iSCHRUNK – In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks

by

Stefano Andreozzi[1,2], Ljubisa Miskovic[1,2,*] and Vassily Hatzimanikatis[1,2,*]

[1]Laboratory of Computational Systems Biotechnology (LCSB), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, CH-1015 Switzerland.

[*]Corresponding authors:

Vassily Hatzimanikatis,
Laboratory of Computational Systems Biotechnology (LCSB),
Swiss Federal Institute of Technology (EPFL),
CH-1015 Lausanne, Switzerland
Email: vassily.hatzimanikatis@epfl.ch
Phone: + 41 (0)21 693 98 70 Fax: +41 (0)21 693 98 75

Ljubisa Miskovic
Laboratory of Computational Systems Biotechnology (LCSB),
Swiss Federal Institute of Technology (EPFL),
CH-1015 Lausanne, Switzerland
Email: ljubisa.miskovic@epfl.ch
Phone: + 41 (0)21 693 98 92 Fax: +41 (0)21 693 98 75

Keywords

Large-scale kinetic models, kinetic parameters, enzyme saturations, uncertainty reduction, Monte Carlo sampling, Machine learning

List of abbreviations

ORACLE – Optimization and Risk Analysis of Complex Living Entities

TFBA – Thermodynamics-based Flux Balance Analysis

TMFA - Thermodynamics-based Metabolic Flux Analysis

GEM – GEnome-scale Model

GP – Given Property

FI – Feasibility Index

Abstract

Accurate determination of physiological states of cellular metabolism requires detailed information about metabolic fluxes, metabolite concentrations and distribution of enzyme states. Integration of fluxomics and metabolomics data, and thermodynamics-based metabolic flux analysis contribute to improved understanding of steady-state properties of metabolism. However, knowledge about kinetics and enzyme activities though essential for quantitative understanding of metabolic dynamics remains scarce and involves uncertainty. Here, we present a computational methodology that allow us to determine and quantify the kinetic parameters that correspond to a certain physiology as it is described by a given metabolic flux profile and a given metabolite concentration vector. Though we initially determine kinetic parameters that involve a high degree of uncertainty, through the use of kinetic modeling and machine learning principles we are able to obtain more accurate ranges of kinetic parameters, and hence we are able to reduce the uncertainty in the model analysis. We computed the distribution of kinetic parameters for glucose-fed *E. coli* producing 1,4-butanediol and we discovered that the observed physiological state corresponds to a narrow range of kinetic parameters of only a few enzymes, whereas the kinetic parameters of other enzymes can vary widely. Furthermore, this analysis suggests which are the enzymes that should be manipulated in order to engineer the reference state of the cell in a desired way. The proposed approach also sets up the foundations of a novel type of approaches for efficient, non-asymptotic, uniform sampling of solution spaces.

1. Introduction

Mathematical modeling and computational analysis of cellular metabolism have become indispensable tools for understanding living organisms at a system level. Genome-scale stoichiometric models (GEMs) of metabolism are now widely available for many organisms (Henry et al., 2010; Herrgard et al., 2008; Orth et al., 2011; Osterlund et al., 2013; Sohn et al., 2010; Thiele et al., 2013) and they have been used in studies of cellular physiology and metabolic engineering (Asadollahi et al., 2009; Borodina et al., 2015; Bro et al., 2006; Dash et al., 2014; King and Feist, 2014; Snitkin et al., 2008). However, these models are not suitable for predicting the responses of metabolism to changes in enzyme expression because they are lacking information about enzyme kinetics (Miskovic and Hatzimanikatis, 2010). The research community has long appreciated this limitation and there are recent intensive efforts towards large- and genome-scale kinetic models of metabolism (Bakker et al., 2010; Chakrabarti et al., 2013; Chowdhury et al., 2014; Jamshidi and Palsson, 2010; Khodayari et al., 2014; Miskovic and Hatzimanikatis, 2010; Miskovic and Hatzimanikatis, 2011; Murabito et al., 2014; Soh et al., 2012; Stanford et al., 2013; Wang et al., 2004; Wang and Hatzimanikatis, 2006a; Wang and Hatzimanikatis, 2006b). Although the methodologies for constructing consistent large-scale kinetic models are becoming available, many challenges remain to be addressed.

When constructing kinetic models we usually start with a flux and a concentration profile, and we must find enzyme kinetics (rate expressions and parameters) that are consistent with these profiles (Chakrabarti et al., 2013; Soh et al., 2012). The two main issues that hamper development of kinetic models is uncertainty associated with the available data acquired at several biological levels:

a) *Uncertainty in metabolite concentration levels and thermodynamic displacement:* The introduction of thermodynamics in the context of flux balance analysis (FBA) have resulted in important reduction of the flux solution space (Soh and Hatzimanikatis, 2010; Soh and Hatzimanikatis, 2014); the thermodynamic properties couple the directionality of the fluxes and the levels of metabolite concentrations and therefore they impose additional constraints on the space of metabolite concentrations (Soh and Hatzimanikatis, 2014; Soh et al., 2012); uncertainties in metabolite measurements and in the estimated thermodynamic properties of reactions (Gibbs free energies of reactions) can impact the conclusions about the displacement of reactions from thermodynamic equilibrium and ultimately the conclusions about the kinetic parameters of the corresponding enzymes;

b) *Uncertainty in kinetic properties of enzymes:* The lack and uncertainty of information about enzyme kinetics has been acknowledged as the single most important obstacle for developing kinetic models (Miskovic and Hatzimanikatis, 2010; Wang et al., 2004); Uncertainties of this type can be either structural, e.g. incomplete knowledge of kinetic mechanisms, or quantitative, e.g. absent or incomplete knowledge about the values of the kinetic parameters of enzymes (Miskovic and Hatzimanikatis, 2011);

Due to the complex interactions between metabolic fluxes, metabolite concentrations, thermodynamics and kinetics, uncertainties in each of these quantities propagate to the kinetic parameter space thus making a reliable direct identification of kinetic parameters a difficult task (Almquist et al., 2014). This inspired the development of new modeling frameworks that exploit the sets of additional thermodynamic and physicochemical

constraints and integrate available data coming from several levels to reduce the space of admissible parameter values (Chakrabarti et al., 2013; Jamshidi and Palsson, 2010; Miskovic and Hatzimanikatis, 2010; Miskovic and Hatzimanikatis, 2011; Soh et al., 2012; Tran et al., 2008; Wang et al., 2004; Wang and Hatzimanikatis, 2006a; Wang and Hatzimanikatis, 2006b). Some of these approaches use Monte Carlo sampling techniques to extract populations of parameter sets capable of reproducing the observed physiology (Birkenmeier et al., 2015a; Birkenmeier et al., 2015b; Chakrabarti et al., 2013; Miskovic and Hatzimanikatis, 2010; Murabito et al., 2014; Soh et al., 2012; Tran et al., 2008; Wang et al., 2004; Wang and Hatzimanikatis, 2006a; Wang and Hatzimanikatis, 2006b). However, the sheer size of the admissible space that spans through the spaces of kinetic parameters, metabolite concentrations and metabolic fluxes along with the intrinsic nonlinearities of enzyme kinetics require tailored formulations and efficient parameter estimation techniques that are scalable and that can ultimately provide a detailed description of the metabolism. In the identification of population of models using sampling methods, a highly efficient method should be able to generate a very large number of models. This is a daunting task as the size of the models and the nonlinearities in the models increase. To overcome these challenges we have developed the ORACLE framework (Optimization and Risk Analysis of Complex Living Entities) (Chakrabarti et al., 2013; Miskovic and Hatzimanikatis, 2010; Miskovic and Hatzimanikatis, 2011; Soh et al., 2012; Wang et al., 2004; Wang and Hatzimanikatis, 2006a; Wang and Hatzimanikatis, 2006b), which uses nonlinear mechanistic rate laws to model reaction kinetics and the model parameters are computed through Monte Carlo sampling, integration of partial data, and a sequence of algebraic operations. ORACLE offers a significant computational advantage over the

parameter estimation methods that: (i) require solving systems of ordinary differential equations (ODEs); and/or (ii) use dynamic optimization techniques. The efficient way of sampling of kinetic parameter space (Miskovic and Hatzimanikatis, 2010; Wang et al., 2004) and low computational requirements make ORACLE scalable and more suitable for modeling of large-scale and genome-scale kinetic networks.

In the most common studies, we have a reference steady state flux profile and a vector of metabolite concentration levels, and we want to derive the corresponding *feasible kinetic model*. We consider as a feasible kinetic model the model that is: (i) consistent with the observed metabolic fluxes and metabolite concentration levels; (ii) locally stable around the reference steady-state (Wang et al., 2004); and (iii) consistent with any additional experimental observations and available expert knowledge (Miskovic and Hatzimanikatis, 2010). Nevertheless, it is impossible to identify a unique kinetic model consistent with the observed physiology due to limited amount of available data relative to the number of model parameters. Instead, in the ORACLE framework we derive *a population* of feasible kinetic models (Figure 1). This population of models involves large uncertainties in the sense that the exact parameter values of the models are typically unknown, and the ranges of some of the parameter values are large and not well characterized. To address this issue we introduced in this manuscript a new approach for characterization and reduction of uncertainty. This approach makes use of ORACLE and machine learning classification techniques to identify the values of the enzyme saturation levels and the corresponding values of kinetic parameters that give rise to feasible kinetic models. As a result of this approach, we obtain a set of kinetic models with well-constrained parameters, i.e. the uncertainty in kinetic parameters is reduced.

