic Acids Res, 42(Database issue), D199–D205.

[17] Caspi R., Altman T., Billington R., Dreher K., Foerster H., Fulcher C. A., Holland T. A., Keseler I. M., Kothari A., Kubo A., Krummenacker M., Latendresse M., Mueller L. A., Ong Q., Paley S., Subhraveti P., Weaver D. S., Weerasinghe D., Zhang P., Karp P. D., (2014), “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”, Nucleic Acids Res, 42(Database issue), D459-471.

[18] Croft D., Mundo A. F., Haw R., Milacic M., Weiser J., Wu G., Caudy M., Garapati P., Gillespie M., Kamdar M. R., Jassal B., Jupe S., Matthews L., May B., Palatnik S., Rothfels K., Shamovsky V., Song H., Williams M., Birney E., Hermjakob H., Stein L., D’Eustachio P., (2014), “The Reactome pathway knowledgebase”, Nucleic Acids Res, 42(Database issue), D472-477.

[19] Alcántara R., Axelsen K. B., Morgat A., Belda E., Coudert E., Bridge A., Cao H., de Matos P., Ennis M., Turner S., Owen G., Bougueleret L., Xenarios I., Steinbeck C., (2012), “Rhea--a manually curated resource of biochemical reactions”, Nucleic Acids Res, 40(Database issue), D754-760.

[20] Kumar A., Suthers P. F., Maranas C. D., (2012), “MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases”, BMC Bioinformatics, 13(1), 6.

[21] Thiele I., Palsson B. Ø., (2010), “A protocol for generating a high-quality genome-scale metabolic reconstruction”, Nat Protoc, 5(1), 93–121.

[22] Blazier A. S., Papin J. A., (2012), “Integration of expression data in genome-scale metabolic network reconstructions”, Front Physiol, 3, 299.

[23] Wiechert W., Möllney M., Petersen S., de Graaf A. A., (2001), “A universal framework for 13C metabolic flux analysis”, Metab Eng, 3(3), 265–283.

[24] Sauer U., (2006), “Metabolic networks in motion: 13C-based flux analysis”, Mol Syst Biol, 2, 62.

[25] Mueller D., Heinzle E., (2013), “Stable isotope-assisted metabolomics to detect metabolic flux changes in mammalian cell cultures”, Curr Opin Biotechnol, 24(1), 54–59.

[26] Antoniewicz M. R., Kelleher J. K., Stephanopoulos G., (2007), “Elementary Metabolite Units (EMU): a novel framework for modeling isotopic distributions”, Metab Eng, 9(1), 68–86.

[27] Zamboni N., Fendt S.-M., Rühl M., Sauer U., (2009), “(13)C-based metabolic flux analysis”, Nat Protoc, 4(6), 878–892.

[28] Schuetz R., Kuepfer L., Sauer U., (2007), “Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli”, Mol Syst Biol, 3.

[29] Tarlak F., Sadıkoğlu H., Çakır T., (2014), “The role of flexibility and optimality in the prediction of intracellular fluxes of microbial central carbon metabolism”, Mol Biosyst, 10(9), 2459–2465.

[30] Zamboni N., Fischer E., Sauer U., (2005), “FiatFlux--a software for metabolic flux analysis from 13C-glucose experiments”, BMC Bioinformatics, 6, 209.

[31] Quek L.-E., Wittmann C., Nielsen L. K., Krömer J. O., (2009), “OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis”, Microb Cell Factories, 8, 25.

[32] Weitzel M., Nöh K., Dalman T., Niedenführ S., Stute B., Wiechert W., (2013), “13CFLUX2--high-performance software suite for (13)C-metabolic flux analysis”, Bioinforma Oxf Engl, 29(1), 143–145.

[33] Schaub J., Mauch K., Reuss M., (2008), “Metabolic flux analysis in Escherichia coli by integrating isotopic dynamic and isotopic stationary 13C labeling data”, Biotechnol Bioeng, 99(5), 1170–1185.

[34] Young J. D., Walther J. L., Antoniewicz M. R., Yoo H., Stephanopoulos G., (2008), “An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis”, Biotechnol Bioeng, 99(3), 686–699.

[35] Wiechert W., Nöh K., (2013), “Isotopically non-stationary metabolic flux analysis: complex yet highly informative”, Curr Opin Biotechnol, 24(6), 979–986.

[36] van Winden W. A., van Dam J. C., Ras C., Kleijn R. J., Vinke J. L., van Gulik W. M., Heijnen J. J., (2005), “Metabolic-flux analysis of Saccharomyces cerevisiae CEN.PK113-7D based on mass isotopomer measurements of (13)C-labeled primary metabolites”, FEMS Yeast Res, 5(6–7), 559–568.

[37] Toya Y., Ishii N., Hirasawa T., Naba M., Hirai K., Sugawara K., Igarashi S., Shimizu K., Tomita M., Soga T., (2007), “Direct measurement of isotopomer of intracellular metabolites using capillary electrophoresis time-of-flight mass spectrometry for efficient metabolic flux analysis”, J Chromatogr A, 1159(1–2), 134–141.

[38] Millard P., Massou S., Wittmann C., Portais J.-C., Létisse F., (2014), “Sampling of intracellular metabolites for stationary and non-stationary (13)C metabolic flux analysis in Escherichia coli”, Anal Biochem, 465, 38–49.

[39] Mahadevan R., Schilling C. H., (2003), “The effects of alternate optimal solutions in constraint-based genome-scale metabolic models”, Metab Eng, 5(4), 264–276.

[40] Müller A. C., Bockmayr A., (2013), “Fast thermodynamically constrained flux variability analysis”, Bioinforma Oxf Engl, 29(7), 903–909.

[41] Varma A., Palsson B. O., (1994), “Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.”, Appl Environ Microbiol, 60(10), 3724–3731.

[42] Feist A. M., Palsson B. O., (2010), “The biomass objective function”, Curr Opin Microbiol, 13(3), 344–349.

[43] Lewis N. E., Hixson K. K., Conrad T. M., Lerman J. A., Charusanti P., Polpitiya A. D., Adkins J. N., Schramm G., Purvine S. O., Lopez-Ferrer D., Weitz K. K., Eils R., König R., Smith R. D., Palsson B. Ø., (2010), “Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models”, Mol Syst Biol, 6, 390.

[44] Lewis N. E., Nagarajan H., Palsson B. O., (2012), “Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods”, Nat Rev Microbiol, 10(4), 291–305.

[45] Akesson M., Förster J., Nielsen J., (2004), “Integration of gene expression data into genome-scale metabolic models”, Metab Eng, 6(4), 285–293.

[46] Becker S. A., Palsson B. O., (2008), “Context-Specific Metabolic Networks Are Consistent with Experiments”, PLOS Comput Biol, 4(5), e1000082.

[47] Shlomi T., Cabili M. N., Herrgård M. J., Palsson B. Ø., Ruppin E., (2008), “Network-based prediction of human tissue-specific metabolism”, Nat Biotechnol, 26(9), 1003–1010.

[48] Lee D., Smallbone K., Dunn W. B., Murabito E., Winder C. L., Kell D. B., Mendes P., Swainston N., (2012), “Improving metabolic flux predictions using absolute gene expression data”, BMC Syst Biol, 6(1), 73.

[49] Machado D., Herrgård M., (2014), “Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism”, PLOS Comput Biol, 10(4), e1003580.

[50] Bordbar A., Mo M. L., Nakayasu E. S., Schrimpe-Rutledge A. C., Kim Y.-M., Metz T. O., Jones M. B., Frank B. C., Smith R. D., Peterson S. N., Hyduke D. R., Adkins J. N., Palsson B. O., (2012), “Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation”, Mol Syst Biol, 8, 558.

[51] Agren R., Bordel S., Mardinoglu A., Pornputtapong N., Nookaew I., Nielsen J., (2012), “Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT”, PLOS Comput Biol, 8(5), e1002518.

[52] 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.

[53] Hoppe A., (2012), “What mRNA Abundances Can Tell us about Metabolism”, Metabolites, 2(3), 614–631.

[54] Jensen P. A., Papin J. A., (2011), “Functional integration of a metabolic network model and expression data without arbitrary thresholding”, Bioinforma Oxf Engl, 27(4), 541–547.

[55] Navid A., Almaas E., (2012), “Genome-level transcription data of Yersinia pestis analyzed with a New metabolic constraint-based approach”, BMC Syst Biol, 6(1), 150.