The proposed methodology also introduces a new way for sampling efficiently and non-asymptotically the space of parameters and for computing the volume of this space. Indeed, this approach approximates the space of the parameters that are consistent with an observed physiology as a set of hyper-boxes, starting from an initial population of models derived through ORACLE. For each of hyper-boxes we can perform sampling and volume computation independently, and we then combine the resulting sampling sets and volumes to obtain a sampling set and a volume that characterize the whole space of parameters.

We illustrated the utility of our approach through a study of glucose-fed *E. coli* producing 1,4-butanediol, and we computed the distribution of enzyme saturations and parameters that are consistent with the observed physiology. We used the machine learning algorithm to identify the subspace of the kinetic parameter space wherein the kinetic models are likely to be feasible, and we discovered that feasible kinetic models can be constructed by constraining the saturations (and the corresponding kinetic parameters) of only 27 out of 153 enzymes within specific ranges, while the other enzymes could operate in any regime. This finding appears to be consistent with studies by Sethna and colleagues where it was shown that in many systems biology models, which they call "sloppy models", most directions in parameter space do not affect the model output but that there are a few so-called stiff directions that change the model behavior (Daniels et al., 2008; Gutenkunst et al., 2007).

2. Materials and Methods

2.1. *Parameter classification problem*

The parameter classification problem is the identification of a subspace of the parameter space wherein the parameters satisfy a given property (GP). If we consider a *n*-dimensional space of parameters, $p_1, p_2, \dots, p_n$, and we assume that the GP is satisfied if a function of these parameters, $f(p_1, p_2, \dots, p_n)$, satisfies $f(p_1, p_2, \dots, p_n) < 0$, then, for this parameter space the parameter classification problem is defined as:

Given:

(i) an ensemble of parameter sets $(p_1, p_2, \dots, p_n)$, and

(ii) the information which of these parameter sets satisfies GP,

can we find ranges of $p_1, p_2, \dots, p_n$, for which GP is satisfied without knowing the exact functional form of $f(p_1, p_2, \dots, p_n)$?

2.1.1. Classification algorithm

Decision-tree learning algorithms use values of observed data samples to infer the rules, such as parameter ranges, that predict if the data satisfy a GP (Bishop, 2006; Han et al., 2012; Quinlan, 1993). For the purpose of this work we have used the CART algorithm (Breiman et al., 1984) implemented in the MATLAB software package. We perform the classification procedure outlined in the following three steps:

Step C1: We form a *training set* of data and use it to infer a set of classification rules. A training set consists of an *input set* and an *answer set*. In this work, the input set contains the sets of parameters whereas the answer set contains the information if GP is satisfied or not for each parameter vector;

Step C2: We compute the feasibility index (FI) as the ratio between the number of parameter sets that satisfy GP and the overall number of generated parameter sets. FI is a

measure of the uncertainty in the parameter space: if all parameter sets satisfy GP, FI is equal to 1, whereas if no parameter set satisfies GP, FI equals 0.

Step C3: We generate a set of data based on the inferred rules from Step C1, and which is independent of the training set (validation set). We compute then FI over the validation set, and we compare the obtained FI score with the one from Step C2: if we obtain an improved FI, the rules from Step C1 are validated.

The CART algorithm produces a binary tree of classification rules (Figure 2, panel B). Each branch of the tree represents a rule, i.e. a sequence of conditions on parameters, $p_1, p_2, \dots, p_n$, that infers whether GP is satisfied or not (Figure 2, panel B). Each classification rule with the leaf label "y'' (yes) envelops the samples that satisfy GP (Figure 2, panel B). In the space of parameters, inferred classification rules correspond to hyper-boxes, and the bounds for each hyper-box are defined as inequalities on the individual parameters. As a set of hyper-boxes approximates the subspace of the parameter space that satisfies GP, there is interplay between the geometric complexity of the subspace defined by GP and the number of hyper-boxes needed to approximate this subspace, i.e. the more complex the shape of the subspace satisfying GP is, the more hyper-boxes are needed to approximate it.

After the computation of the rules and the associated hyper-boxes we are able to perform two very important operations on the parameter space. First, we are able to compute in an efficient way the *n*-dimensional volume of the subspace of kinetic parameters that satisfy GP by summing up the *n*-dimensional volumes of the individual hyper-boxes. Second, due to the nature of the parameter subspaces (hyper-boxes), we can efficiently sample the subspace of kinetic parameters with uniform distribution by sampling the individual hyper-boxes.

The number of rules returned by the algorithm depends on several factors such as the shape of the subspace defined by the GP and the number of samples. An important parameter of the classification algorithm is the cut-off threshold *Tk*, a minimal number of samples (parameter sets) that the algorithm can use to form a rule (Duda et al., 2001; Han et al., 2012). More precisely, the classification algorithm retains only the rules that are based on at least *Tk* samples (parameter sets), whereas all the rules that are inferred on less than *Tk* samples are discarded. Therefore, the higher *Tk* is, the fewer rules are inferred.

We use subsequently the following *toy example* to define the parameters and metrics employed in parameter classification:

> For the space of two parameters, p_1 and p_2, which are bounded in the range between 0 and 1, solve the above defined parameter classification problem with a function $f(p_1, p_2) = (p_1 - 0.1)^2 + 4(p_2 - 0.1)^2 - 0.64$. That is, find the set of ranges of p_1 and p_2 if GP is satisfied for $f(p_1, p_2) < 0$ (Figure 2, Panel A). The area of the subspace that satisfies GP is subsequently referred to as the "true area".

We applied the CART algorithm to the toy example. We generated the *input* set of 500 random samples from this space, and we formed the *answer* set by assessing whether GP was satisfied or not (Figure 2, panel A). These two sets were provided to the CART algorithm as the input. We choose the cut-off threshold to be *Tk* = 1, which means that all the rules generated by the classification algorithms are kept. The algorithm produced 5 rules on parameters p_1 and p_2 that can be used to infer whether GP is satisfied or not (Figure 2, panel B). For example, the rule deduced from the solid branch in Figure 2, panel B, with $0.768 < p_1 < 0.858$ and $p_2 < 0.271$ corresponds to the hyper-box III in panel A. This

rule implies if the value of p_1 is between 0.768 and 0.858, and if the value of p_2 is less than 0.271 then there is a "high certainty" that GP will be satisfied. The other rules satisfying GP are shown as hatched boxes I, II, IV, and V (Figure 2, panel A). With this set of classification rules, the algorithm approximated the function $(p_1 - 0.1)^2 + 4(p_2 - 0.1)^2 - 0.64 < 0$. We choose 1000 parameter sets according to the approximate rules, i.e. we formed the validation set from Step C3, and we tested if GP was satisfied. We obtained FI of 0.976, i.e. for 976 out of 1000 parameter sets (97.6%,) the algorithm correctly predicted the outcome, and there were 24 (2.4%) false positives (parameter sets which are predicted by the algorithm to satisfy GP but they do not).

2.1.2. Ranking of classification rules

For a precise determination of the frontier of separation between the samples satisfying GP and those that are not, we need a large number of samples that will result in larger number and more accurate rules. In general, we should expect that the rules inferred on the basis of only a few samples are more likely to be imprecise, i.e. they might envelop a considerable part of the parameter space where GP is not satisfied. Moreover, if we sample the parameter space uniformly, we can also expect that rules enveloping more samples approximate a larger portion of the parameter space that satisfies GP. Based on this reasoning, we rank the rules according to the number of samples they envelop – the more samples, the higher ranked the rule. The ranking allows us to identify and discard the rules that: (i) are more likely to provide erroneous classification; and (ii) approximate negligible portions of the parameter space satisfying GP. Following this ranking, we eventually end up with *simpler and more reliable* rules.

We ranked the rules inferred for the toy example, and the box corresponding to the top rule contained 147 samples whereas the one for the lowest ranked rule contained only two samples (Figure 2, panel A, boxes I – V).

Table 1: Successive application of the rules computed for the toy example (Figure 2).

Rules	Total enclosed samples	Enclosed samples of added rule	FI	Enclosed area
I	147	147	0.989	0.282
I+II	165	18	0.978	0.336
I+II+III	175	10	0.978	0.360
I+II+III+IV	180	5	0.976	0.378
I+II+III+IV+V	182	2	0.976	0.382

For the inferred rules we computed the feasibility index (FI). For a chosen set of hyper-boxes, FI represents a ratio between the parameter sets with the correct predictions of GP and the total samples contained in these boxes. For the top ranked rule (Box I) we computed FI of 0.989, i.e. if we sample within this box 98.9% of parameter sets would satisfy GP. As we successively added one-by-one rules according to the ranking starting from the top rule we observed that FI was slightly decreasing so that for all five rules FI was 0.976 (Table 1).

Since we know the functional form of $f(p_1, p_2)$ for this toy example, we were able to compute the exact value for the *true area* (0.3804). Then, for each hyper-box, we used the information about the ranges of parameters to compute its corresponding area. For example, for hyper-box I with the parameters ranging $0 < p_1 < 0.768$ and $0 < p_2 < 0.368$ we computed the area of 0.282, which approximated 74% of the *true area* (Table 1 and Figure 2). As we successively added the other rules according to the ranking, the covered area was increasing so that for all 5 rules the covered area was 0.382 (Table 1). Indeed, the covered

area was slightly larger than the *true area*, which explains the 2.4% of false positives reported in Section 2.1.1.

A comparison of the evolution of FI and the coverage areas in Table 1 suggested that there was a trade-off between the reliability of the predictions (in terms of FI) and the coverage of the space that satisfies GP. Specifically, the larger the covered space was, i.e. the larger number of rules was considered, the smaller the feasibility index was obtained.