[56] Daran-Lapujade P., Rossell S., van Gulik W. M., Luttik M. A. H., de Groot M. J. L., Slijper M., Heck A. J. R., Daran J.-M., de Winde J. H., Westerhoff H. V., Pronk J. T., Bakker B. M., (2007), “The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels”, Proc Natl Acad Sci U S A, 104(40), 15753–15758.

[57] Postmus J., Canelas A. B., Bouwman J., Bakker B. M., van Gulik W., de Mattos M. J. T., Brul S., Smits G. J., (2008), “Quantitative analysis of the high temperature-induced glycolytic flux increase in Saccharomyces cerevisiae reveals dominant metabolic regulation”, J Biol Chem, 283(35), 23524–23532.

[58] Nikerel E., Berkhout J., Hu F., Teusink B., Reinders M. J. T., Ridder D. de, (2012), “Understanding Regulation of Metabolism through Feasibility Analysis”, PLOS ONE, 7(7), e39396.

[59] Chubukov V., Uhr M., Le Chat L., Kleijn R. J., Jules M., Link H., Aymerich S., Stelling J., Sauer U., (2013), “Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis”, Mol Syst Biol, 9, 709.

[60] Henry C. S., Jankowski M. D., Broadbelt L. J., Hatzimanikatis V., (2006), “Genome-scale thermodynamic analysis of Escherichia coli metabolism”, Biophys J, 90(4), 1453–1461.

[61] Hoppe A., Hoffmann S., Holzhütter H.-G., (2007), “Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks”, BMC Syst Biol, 1, 23.

[62] Bennett B. D., Kimball E. H., Gao M., Osterhout R., Van Dien S. J., Rabinowitz J. D., (2009), “Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*”, Nat Chem Biol, 5(8), 593–599.

[63] Soh K. C., Hatzimanikatis V., (2010), “Network thermodynamics in the post-genomic era”, Curr Opin Microbiol, 13(3), 350–357.

[64] Hamilton J. J., Dwivedi V., Reed J. L., (2013), “Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models”, Biophys J, 105(2), 512–522.

[65] Cakir T., Efe C., Dikicioglu D., Hortaçsu A., Kirdar B., Oliver S. G., (2007), “Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains”, Biotechnol Prog, 23(2), 320–326.

[66] Mo M. L., Palsson B. Ø., Herrgård M. J., (2009), “Connecting extracellular metabolomic measurements to intracellular flux states in yeast”, BMC Syst Biol, 3(1), 37.

[67] Zelezniak A., Sheridan S., Patil K. R., (2014), “Contribution of Network Connectivity in Determining the Relationship between Gene Expression and Metabolite Concentration Changes”, PLOS Comput Biol, 10(4), e1003572.

[68] Teusink B., Passarge J., Reijenga C. A., Esgalhado E., van der Weijden C. C., Schepper M., Walsh M. C., Bakker B. M., van Dam K., Westerhoff H. V., Snoep J. L., (2000), “Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry”, Eur J Biochem, 267(17), 5313–5329.

[69] Chassagnole C., Noisommit-Rizzi N., Schmid J. W., Mauch K., Reuss M., (2002), “Dynamic modeling of the central carbon metabolism of Escherichia coli”, Biotechnol Bioeng, 79(1), 53–73.

[70] Link H., Christodoulou D., Sauer U., (2014), “Advancing metabolic models with kinetic information”, Curr Opin Biotechnol, 29, 8–14.

[71] Visser D., Schmid J. W., Mauch K., Reuss M., Heijnen J. J., (2004), “Optimal re-design of primary metabolism in Escherichia coli using linlog kinetics”, Metab Eng, 6(4), 378–390.

[72] Sorribas A., Hernández-Bermejo B., Vilaprinyo E., Alves R., (2007), “Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations”, Biotechnol Bioeng, 97(5), 1259–1277.

[73] Smallbone K., Simeonidis E., Swainston N., Mendes P., (2010), “Towards a genome-scale kinetic model of cellular metabolism”, BMC Syst Biol, 4(1), 6.

[74] Chakrabarti A., Miskovic L., Soh K. C., Hatzimanikatis V., (2013), “Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints”, Biotechnol J, 8(9), 1043–1057.

[75] Stanford N. J., Lubitz T., Smallbone K., Klipp E., Mendes P., Liebermeister W., (2013), “Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks”, PLOS ONE, 8(11), e79195.

[76] Steuer R., Gross T., Selbig J., Blasius B., (2006), “Structural kinetic modeling of metabolic networks”, Proc Natl Acad Sci U S A, 103(32), 11868–11873.

[77] Tran L. M., Rizk M. L., Liao J. C., (2008), “Ensemble Modeling of Metabolic Networks”, Biophys J, 95(12), 5606–5617.

[78] Khodayari A., Zomorrodi A. R., Liao J. C., Maranas C. D., (2014), “A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data”, Metab Eng, 25, 50–62.

[79] Khazaei T., McGuigan A., Mahadevan R., (2012), “Ensemble modeling of cancer metabolism”, Front Physiol, 3, 135.

[80] Crampin E. J., Schnell S., McSharry P. E., (2004), “Mathematical and computational techniques to deduce complex biochemical reaction mechanisms”, Prog Biophys Mol Biol, 86(1), 77–112.

[81] Chou I.-C., Voit E. O., (2009), “Recent developments in parameter estimation and structure identification of biochemical and genomic systems”, Math Biosci, 219(2), 57–83.

[82] Lecca P., Priami C., (2013), “Biological network inference for drug discovery”, Drug Discov Today, 18(5), 256–264.

[83] Arkin A., Ross J., (1995), “Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series”, J Phys Chem, 99(3), 970–979.

[84] Arkin A., Shen P., Ross J., (1997), “A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements”, Science, 277(5330), 1275–1279.

[85] Samoilov M., Arkin A., Ross J., (2001), “On the deduction of chemical reaction pathways from measurements of time series of concentrations”, Chaos Woodbury N, 11(1), 108–114.

[86] Lecca P., Morpurgo D., Fantaccini G., Casagrande A., Priami C., (2012), “Inferring biochemical reaction pathways: the case of the gemcitabine pharmacokinetics”, BMC Syst Biol, 6(1), 51.

[87] Villaverde A. F., Ross J., Morán F., Banga J. R., (2014), “MIDER: Network Inference with Mutual Information Distance and Entropy Reduction”, PLOS ONE, 9(5), e96732.

[88] Vance W., Arkin A., Ross J., (2002), “Determination of causal connectivities of species in reaction networks”, Proc Natl Acad Sci, 99(9), 5816–5821.

[89] Torralba A. S., Yu K., Shen P., Oefner P. J., Ross J., (2003), “Experimental test of a method for determining causal connectivities of species in reactions”, Proc Natl Acad Sci, 100(4), 1494–1498.

[90] Schmidt H., Cho K.-H., Jacobsen E. W., (2005), “Identification of small scale biochemical networks based on general type system perturbations”, FEBS J, 272(9), 2141–2151.

[91] Astola L., Groenenboom M., Gomez Roldan V., van Eeuwijk F., Hall R. D., Bovy A., Molenaar J., (2011), “Metabolic Pathway Inference from Time Series Data: A Non Iterative Approach”. In: Loog M, Wessels L, Reinders MJT, de Ridder D

(eds) Pattern Recognition in Bioinformatics. Springer Berlin Heidelberg, pp 97–108.

[92] Savageau M. A., Voit E. O., (1987), “Recasting nonlinear differential equations as S-systems: a canonical nonlinear form”, Math Biosci, 87(1), 83–115.

[93] Liu P.-K., Wang F.-S., (2008), “Inference of biochemical network models in S-system using multiobjective optimization approach”, Bioinformatics, 24(8), 1085–1092.

[94] Ando S., Sakamoto E., Iba H., (2002), “Evolutionary modeling and inference of gene network”, Inf Sci, 145(3), 237–259.

[95] Cho D.-Y., Cho K.-H., Zhang B.-T., (2006), “Identification of biochemical networks by S-tree based genetic programming”, Bioinforma Oxf Engl, 22(13), 1631–1640.

[96] Schmidt M. D., Vallabhajosyula R. R., Jenkins J. W., Hood J. E., Soni A. S., Wikswo J. P., Lipson H., (2011), “Automated refinement and inference of analytical models for metabolic networks”, Phys Biol, 8(5), 055011.