2.2. Computational procedure for characterization of uncertainty

The computational procedure for the characterization of uncertainty is based on the ORACLE framework (Chakrabarti et al., 2013; Miskovic and Hatzimanikatis, 2010; Miskovic and Hatzimanikatis, 2011; Soh et al., 2012; Wang et al., 2004) and involves a set of successive computational procedures that allows us to consistently integrate omics data, and physicochemical and thermodynamic constraints into kinetics models of sizes scalable to genome-scale metabolic networks. The procedure involves 8 steps where Steps 1-5 stem from the original ORACLE framework, whereas Steps 6 to 8 correspond to Steps C1 to C3 of the classification procedure presented in Section 2.1.1. We outline the procedure as follows (Fig. 1):

Step 1: We define the stoichiometry and the thermodynamic constraints followed by the integration of the metabolomics, fluxomics, and physiology data and we perform the Thermodynamics-based Metabolic Flux Analysis (TMFA) (Henry et al., 2007), also called Thermodynamics-Based Flux Balance Analysis (TFBA)(Soh and Hatzimanikatis, 2010; Soh and Hatzimanikatis, 2014; Soh et al., 2012). Since the TFBA problem might have multiple optimal solutions, i.e. multiple sets of flux and concentrations vectors can explain the

observed measurements for the same value of the objective function, we choose a metabolic flux vector based on expert knowledge, a hypothesis, or we perform PCA (Jolliffe, 2002) to find a representative steady-state flux profile consistent with the observed physiology.

Step 2: We sample the space of metabolite concentrations that are thermodynamically consistent with the steady-state flux profile determined in Step 1, and we compute the displacements from the thermodynamic equilibrium of all reactions in the metabolic network.

Step 3: We integrate the available information about the kinetic mechanisms (Segel, 1975) and values of kinetic parameters from the literature and the databases (Schomburg et al., 2013; Wittig et al., 2012). For the reactions with unknown kinetic mechanisms, we use approximate rate laws such as convenience kinetics (Liebermeister and Klipp, 2006) and reversible Hill kinetics (Hofmeyr and Cornish-Bowden, 1997). For enzymes with no or incomplete information about their kinetic parameters we use Monte Carlo sampling techniques (Gentle, 2003; Gilks et al., 1998) to sample the space of kinetic properties in the form of enzyme states (Miskovic and Hatzimanikatis, 2011) or the degree of saturation of enzyme (Wang et al., 2004).

Step 4: We use the results acquired in Steps 1-3 to parameterize a population of kinetic models of metabolism of the same structure. The structure of the models can be either of the following types: nonlinear models, log-linear models (Hatzimanikatis and Bailey, 1996; Hatzimanikatis and Bailey, 1997; Wang et al., 2004; Wang and Hatzimanikatis, 2006a; Wang and Hatzimanikatis, 2006b), BST models (Savageau, 1969a; Savageau, 1969b; Savageau, 1970), etc.

Step 5: We perform the feasibility test, i.e. assuming that the observable flux and metabolite profiles we want to capture are at steady or quasi-steady state, we verify the

stability and we impose consistency of the obtained models with the experimentally observed data and literature. The feasibility test tells us if the GP of the parameter classification problem defined in Section 2.1 is satisfied.

Step 6: We form the *input set* with the parameters obtained in Step 4, and the *answer set* with the feasibility test results obtained in Step 5 (see Step C1 in Section 2.1.1). We use these two sets as a training set for the CART machine learning algorithm (Han et al., 2012) to extract the rules on the ranges of kinetic parameters that give rise to feasible kinetic models.

Step 7: We compute the feasibility index (FI), which is a measure of the uncertainty in the kinetic parameter space. FI is computed as the ratio between the number of kinetic parameter sets that passed the feasibility check in Step 5 and the overall number of generated parameter sets.

Step 8: We use the inferred rules from Step 6 to generate an independent population of the kinetic models in Steps 3 and 4. We then use this population of models as the validation set. If we observe an increased FI of the validation set compared to FI of the training set, the rules are validated.

These rules can then be used to generate new populations of kinetic models with improved certainty of being feasible, i.e. being locally stable, consistent with the studied fluxomics, metabolomics, physiology data, and consistent with the available expert knowledge and postulated hypotheses.

2.3. Ranking of classification rules, enzyme saturations and enzymes

In the ORACLE framework, instead of directly sampling the parameter space, we first sample the enzyme states, or the enzyme saturations, which are always within well-defined bounds. We then use the corresponding metabolite concentrations to compute the parameters from the saturation samples.

For the system we present here, and for all the systems we have studied, in the rules of the classification procedure (Steps 6-8) only a few of the enzyme saturations must be constrained within narrow bounds in order to derive feasible kinetic models for a given physiology, while the rest of enzyme saturations can range widely. Therefore, by narrowly constraining only a few enzyme saturations while choosing the values of the remaining enzyme saturations in a random manner we can obtain a population of models with an improved FI. We ranked the enzyme saturations and the enzymes according to the aforementioned improvement in FI.

2.3.1. Ranking of classification rules

We rank the classification rules based on the number of kinetic parameter sets that they enclose (see the discussion in Section 2.1.2). The higher the number of enclosed parameter sets, the higher ranked the rule.

2.3.2. Ranking of enzyme saturations within a rule

For a given rule, we choose an enzyme saturation and we extract its bounds within this rule. Next, we form a subspace of the parameter space that is defined within the extracted bounds, while the rest of enzyme saturations can range over all admissible values. We then evaluate FI within this subspace, i.e. we evaluate FI over all samples from the training set that satisfy the ranges of the chosen enzyme saturation.

We repeat this procedure for all enzyme saturations and we rank them according to the obtained values of FI from the highest FI towards the lowest FI.

2.3.3. Ranking of enzyme saturations over the Top 10 rules

We want to screen out the enzyme saturations which, when constrained, give the highest improvement of FI over the Top 10 rules. We perform the ranking as follows. First, for each of the Top 10 rules, we evaluate FI for all enzyme saturations as described in Section 2.3.2. Second, for each of enzyme saturations we multiply the FI value in each of the Top 10 rules by the number of samples the corresponding rule envelops and we sum the obtained values. Finally, we rank the enzyme saturations over Top 10 rules according to the obtained sums starting from the highest sum (Supplementary File 3).

2.3.4. Ranking of enzymes with a rule and over the Top 10 rules

The ranking of enzymes within a rule is performed in a similar way to that of enzyme saturations (Section 2.3.2.). We start by extracting the bounds on *all the saturations pertaining to a chosen enzyme*. We then form a parameter subspace that is constrained by the extracted bounds where the saturations of other enzymes can range over all admissible values, and we evaluate FI within this subspace. We repeat this procedure for all enzymes and we rank them according to the obtained FI values.

The ranking of enzymes over the Top 10 rules is performed analogously to the procedure presented in Section 2.3.3.

3. Results and Discussion

3.1. Characterization of feasible kinetic parameter space

We used a reduced stoichiometric model of 1,4-butanediol producing *E. coli* obtained from the genome-scale model of *E. Coli*, iJO1366 (Orth et al., 2011). The reduced model includes the core metabolic pathways glycolysis, pentose phosphate pathway, tri-carboxylic cycle (TCA) and electron transport chain (ETC) along with the engineered 1,4-butanediol production pathway. The model contains 175 intracellular reactions and mass balances for 106 metabolites in the cytosol and the extracellular space (Supplementary File 1, Figure 1 and Tables 1 and 2). We assigned kinetic mechanisms such as reversible Michaelis-Menten kinetics, Uni-Bi, ordered Bi-Bi, Bi-Ter, Ter-Bi etc., to 153 enzymatic reactions of the metabolic network (Segel, 1975) (Supplementary File 1, Table 3). If for some reactions the kinetic mechanism was unknown, we used generalized reversible Hill kinetics (Hofmeyr and Cornish-Bowden, 1997) or convenience kinetics (Liebermeister and Klipp, 2006). The obtained kinetic space, subsequently referred to as *original kinetic parameter space,* consisted of 527 enzyme saturations corresponding to 527 K_m values.

We randomly generated a set of 1000 metabolite concentration vectors (Supplementary File 2) that are thermodynamically consistent with the chosen metabolic flux (Supplementary File 1, Table 5), and we randomly picked one sample from this set (subsequently referred to as *chosen metabolite concentration vector*).

From Brenda and SABIO-RK database (Schomburg et al., 2013; Wittig et al., 2012) we extracted experimental information about 69 Michaelis constants, K_m, which corresponded to 37 enzymes in our model (Supplementary File 1, Table 4). The K_m values for every substrate and product were available only for 8 enzymes. We used the experimental K_m values to compute the bounds on enzyme saturations from the samples of metabolite concentrations. For the remaining enzymes with incomplete or no information about K_m

values we set the lower bound on enzyme saturations to 0 (non-saturation) and the upper bound to 1 (full saturation) (Wang et al., 2004). We then sampled the enzyme saturations by assigning uniformly random numbers between the assigned bounds (Wang et al., 2004). This way, we generated a set of 150000 enzyme saturations vectors, subsequently referred to as *validation set*, and for the chosen metabolite concentration vector we performed the feasibility test over the validation set. In ORACLE all generated parameter vectors are consistent with the observed metabolic fluxes and metabolite concentration levels, and therefore the results of the feasibility test depended on the local stability of models around the reference steady state. The computed FI was 0.477, i.e. 71623 (47.7%) parameter vectors out of 150000 in the validation set gave rise to feasible kinetic models.

3.2. Quantification of uncertainty in the kinetic parameter space

We generated the input set for the classification algorithm consisting of a uniformly distributed random set of 100000 enzyme saturations. We performed next the feasibility test over this input set for the chosen metabolite concentration vector, and we obtained an FI of 0.481. We then generated the answer set based on the criterion if the feasibility test was satisfied or not for each of input set saturations. With the input and the answer set we formed the training set (Materials and Methods, Sections 2.1.1 and 2.2). We choose the cut-off threshold to be $Tk = 10$, which means that the classification algorithm retained only the rules that are based on at least 10 parameter sets (Section 2.1.1). The algorithm returned 3801 classification rules on 527 enzyme saturations that lead to an improved FI, and we tested these rules over the validation set. The kinetic parameter subspace defined by these rules had an improved FI of 0.595 compared to the original kinetic parameter space, which

had an FI of 0.477 (Table 2). We ranked the rules according to the number of samples they contain (Material and Methods, section 2.1.2), discarded the ones with the lowest ranking, and tested FI of the retained rules over the validation set. Similarly to the toy model case discussed in section 2.1.2, as we consider a smaller number of more reliable rules, the incidence of correct predictions increase, so that for 10 top rules FI was 0.845 whereas for the most reliable rule it climbed up to 0.884 (Table 2). Simultaneously, the smaller the number of rules we consider, the smaller the considered kinetic parameter subspace, so that 10 top rules covered the subspace that contained 8802 samples, i.e. 5.8% of the samples in the validation set, whereas the top rule covered 2183 samples, i.e. 1.45% of the validation set (Table 2). The ranking of the rules and its successive application allowed us to *map the kinetic parameters space* according to FI. More specifically, by starting with the most dominant rule and then adding successively one-by-one the remaining rules we were able to demarcate the regions of the kinetic parameter space based on the value of FI. The evolution of FI as we added the first 50 rules according to the ranking is provided in (Supplementary File 1, Table 6).

Table 2: Feasibility index (FI) of the obtained classification rules over the validation set for the chosen metabolite concentration vector.

Number of rules	3801		50		10		1	
	FI	Number of models	FI	Number of models	FI	Number of models	FI	Number of models
Quantitative rules	0.595	42887	0.787	20250	0.845	8802	0.884	2183
Discrete rules	0.477	149995	0.527	124753	0.573	93266	0.666	26675

We next analyzed the kinetic parameter subspace defined by the top rule and we found that only 94 out of 527 enzyme saturations are constrained within narrow ranges, and the

remaining 433 enzyme saturations could take any value between 0 (linear regime) to 1 (full saturation). Interestingly, these 94 saturations corresponded to only 27 out of 153 enzymes. This striking result implies that it is sufficient to constrain only a small number of highly ranked enzyme saturations to improve FI compared to the one of the original kinetic parameter space. We ranked these 94 enzyme saturations as presented in Materials and Methods, section 2.3. For the top ranked saturation, i.e. saturation of AKGDH (2-oxoglutarate dehydrogenase) by succinate-CoA, we constrained its value between 0.594 and 1, and we obtained FI of 0.569 over the validation set (Figure 3 and Supplementary File 1, Table 7).

We continued the analysis by studying the kinetic parameter subspace defined by the top 10 rules and we found the same 94 enzyme saturations are narrowly constrained within this subspace, and they were related to the same 27 (out of 153) enzymes. We analyzed the values of the Top 10 ranked enzyme saturations (Materials and Methods, Section 2.3) that give rise to feasible kinetic models in this subspace, and we discovered the ranges of these saturations that lead to high values of FI. Specifically, we observed that when the values of the saturation of AKGDH by succinate-CoA were in the medium-to-high range, i.e. when the values of this saturation were higher than 0.5 (Figure 3), FI was increased to the value of 0.574. When this saturation had the low-to-medium values (Figure 3), the corresponding FI was deteriorated to the value of 0.418. We observed similar patterns for the saturations of PFK by FdP, GLCptspp by G6P, AKGDH by CoA and GLCptspp by PEP (Figure 3). In contrast, we observed improved values for FI (ranging from 0.517 to 0.535) whenever the saturation of AKGDH by NAD ranged in the low-to-medium range, and deteriorated values of FI, for example 0.405 for the rule 10, when this saturation was in the medium-to-high range

(Figure 3). Some of the top ranked enzyme saturations, such as AKGDH by CoA and GLCptspp by PEP, were not constrained in some of the top 10 rules (Figure 3).

Once we have derived the top rules we can further perform meta-analysis to investigate combinations of ranges of enzyme saturations within the top enzymes that would give an improved FI, and then we can use any of these combinations to form *synthetic rules.* For example, we analyzed the saturation of PFK by FdP and we obtained an FI of 0.541 when this saturation was constrained within the range defined by Rules 4 and 10, while the rest of enzyme saturations ranged over all admissible values (Figure 3). This FI was superior to the one obtained when this saturation was constrained within the range defined by Rule 1 (0.534). Similarly, for the saturation of AKGDH by NAD we found that if we constrain this saturation according to Rule 4, we would obtain the highest FI (0.535).

Based on this analysis, we constructed a synthetic rule (Figure 3). For each of the analyzed top 6 enzyme saturations we took the corresponding ranges that would give the highest FI among the top 10 rules. We tested the successive application of constraints of top 6 enzyme saturations over the validation set for the top 10 inferred rules and for the synthetic rule. The cumulative FI of the synthetic rule of 0.855 was far superior to all other rules (Figure 3). Even more striking was that a synthetic rule composed by narrow ranges of only 6 enzyme saturations was comparable in terms of FI to the top rule I which had 94 saturations constrained (0.855 for the synthetic rule versus 0.883 for the top rule I). In comparison, successive application of constraints on the top 10 ranked enzyme saturations over the top 10 rules provided a FI of 0.75 (Figure 4, panel A).

This analysis indicated that for each of the enzyme saturations there was a well-defined range that gave rise to improved FI. This analysis also suggested that the individual

improvements of FI, obtained by constraining each of the enzyme saturations to these ranges, were synergistic. That is, when we successively constrained one-by-one the ranges of enzyme saturations in the synthetic rule we obtained a monotonic increase of FI (Figure 4, panel A).

Next, we ranked the enzymes in the network as discussed in Materials and Methods, Section 2.3, and the Top 5 enzymes were AKGDH, GLCptspp, PFK, THD2pp, GLYCLTDy (Figure 4, panel B). We constrained the enzyme saturations that pertained to AKGDH, and we obtained FI of 0.613. By additionally constraining the saturations related to GLCptspp, FI increased to 0.680. Furthermore, by constraining 8 top ranked enzymes, i.e. AKGDH, GLCptspp, PFK, THD2pp, GLYCLTDy, AKGD, CS and PGI, we obtained FI of 0.818 that is close to FI of 0.845 when all enzymes are constrained (Figure 4, panel B). This analysis identifies the enzymes whose kinetics must be determined to characterize precisely the analyzed feasible kinetic subspace.

3.3. Semi-Quantitative uncertainty characterization in the kinetic parameter space

Available information about enzymes can be available in a "semi-quantitative" form. For example, if we consider enzyme saturations, then this information is sometimes communicated as follows: an enzyme operates in (i) low saturation (linear regime); (ii) medium saturation; and (iii) high saturation. Therefore, we proposed the following two procedures that allowed us to analyze semi-quantitative data. In the first procedure we first discretize the parameter space and we then build the rules, while in the second procedure we first build the quantitative rules and then we discretize them to construct the discrete rules.

3.3.1. Integration of semi-quantitative information into the kinetic models

We first split the space of saturation in three *discrete* intervals: (i) interval of low saturation – the quantitative values of saturation range between 0 and 0.25; (ii) interval of medium saturation – the quantitative values of saturation range between 0.25 and 0.75; and (iii) interval of high saturation – the quantitative values of saturation are larger than 0.75. Then, for each provided semi-quantitative information we assign one of the three intervals (or their combination) and we sample within such interval. Finally the samples are provided to the machine classification algorithm as inputs.

3.3.2. Semi-quantitative characterization of the subspaces with reduced uncertainty

For each rule, we consider constraints on enzyme saturations and for each computed range of the saturations we compute the discretized interval (or their combination) that encloses this range. For example, if in a classification rule enzyme saturation ranges from 0.12 to 0.64, then in the discrete rule this saturation will cover interval of low and medium saturation described in Section 3.3.1, i.e. it will range between 0 and 0.75.

We discretized the set of quantitative rules obtained in Section 3.2 and then we tested FI over the validation set. The discrete rules provided improved FI (ranging from 0.477 to 0.666) compared to the original kinetic parameter space (0.477), but this improvement was inferior to the one when the quantitative rules are applied (with FI ranging from 0.595 to 0.884) (Table 2). This was expected, as the discrete rules defined a parameter subspace that enclosed the portions of the parameter space that had a lower FI compared to the FI of the space defined by the quantitative rules. For the top discrete rule we obtained FI of 0.666, which was inferior to 50 top quantitative rules with FI of 0.787 (Table 2). Simultaneously, the top discrete rule enveloped a bigger portion of the original kinetic

parameter space (26675 samples, i.e. 17.8% of the samples in the validation set) than 50 top quantitative rules (20250 samples, i.e. 13.5% of the samples in the validation set). The discrete rules were so approximate that 3801 discrete rules enveloped practically the whole original kinetic parameter space (Table 2). In comparison, the quantitative rules demarcate very precisely the portions of the kinetic parameter space with high and low FI.