[97] Mourão M. A., Srividhya J., McSharry P. E., Crampin E. J., Schnell S., (2011), “A graphical user interface for a method to infer kinetics and network architecture (MIKANA)”, PloS One, 6(11), e27534.

[98] Gormley P., Li K., Wolkenhauer O., Irwin G. W., Du D., (2013), “Reverse Engineering of Biochemical Reaction Networks Using Co-evolution with Eng-Genes”, Cogn Comput, 5(1), 106–118.

[99] Steuer R., Kurths J., Fiehn O., Weckwerth W., (2003), “Observing and interpreting correlations in metabolomic networks”, Bioinforma Oxf Engl, 19(8), 1019–1026.

[100] Camacho D., de la Fuente A., Mendes P., (2005), “The origin of correlations in metabolomics data”, Metabolomics, 1(1), 53–63.

[101] Krumsiek J., Suhre K., Illig T., Adamski J., Theis F. J., (2011), “Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data”, BMC Syst Biol, 5(1), 21.

[102] Nemenman I., Escola G. S., Hlavacek W. S., Unkefer P. J., Unkefer C. J., Wall M. E., (2007), “Reconstruction of metabolic networks from high-throughput metabolite profiling data: in silico analysis of red blood cell metabolism”, Ann N Y Acad Sci, 1115, 102–115.

[103] Margolin A. A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Favera R. D., Califano A., (2006), “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context”, BMC Bioinformatics, 7(Suppl 1), S7.

[104] Bandaru P., Bansal M., Nemenman I., (2011), “Mass conservation and inference of metabolic networks from high-throughput mass spectrometry data”, J Comput Biol J Comput Mol Cell Biol, 18(2), 147–154.

[105] Diệp N. Q., Hoan P. T., Bảo H. T., Hùng T. Đ., Thắng P. Q., (2011), “Computational reconstruction of metabolic networks from high-throughput profiling data”, J Comput Sci Cybern, 27(1), 23–35.

[106] Kampen N. G. V., (1992), Stochastic Processes in Physics and Chemistry. Elsevier.

[107] Oshiro M., Shinto H., Tashiro Y., Miwa N., Sekiguchi T., Okamoto M., Ishizaki A., Sonomoto K., (2009), “Kinetic modeling and sensitivity analysis of xylose metabolism in Lactococcus lactis IO-1”, J Biosci Bioeng, 108(5), 376–384.

[108] Parachin N. S., Bergdahl B., van Niel E. W. J., Gorwa-Grauslund M. F., (2011), “Kinetic modelling reveals current limitations in the production of ethanol from xylose by recombinant Saccharomyces cerevisiae”, Metab Eng, 13(5), 508–517.

[109] Hoefnagel M. H. N., Starrenburg M. J. C., Martens D. E., Hugenholtz J., Kleerebezem M., Van Swam I. I., Bongers R., Westerhoff H. V., Snoep J. L., (2002), “Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis”, Microbiol Read Engl, 148(Pt 4), 1003–1013.

[110] Cintolesi A., Clomburg J. M., Rigou V., Zygourakis K., Gonzalez R., (2012), “Quantitative analysis of the fermentative metabolism of glycerol in Escherichia coli”, Biotechnol Bioeng, 109(1), 187–198.

[111] Marchisio M. A., Stelling J., (2009), “Computational design tools for synthetic biology”, Curr Opin Biotechnol, 20(4), 479–485.

[112] Torella J. P., Chait R., Kishony R., (2010), “Optimal Drug Synergy in Antimicrobial Treatments”, PLOS Comput Biol, 6(6), e1000796.

[113] Miskovic L., Tokic M., Fengos G., Hatzimanikatis V., (2015), “Rites of passage: requirements and standards for building kinetic models of metabolic phenotypes”, Curr Opin Biotechnol, 36, 146–153.

[114] Kim O. D., Rocha M., Maia P., (2018), “A Review of Dynamic Modeling Approaches and Their Application in Computational Strain Optimization for Metabolic Engineering”, Front Microbiol, 9.

[115] Schnell S., Turner T. E., (2004), “Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws”, Prog Biophys Mol Biol, 85(2), 235–260.