3.3. Robustness of the classification rules

We considered a set of 1000 metabolite concentration vectors and we analyzed the feasibility of kinetic models for each concentration vector over the validation set consisting of 150000 parameter sets (see section 3.1). The computed FI ranged from 0.239, i.e. 23.9% of the samples from the validation set formed a feasible kinetic model with the analyzed concentration vector, to 0.851 with a symmetrical distribution around the mean value of 0.531 (Figure 5). The distribution of the FI over the metabolite concentration levels suggested that these quantities are important factor in forming the feasible kinetic space. This observation is consistent with previous studies by Chakrabarti et al., where it was reported that the stability of the kinetic models was affected predominantly by metabolite concentration levels and enzyme saturations and to a lesser extent by metabolic flux levels (Chakrabarti et al., 2013).

We then tested if the rules obtained in Section 3.2 for the chosen metabolite concentration vector will also give high FI when they are used for generating populations of kinetic models that correspond to the same flux profiles but with different metabolite concentration vectors. For each vector from the set of 1000 metabolite concentrations, we considered the top 10 rules and assessed FI over the samples from the validation set that

satisfy these rules. The obtained distribution of FI over the set of metabolite concentration showed that the obtained rules were robust throughout the metabolite concentration space (Figure 5). FI was improved for the whole population compared to the original kinetic parameter space, i.e. it ranged from 0.293 to 0.931 with an asymmetrical distribution around the mean value of 0.734 (Figure 5).

We repeated the same analysis for the semi-quantitative rules obtained in Section 3.3.1. The distribution of FI for top 10 semi-quantitative rules showed the robustness of these rules throughout the metabolite concentration space (Figure 5). FI ranged from 0.276 to 0.883 with a symmetrical distribution around the mean value of 0.614 (Figure 5).

Since the values of metabolite concentrations and regimes in which enzymes operate are intrinsically inseparable, these results indicate that the proposed method allows us to identify the regions in both enzyme saturation and metabolite concentration space which are more likely to give rise to feasible kinetic models.

4. Conclusions

In this work, we introduce the first computational methodology capable of determining the important enzymes in the network along with the operating ranges of their saturation by substrates and products and their parameters that correspond to a given metabolic flux and a given metabolite concentration. The proposed approach is based on the ORACLE framework and machine learning methods and it offers information about enzymes that supplements the one obtained by experimental techniques. The obtained bounds on kinetic parameter values, and enzyme saturation levels can be used for postulation of hypotheses around observed physiological condition.

The proposed method can be considered also as a new parameter estimation procedure since it can identify enzymes whose saturations, if constrained to a narrow range, allow us to build the kinetic models capable to describe the studied physiology, and by this mean to provide accurate estimates of ranges of kinetic parameters relevant for the studied physiology.

This approach can also be used for efficient stratified sampling of solution spaces as it has been previously done for other biological systems (Zamora-Sillero et al., 2011). The proposed methodology represents the solution space, in this case the space of the parameters that are consistent with an observed physiology, as a set of multidimensional hyper-boxes. We can sample each of hyper-boxes independently, and the union of these samples will span the solution space. As a consequence, in contrast to commonly used methods for sampling of solution spaces such as artificial centering hit-and-run (Kaufman and Smith, 1998) the proposed method offers a possibility to perform uniform sampling in an efficient and *non-asymptotic* fashion. Another important feature of this approach is that it provides a very general and efficient way to compute the volume of the solution space as the sum of the volumes of the hyper-boxes. Once the ranges of parameters in the solution space are estimated, practically there are no additional computational requirements to compute the volume, which is an advantage with respect to the Monte Carlo based methods for the volume calculation (Wiback et al., 2004). This offers new possibilities to study large- and genome-scale metabolic networks and to analyze their properties. For example, with the proposed method we can analyze how the shape and the size of various solution spaces, such as the space of the steady-state fluxes, change under different physiological conditions.

Finally, it is important to note that the results of the proposed method can be used with any chosen distribution of samples provided that the samples span the solution space.

Acknowledgements

S.A. was supported by the Swiss National Science Foundation. L.M and V.H. were supported by the Ecole Polytechnique Fédérale de Lausanne (EPFL), and the RTD grants MalarX and BattleX, both within SystemsX.ch, the Swiss Initiative for Systems Biology evaluated by the Swiss National Science Foundation. We thank Genomatica for provided experimental data. We thank Joana Pinto Vieira for her help in coining the iSCHRUNK acronym.

Figures with captions

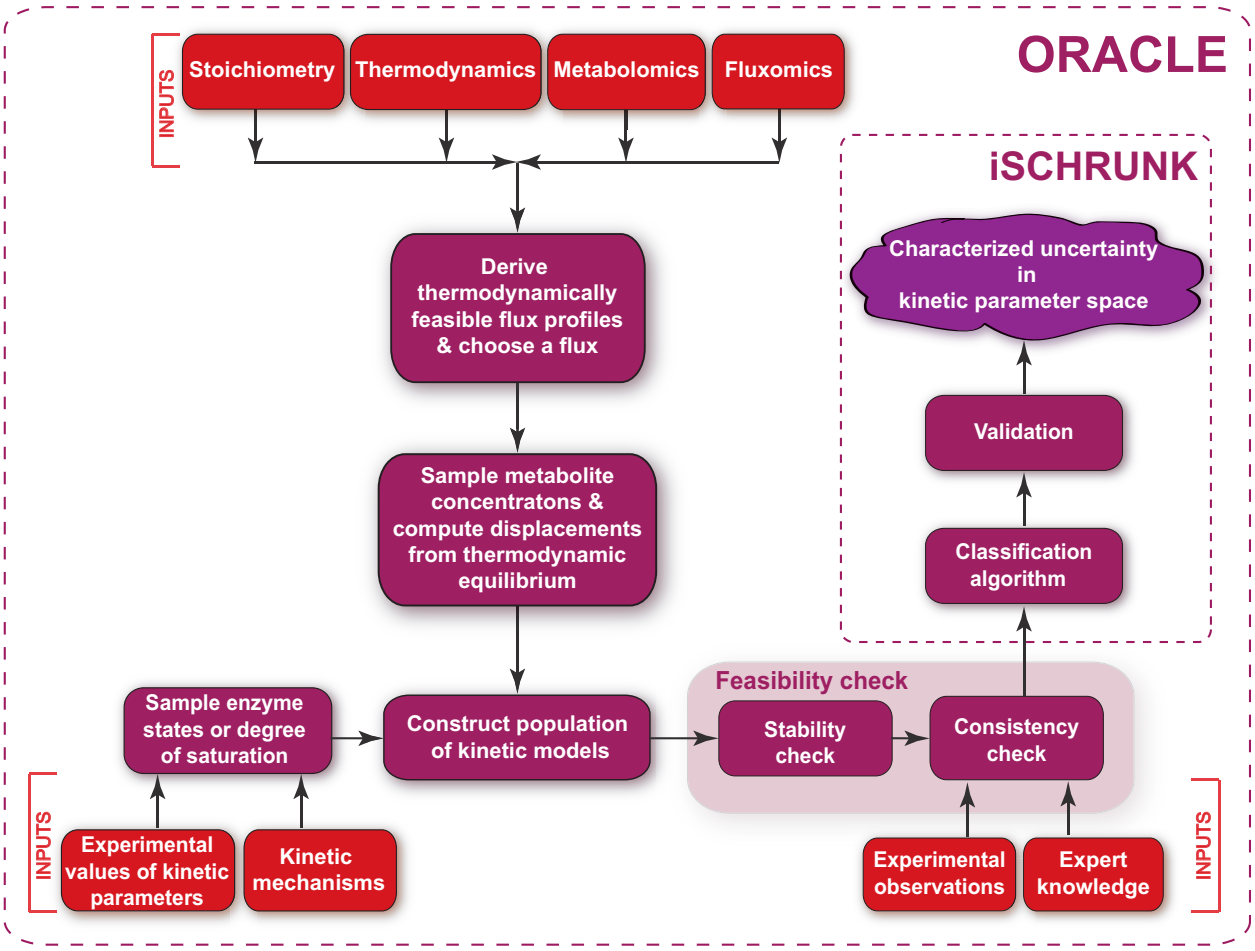

Figure 1: Generalized workflow of a computational procedure for reduction of uncertainty in kinetic parameter space (for details see main text).

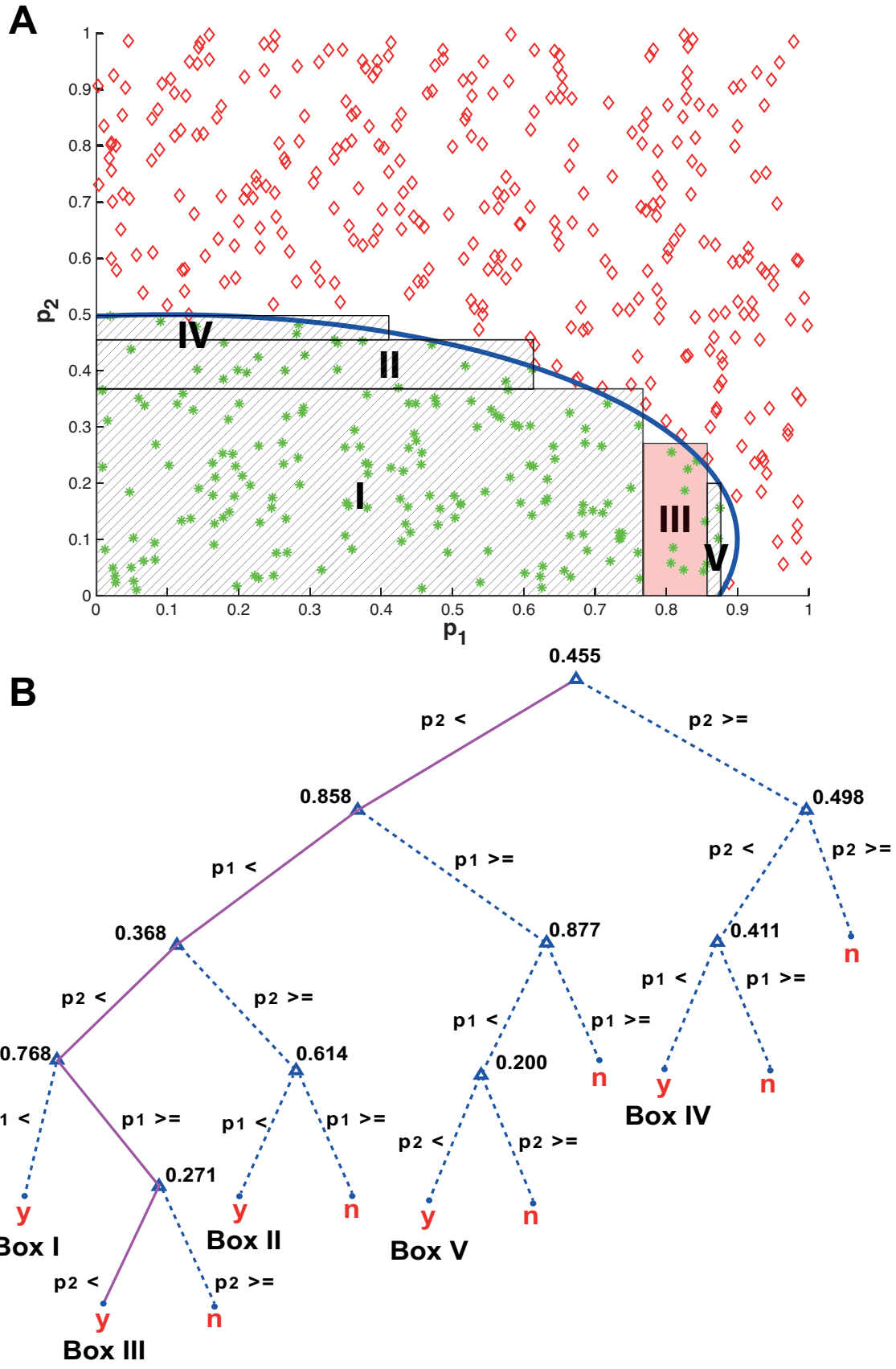
A
IV
II
I
III
V
p2
p1
B
0.455
p2 <
p2 >=
0.858
0.498
p1 <
p1 >=
p2 <
p2 >=
0.368
0.877
0.411
n
p2 <
p2 >=
p1 <
p1 >=
p1 <
p1 >=
0.768
0.614
0.200
n
y
n
p1 <
p1 >=
p1 <
p1 >=
p2 <
p2 >=
0.271
y
Box IV
n
y
Box I
y
Box II
n
y
Box V
n
p2 <
p2 >=
y
n
Box III

Figure 2: Toy example. Panel A – Random samples satisfying (green stars) and not satisfying (red diamonds) the given property (GP) in the parameter space of p_1 and p_2. The frontier between the two sets of samples is shown as the blue line. The decision-tree learning algorithm constructs the approximate rules (solid and hatched boxes) around the samples satisfying GP. Panel B – Binary decision tree with each branch (from the top of the tree till a leaf) representing a rule in the form of a sequence of conditions. Labels on leafs denote if the algorithm predicts that GP is satisfied ('y') or not ('n') within a rule. The rule with the solid line corresponds to the solid box in panel A.

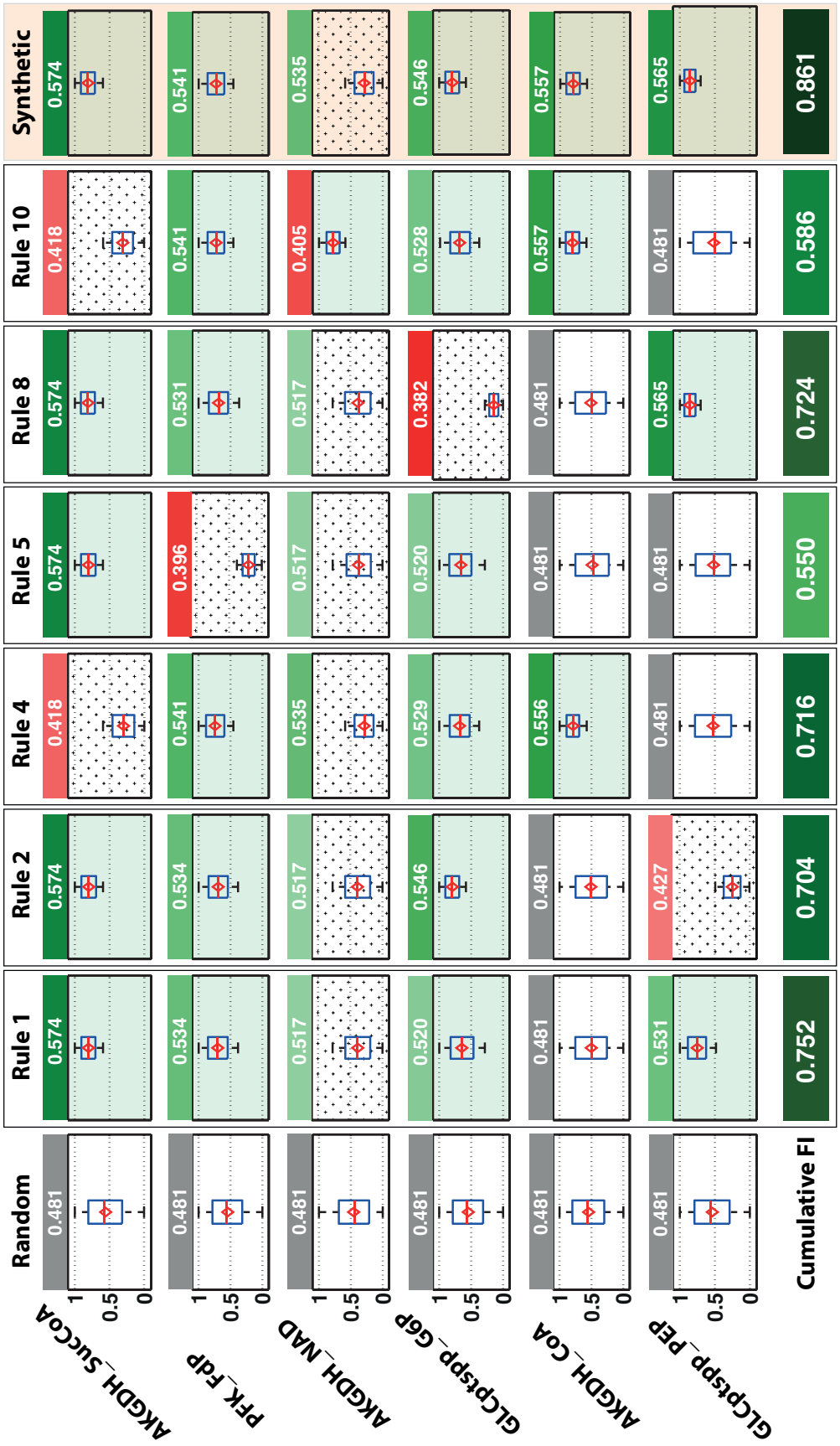
Random
Rule 1
Rule 2
Rule 4
Rule 5
Rule 8
Rule 10
Synthetic
Cumulative FI
AKGDH_Succoa
PFK_Fdp
AKGDH_NAD
GLCptspp_G6P
AKGDH_CoA
GLCptspp_PEP
0.481
0.574
0.574
0.418
0.574
0.574
0.418
0.574
0.752
0.481
0.534
0.534
0.541
0.396
0.531
0.541
0.541
0.541
0.704
0.481
0.517
0.517
0.535
0.517
0.517
0.405
0.535
0.716
0.481
0.520
0.546
0.529
0.520
0.382
0.528
0.546
0.550
0.481
0.481
0.481
0.556
0.481
0.481
0.557
0.557
0.724
0.481
0.531
0.427
0.481
0.481
0.565
0.481
0.565
0.586
0.861
1
0.5
0

Figure 3: Boxplots of the top 6 enzyme saturations constrained according to the chosen 6 inferred rules. The enzyme saturations either: (i) were not constrained, i.e. ranged from 0 to 1 (white background); or (ii) ranged in the medium-to-high saturation region (light green background); or (iii) ranged in the low-to-medium saturation region (hatched background). The colored stripes and numbers on the top of each boxplot correspond to FI computed over the training set when the corresponding range of enzyme saturation was used to constrain the parameter space, whereas the other enzyme saturations were allowed to range all admissible values. With respect to the random set of parameters (with FI of 0.481), the resulting FI could be: (i) improved (green stripes, darker the green stripes, higher the improvement); (ii) remain the same (gray stripes); or deteriorated (red stripes, darker the red stripes, worse the deterioration). The first column depicts the distribution of the top 6 enzyme saturations in the original kinetic parameter space. Notation for the labels on the vertical axis was as follows: "Enzyme"_"Metabolite that saturates the enzyme". Enzymes: AKGDH, 2-Oxoglutarate dehydrogenase; PFK, phoshofructokinase; GLCptspp, glucose transport via the phosphoenolpyruvate-pyruvate phosphotransferase system; Metabolites: SucCoA, Succinyl-CoA; FdP, D-Fructose 1,6-bisphosphate; NAD, Nicotinamide adenine dinucleotide; G6P, D-Glucose 6-phosphate; CoA, Coenzyme A; PEP, Phosphoenolpyruvate.