[116] Magnus J. B., Hollwedel D., Oldiges M., Takors R., (2006), “Monitoring and modeling of the reaction dynamics in the valine/leucine synthesis pathway in Corynebacterium glutamicum”, Biotechnol Prog, 22(4), 1071–1083.

[117] Nikerel I. E., van Winden W. A., van Gulik W. M., Heijnen J. J., (2006), “A method for estimation of elasticities in metabolic networks using steady state and dynamic metabolomics data and linlog kinetics”, BMC Bioinformatics, 7(1), 540.

[118] Heijnen J. J., Verheijen P. J. T., (2013), “Parameter identification of in vivo kinetic models: limitations and challenges”, Biotechnol J, 8(7), 768–775.

[119] Dräger A., Planatscher H., (2013), “Parameter Estimation, Metabolic Network Modeling”. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of Systems Biology. Springer New York, New York, NY, pp 1627–1631.

[120] Cornish-Bowden A., (2015), “One hundred years of Michaelis–Menten kinetics”, Perspect Sci, 4, 3–9.

[121] Henri V., (1903), Lois générales de l’action des diastases. Paris : Librairie Scientifique A. Hermann.

[122] Michaelis L., Menten M. L., (1913), “Kinetik der Invertinwirkung”, Biochem Ztg, 49, 333–369.

[123] Schomburg I., Chang A., Schomburg D., (2002), “BRENDA, enzyme data and metabolic information”, Nucleic Acids Res, 30(1), 47–49.

[124] Rios G. M., Belleville M. P., Paolucci D., Sanchez J., (2004), “Progress in enzymatic membrane reactors – a review”, J Membr Sci, 242(1), 189–196.

[125] Nguyen H. H., Lee S. H., Lee U. J., Fermin C. D., Kim M., (2019), “Immobilized Enzymes in Biosensor Applications”, Materials, 12(1).

[126] Rocchitta G., Spanu A., Babudieri S., Latte G., Madeddu G., Galleri G., Nuvoli S., Bagella P., Demartis M. I., Fiore V., Manetti R., Serra P. A., (2016), “Enzyme Biosensors for Biomedical Applications: Strategies for Safeguarding Analytical Performances in Biological Fluids”, Sensors, 16(6).

[127] Cronwright G. R., Rohwer J. M., Prior B. A., (2002), “Metabolic control analysis of glycerol synthesis in Saccharomyces cerevisiae”, Appl Environ Microbiol, 68(9), 4448–4456.

[128] Polisetty P. K., Gatzke E. P., Voit E. O., (2008), “Yield optimization of regulated metabolic systems using deterministic branch-and-reduce methods”, Biotechnol Bioeng, 99(5), 1154–1169.

[129] Tan Y., Lafontaine Rivera J. G., Contador C. A., Asenjo J. A., Liao J. C., (2011), “Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux”, Metab Eng, 13(1), 60–75.

[130] Costa R. S., Hartmann A., Gaspar P., Neves A. R., Vinga S., (2014), “An extended dynamic model of Lactococcus lactis metabolism for mannitol and 2,3-butanediol production”, Mol Biosyst, 10(3), 628–639.

[131] Contador C. A., Rizk M. L., Asenjo J. A., Liao J. C., (2009), “Ensemble modeling for strain development of l-lysine-producing Escherichia coli”, Metab Eng, 11(4), 221–233.

[132] Rizk M. L., Liao J. C., (2009), “Ensemble Modeling for Aromatic Production in Escherichia coli”, PLOS ONE, 4(9), e6903.

[133] Oh E., Lu M., Park C., Park C., Oh H. B., Lee S. Y., Lee J., (2011), “Dynamic modeling of lactic acid fermentation metabolism with Lactococcus lactis”, J Microbiol Biotechnol, 21(2), 162–169.

[134] Nishio Y., Usuda Y., Matsui K., Kurata H., (2008), “Computer-aided rational design of the phosphotransferase system for enhanced glucose uptake in Escherichia coli”, Mol Syst Biol, 4, 160.

[135] Machado D., Costa R. S., Ferreira E. C., Rocha I., Tidor B., (2012), “Exploring the gap between dynamic and constraint-based models of metabolism”, Metab Eng, 14(2), 112–119.