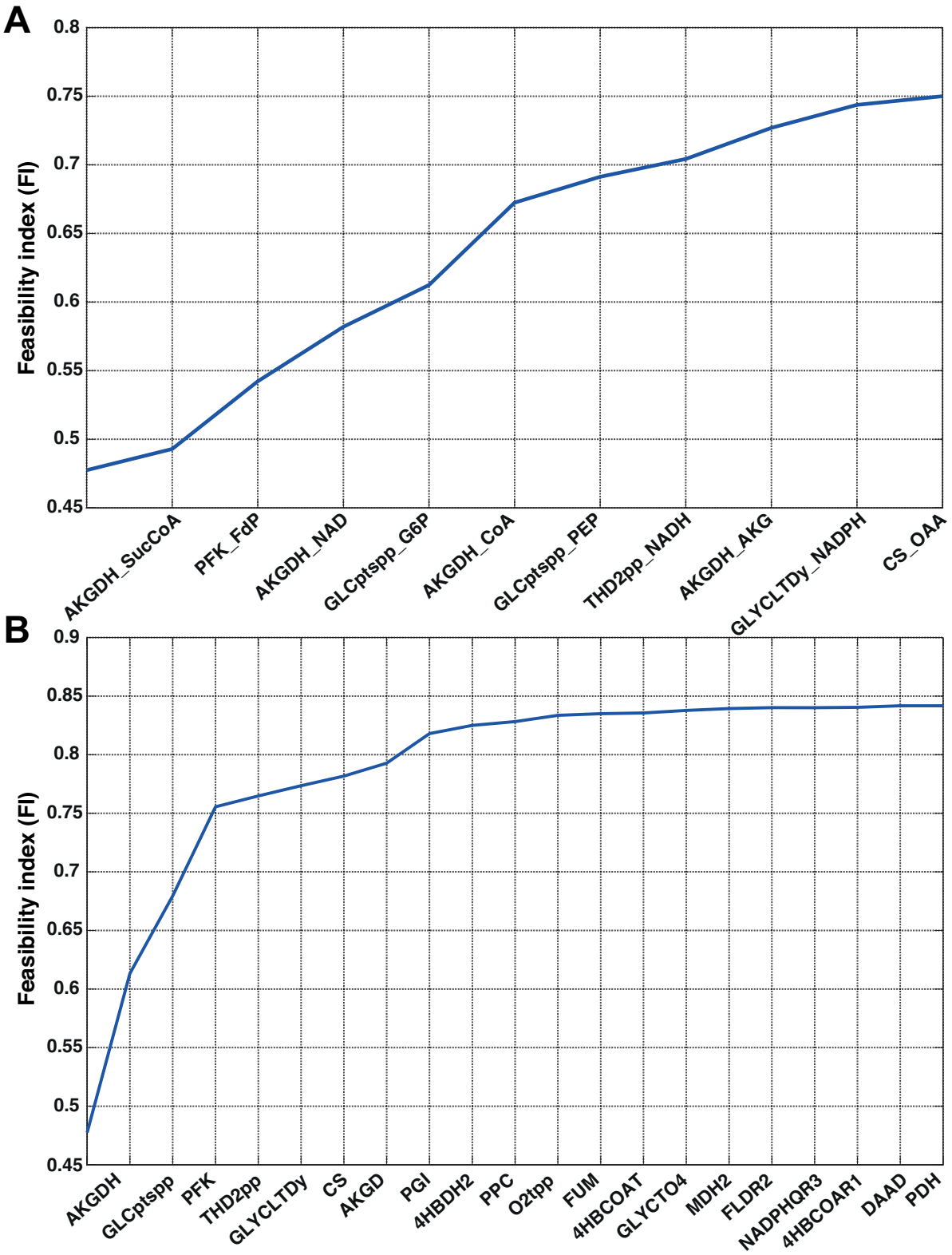
A
Feasibility index (FI)
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
AKGDH_SucCoA
PFK_FdP
AKGDH_NAD
GLCptspp_G6P
AKGDH_CoA
GLCptspp_PEP
THD2pp_NADH
AKGDH_AKG
GLYCLTDy_NADPH
CS_OAA
B
Feasibility index (FI)
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
AKGDH
GLCptspp
PFK
THD2pp
GLYCLTDy
CS
AKGD
PGI
4HBDH2
PPC
O2tpp
FUM
4HBCOAT
GLYCTO4
MDH2
FLDR2
NADPHQR3
4HBCOAR1
DAAD
PDH

Figure 4: Evolution of feasibility index (FI) as the constraints on the top 10 enzyme saturations (panel A) and the top 20 enzymes (panel B) are successively applied to constrain the space of kinetic parameters. Enzymes: AKGDH, 2-Oxoglutarate dehydrogenase; PFK, phoshofructokinase; GLCptspp, glucose transport via the phosphoenolpyruvate-pyruvate phosphotransferase system; THD2pp, NAD(P) transhydrogenase (periplasm); GLYCLTDy, glycolate dehydrogenase (NADP); CS, citrate synthase; AKGD, 2-oxoglutarate carboxy-lyase; PGI, glucose-6-phosphate isomerase; 4HBDH2, NADPH-dependent BDO dehydrogenase; PPC, phosphoenolpyruvate carboxylase; O2tpp, o2 transport via diffusion (periplasm); FUM, fumarase; 4HBCOAT, 4-hydroxybutanoate CoA transferase; GLYCTO4, glycolate oxidase; MDH2, Malate dehydrogenase (ubiquinone 8 as acceptor); FLDR2, NADPH-dependent flavodoxin reductase; NADPHQR3, NADPH Quinone Reductase (Menaquinone-8); 4HBCOAR1 4-hydroxybutanoate aldehyde dehydrogenase; DAAD, D-amino acid dehydrogenase; PDH, pyruvate dehydrogenase; Metabolites: SucCoA, Succinyl-CoA; FdP, D-Fructose 1,6-bisphosphate; NAD, Nicotinamide adenine dinucleotide; G6P, D-Glucose 6-phosphate; CoA, Coenzyme A; PEP, Phosphoenolpyruvate; NADH, Nicotinamide adenine dinucleotide - reduced; AKG, 2-Oxoglutarate; NADPH, Nicotinamide adenine dinucleotide phosphate - reduced; OAA, Oxaloacetate.

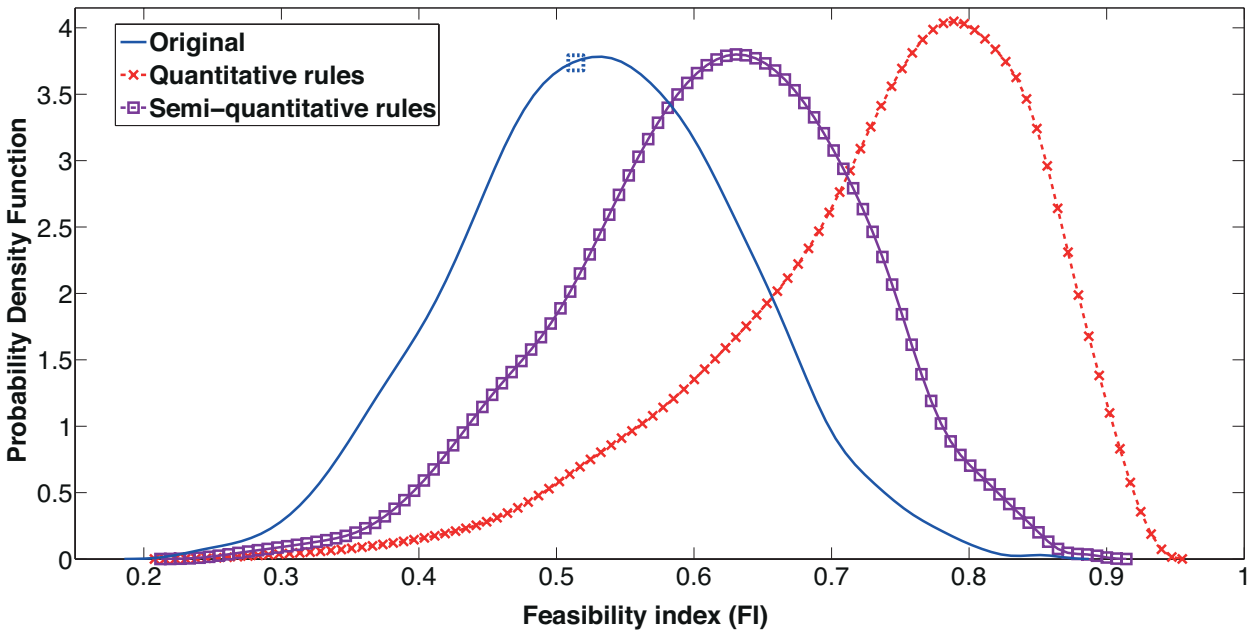

Figure 5: The probability density function (PDF) of feasibility index (FI) of the population of 1000 concentration vectors consistent with the observed physiology and thermodynamics. FI was computed over: (i) the original space of kinetic parameters (blue solid line), (ii) the subspace satisfying quantitative rules (red lines with stars) and (iii) over the subspace satisfying semi-quantitative rules (violet line with boxes). FI of the randomly chosen metabolite concentration vector used for analyses in sections 3.1 and 3.2 is shown as the dashed blue box.