[136] Cotten C., Reed J. L., (2013), “Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models”, BMC Bioinformatics, 14(1), 32.

[137] Heijnen J. J., (2005), “Approximative kinetic formats used in metabolic network modeling”, Biotechnol Bioeng, 91(5), 534–545.

[138] Jamshidi N., Palsson B. Ø., (2008), “Formulating genome-scale kinetic models in the post-genome era”, Mol Syst Biol, 4, 171.

[139] Jamshidi N., Palsson B. Ø., (2010), “Mass Action Stoichiometric Simulation Models: Incorporating Kinetics and Regulation into Stoichiometric Models”, Biophys J, 98(2), 175–185.

[140] Hendriks M. M. W. B., Eeuwijk F. A. van, Jellema R. H., Westerhuis J. A., Reijmers T. H., Hoefsloot H. C. J., Smilde A. K., (2011), “Data-processing strategies for metabolomics studies”, TrAC Trends Anal Chem, 30(10), 1685–1698.

[141] Çakır T., Khatibipour M. J., (2014), “Metabolic Network Discovery by Top-Down and Bottom-Up Approaches and Paths for Reconciliation”, Front Bioeng Biotechnol, 2.

[142] Wu L., Mashego M. R., van Dam J. C., Proell A. M., Vinke J. L., Ras C., van Winden W. A., van Gulik W. M., Heijnen J. J., (2005), “Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards”, Anal Biochem, 336(2), 164–171.

[143] Kresnowati M. T. A. P., van Winden W. A., Almering M. J. H., ten Pierick A., Ras C., Knijnenburg T. A., Daran-Lapujade P., Pronk J. T., Heijnen J. J., Daran J. M., (2006), “When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation”, Mol Syst Biol, 2, 49.

[144] Scott M., (2013), Applied stochastic processes in science and engineering.

[145] Sun X., Weckwerth W., (2012), “COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data”, Metabolomics, 8(1), 81–93.

[146] Kügler P., Yang W., (2014), “Identification of alterations in the Jacobian of biochemical reaction networks from steady state covariance data at two conditions”, J Math Biol, 68(7), 1757–1783.

[147] Sun X., Länger B., Weckwerth W., (2015), “Challenges of Inversely Estimating Jacobian from Metabolomics Data”, Front Bioeng Biotechnol, 3, 188.

[148] Picchini U., (2007), SDE Toolbox: Simulation and Estimation of Stochastic Differential Equations with MATLAB.

[149] de la Fuente A., Bing N., Hoeschele I., Mendes P., (2004), “Discovery of meaningful associations in genomic data using partial correlation coefficients”, Bioinforma Oxf Engl, 20(18), 3565–3574.

[150] Weber M., Henkel S. G., Vlaic S., Guthke R., van Zoelen E. J., Driesch D., (2013), “Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0”, BMC Syst Biol, 7, 1.

[151] Linde J., Schulze S., Henkel S. G., Guthke R., (2015), “Data- and knowledge-based modeling of gene regulatory networks: an update”, EXCLI J, 14, 346–378.

[152] Briggs G. E., Haldane J. B. S., (1925), “A Note on the Kinetics of Enzyme Action”, Biochem J, 19(2), 338–339.

[153] Milo R., (2013), “What is the total number of protein molecules per cell volume? A call to rethink some published values”, Bioessays, 35(12), 1050–1055.

[154] Milo R., Phillips R., (2015), Cell Biology by the Numbers, 1 edition. Garland Science, New York, NY.

[155] “E Coli Core | Systems Biology Research Group”. http://systemsbiology.ucsd.edu/Downloads/EcoliCore. Accessed 10 Dec 2019.

[156] Long C. P., Antoniewicz M. R., (2019), “Metabolic flux responses to deletion of 20 core enzymes reveal flexibility and limits of E. coli metabolism”, Metab Eng, 55, 249–257.

BIOGRAPHY

Mohammad Jafar Khatibipour was born in Jahrom, Iran, in May 1986. He graduated from Shiraz University, Department of Chemical Engineering, in 2008 and took his master's degree in Biotechnology from University of Isfahan in 2011. He started his PhD career at the Gebze Technical University in 2012, focusing on mathematical modeling of intracellular biochemical reaction networks.