References

Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., Jirstrand, M., 2014. Kinetic models in industrial biotechnology - Improving cell factory performance. Metabolic Engineering. 24**,** 38-60.

Asadollahi, M. A., Maury, J., Patil, K. R., Schalk, M., Clark, A., Nielsen, J., 2009. Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering. Metab Eng.

Bakker, B. M., van Eunen, K., Jeneson, J. A. L., van Riel, N. A. W., Bruggeman, F. J., Teusink, B., 2010. Systems biology from micro-organisms to human metabolic diseases: the role of detailed kinetic models. Biochem Soc T. 38**,** 1294-1301.

Birkenmeier, M., Mack, M., Roder, T., 2015a. Erratum to: A coupled thermodynamic and metabolic control analysis methodology and its evaluation on glycerol biosynthesis in Saccharomyces cerevisiae. Biotechnol Lett. 37**,** 317-26.

Birkenmeier, M., Mack, M., Roder, T., 2015b. Thermodynamic and Probabilistic Metabolic Control Analysis of Riboflavin (Vitamin B) Biosynthesis in Bacteria. Appl Biochem Biotechnol.

Bishop, C. M., 2006. Pattern recognition and machine learning. Springer, New York.

Borodina, I., Kildegaard, K. R., Jensen, N. B., Blicher, T. H., Maury, J., Sherstyk, S., Schneider, K., Lamosa, P., Herrgard, M. J., Rosenstand, I., Oberg, F., Forster, J., Nielsen, J., 2015. Establishing a synthetic pathway for high-level production of 3-hydroxypropionic acid in Saccharomyces cerevisiae via beta-alanine. Metabolic Engineering. 27**,** 57-64.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth, Belmont, Calif.

Bro, C., Regenberg, B., Forster, J., Nielsen, J., 2006. In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. Metab Eng. 8**,** 102-11.

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. Biotechnology journal. 8**,** 1043-1057.

Chowdhury, A., Zomorrodi, A. R., Maranas, C. D., 2014. k-OptForce: Integrating Kinetics with Flux Balance Analysis for Strain Design. Plos Comput Biol. 10.

Daniels, B. C., Chen, Y. J., Sethna, J. P., Gutenkunst, R. N., Myers, C. R., 2008. Sloppiness, robustness, and evolvability in systems biology. Current Opinion in Biotechnology. 19**,** 389-395.

Dash, S., Mueller, T. J., Venkataramanan, K. P., Papoutsakis, E. T., Maranas, C. D., 2014. Capturing the response of Clostridium acetobutylicum to chemical stressors using a regulated genome-scale metabolic model. Biotechnol Biofuels. 7.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern classification. Wiley, New York ; Chichester.

Gentle, J., 2003. Random Number Generation and Monte Carlo Methods.

Gilks, W., Richardson, S., Spiegelhalter, D., 1998. Markov Chain Monte Carlo in Practice.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., Sethna, J. P., 2007. Universally sloppy parameter sensitivities in systems biology models. Plos Comput Biol. 3**,** 1871-1878.

Han, J., Kamber, M., Pei, J. P. D., 2012. Data mining : concepts and techniques. Morgan Kaufmann ; [Oxford : Elsevier Science, distributor], Waltham, MA.

Hatzimanikatis, V., Bailey, J. E., 1996. MCA has more to say. J Theor Biol. 182**,** 233-242.

Hatzimanikatis, V., Bailey, J. E., 1997. Effects of spatiotemporal variations on metabolic control: Approximate analysis using (log)linear kinetic models. Biotechnol Bioeng. 54**,** 91-104.

Henry, C., Broadbelt, L., Hatzimanikatis, V., 2007. Thermodynamics-Based Metabolic Flux Analysis. Biophysical Journal.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., Stevens, R. L., 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology. 28**,** 977-U22.

Herrgard, M. J., Swainston, N., Dobson, P., Dunn, W., Arga, K. Y., Arvas, M., Bluethgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Le Novere, N., Li, P., Liebermeister, W., Mo, M., Oliveira, A., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasic, I., Weichart, D., Brent, R., Broomhead, D., Westerhoff, H., Kirdar, B., Penttila, M., Klipp, E., Palsson, B., Sauer, U., Oliver, S., Mendes, P., Nielsen, J., Kell, D., A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol, Vol. 26, 2008, pp. 1155-1160.

Hofmeyr , J., Cornish-Bowden, A., 1997. The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. Comp. Appl. Biosci. 13**,** 377-385.

Jamshidi, N., Palsson, B. O., 2010. Mass Action Stoichiometric Simulation Models: Incorporating Kinetics and Regulation into Stoichiometric Models. Biophysical Journal. 98**,** 175-185.

Jolliffe, I., 2002. Principal component analysis. Springer, New York.

Kaufman, D., Smith, R., 1998. Direction choice for accelerated convergence in hit-and-run sampling. Operations Research.

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. Metabolic Engineering. 25**,** 50-62.

King, Z. A., Feist, A. M., 2014. Optimal cofactor swapping can increase the theoretical yield for chemical production in Escherichia coli and Saccharomyces cerevisiae. Metabolic Engineering. 24**,** 117-128.

Liebermeister, W., Klipp, E., 2006. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. Theoretical Biology and Medical Modeling. 3.

Miskovic, L., Hatzimanikatis, V., 2010. Production of biofuels and biochemicals: in need of an ORACLE. Trends in Biotechnology. 28**,** 391-7.

Miskovic, L., Hatzimanikatis, V., 2011. Modelling of uncertainties in biochemical reactions. Biotechnology and Bioengineering. 108**,** 413-423.

Murabito, E., Verma, M., Bekker, M., Bellomo, D., Westerhoff, H. V., Teusink, B., Steuer, R., 2014. Monte-Carlo modeling of the central carbon metabolism of Lactococcus lactis: insights into metabolic regulation. PLoS One. 9**,** e106453.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., Palsson, B. O., 2011. A comprehensive genome-scale reconstruction of Escherichia coli metabolism. Molecular Systems Biology. 7**,** 535.

Osterlund, T., Nookaew, I., Bordel, S., Nielsen, J., 2013. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. Bmc Systems Biology. 7.

Quinlan, J. R., 1993. C4.5 : programs for machine learning. Morgan Kaufmann Publishers, San Mateo, Calif.

Savageau, M. A., 1969a. Biochemical Systems Analysis .1. Some Mathematical Properties of Rate Law for Component Enzymatic Reactions. Journal of Theoretical Biology. 25**,** 365-&.

Savageau, M. A., 1969b. Biochemical Systems Analysis .2. Steady-State Solutions for an N-Pool System Using a Power-Law Approximation. Journal of Theoretical Biology. 25**,** 370-&.

Savageau, M. A., 1970. Biochemical Systems Analysis .3. Dynamic Solutions Using a Power-Law Approximation. Journal of Theoretical Biology. 26**,** 215-&.

Schomburg, I., Chang, A., Placzek, S., Sohngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M., Schomburg, D., 2013. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Research. 41**,** D764-72.

Segel, I. H., 1975. Enzyme Kinetics.

Snitkin, E., Dudley, A., Janse, D., Wong, K., Church, G., Segrè, D., 2008. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biol. 9**,** R140.

Soh, K. S., Hatzimanikatis, V., 2010. Network thermodynamics in the post-genomic era. Current Opinion Microbiology. 13**,** 350-357.

Soh, K. S., Hatzimanikatis, V., 2014. Constraining the flux space using thermodynamics and integration of metabolomics data. Methods in Molecular Biology. 1191**,** 49-63.

Soh, K. S., Miskovic, L., Hatzimanikatis, V., 2012. From network models to network responses: integration of thermodynamic and kinetic properties of yeast genome-scale metabolic networks. FEMS Yeast Research. 12**,** 129-143.

Sohn, S. B., Kim, T. Y., Park, J. M., Lee, S. Y., 2010. In silico genome-scale metabolic analysis of Pseudomonas putida KT2440 for polyhydroxyalkanoate synthesis, degradation of aromatics and anaerobic survival. Biotechnology Journal. 5**,** 739-750.

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.

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bolling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Le Novere, N., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov, E., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H. G. M., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P., Palsson, B. O., 2013. A community-driven global reconstruction of human metabolism. Nature Biotechnology. 31**,** 419-+.

Tran, L. M., Rizk, M. L., Liao, J. C., 2008. Ensemble Modeling of Metabolic Networks. Biophysical Journal.

Wang, L., Birol, I., Hatzimanikatis, V., 2004. Metabolic Control Analysis under Uncertainty: Framework Development and Case Studies. Biophysical Journal. 87**,** 3750-3763.

Wang, L., Hatzimanikatis, V., 2006a. Metabolic engineering under uncertainty—II: Analysis of yeast metabolism. Metabolic Engineering. 8**,** 142-159.

Wang, L., Hatzimanikatis, V., 2006b. Metabolic engineering under uncertainty. I: Framework development. Metabolic Engineering. 8**,** 133-141.

Wiback, S. J., Famili, I., Greenberg, H. J., Palsson, B. O., 2004. Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. Journal of Theoretical Biology. 228**,** 437-447.

Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., Krebs, O., Bittkowski, M., Wetsch, E., Rojas, I., Muller, W., 2012. SABIO-RK-database for biochemical reaction kinetics. Nucleic Acids Research. 40**,** D790-D796.

Zamora-Sillero, E., Hafner, M., Ibig, A., Stelling, J., Wagner, A., 2011. Efficient characterization of high-dimensional parameter spaces for systems biology. Bmc Systems Biology. 5.