12-2012

Predicting Enzyme Targets for Optimization of Metabolic Networks under Uncertainty

David Christopher Flowers
dflower3@utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by David Christopher Flowers entitled "Predicting Enzyme Targets for Optimization of Metabolic Networks under Uncertainty." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Chemical Engineering.

Tsewei Wang, Cong T. Trinh, Major Professor

We have read this thesis and recommend its acceptance:

J. Douglas Birdwell

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Predicting Enzyme Targets for Optimization of Metabolic Networks under Uncertainty

A Thesis Presented for

The Master of Science

Degree

The University of Tennessee, Knoxville

David Christopher Flowers

August 2012

Dedication

To my mother

Diane

and my father

Mike

Acknowledgements

I would like to thank Drs. Tsewei Wang and J. Douglas Birdwell for employing me in the Laboratory for Information Technologies, giving me much-needed experience in research and for serving on my committee. Additionally I would like to specially thank Drs. Wang and Cong Trinh for serving as my co-advisors and giving advice critical to the advancement of this investigation. Acknowledgement also goes to the Department of Chemical and Biomolecular Engineering for giving me the opportunity to pursue a master's degree, the Newton HPC Program for providing computational resources, and LucidChart.com for providing software for flowchart construction under an educational account.

Abstract

Recently, ensemble modeling was applied to metabolic networks for the sake of predicting the effects of genetic manipulations on the observed phenotype of the system. The ensemble of models is generated from experimental wild-type flux data and screened using phenotypic data from gene overexpression and knockout experiments, leaving predictive models. The need for data from multiple genetic perturbation experiments is an inherent limitation to this approach. In this investigation, ensemble modeling is used alongside elementary mode analysis to attempt to predict those enzymatic perturbations that are most likely to result in an increase in a target yield and a target flux when only the wild-type flux distribution is known. Elementary mode analysis indicates the maximum theoretical yield and its associated steady-state flux distribution(s), and the minimal cut set knockouts are determined that eliminate all but the highest-yield elementary modes. These knockouts and other perturbations are simulated using all of the ensemble models, and the distributions of predicted fluxes and yields over the models are compared to elucidate which reactions and metabolites most likely limit the target yield and flux. Additionally, a systematic method is developed to simultaneously identify multiple reactions that are responsible for bottlenecks after the minimal cut set knockouts are performed. These methods are applied to a metabolic network that models 3-deoxy-D-arabinoheptulosonate-7-phosphate (DAHP) production in *E. coli*. Results show that pyruvate accumulation due to glucose uptake and erythrose-4-phosphate (E4P) shortages resulting from the slow reaction rate of transketolase (*Tkt*) limit

DAHP production. These results are consistent with published data, indicating that a detailed understanding of metabolic networks can be obtained with minimal experimental data. Additionally, the systematic method identifies four enzymes (*Tkt*, *Tal*, *Pps*, and *AroG*) that, when overexpressed experimentally, increase yield to nearly the maximum theoretical limit. Systematic analysis of a toy network also correctly identifies the post-MCS overexpression that results in the largest increases in yield and absolute fluxes. These results indicates that wild-type steady-state flux data can be used to accurately identify enzyme perturbation targets for increasing yield and target flux values.

Contents

List of Tables

List of Figures

Chapter 1

Overview and Background

1.1 Overview

In this report we present the development of a mathematically-based systematic simulation and data analysis of metabolic network structure to identify the top-ranked enzyme candidates whose under- or overexpression will optimize the production of a product produced by the network. The innovation of this approach is that it does not require intermediate experimental results to refine the analysis from one step of the process to the next. The only required experimental results are the steady-state fluxes of the various reactions in the network observed in the wild-type strain, which are often easily estimated from external fluxes given in the literature. Initial results reported here are very promising; the ranked list of candidate enzymes from the simulation match exactly the experimental results reported by other researchers in the literature for the same network system. This approach, if applicable to metabolic networks in general, would represent a significant advancement in the determination of genetic modifications in strain design necessary to increase the yield and productivity of a desired metabolic product, due to the fact that time-consuming and costly experimental results are not required.

The simulation and systematic data analysis methodology (SSDA) requires *a priori* the following data, and the major steps involved are listed below.

1.1.1 Required data set

Steady-state fluxes of network reactions of the wild-type strain are required. Also required are various set of network parameters and structural network data, such as standard Gibbs free energies of the involved reactions, indication of absolutely irreversible reactions, and stoichiometric relationships of the network reactions. The specific requirements are described in Chapter 2. For a chosen cellular system and the metabolic network, these data can usually be attained from literature or from prior work in one's own laboratory.

1.1.2 Major simulation and data analysis steps

1. Determine the set of elementary modes for the network.

2. Choose the mode(s) that gives maximum yield and determine the minimum cut set of enzymes that, upon knockout, leave only the desired elementary modes.

3. Generate a large set of ensemble models based on the *a priori* data described above. The generated ensemble models include the reaction kinetic parameters and fractions of the total enzyme concentrations that are in each complexed enzyme form. Upon simulation of the wild-type enzyme state, all generated models predict the same final steady-state fluxes as those supplied.

4. Subject the results from steps 1 to 3 to the developed systematic simulation and data analysis procedures involving linear algebra to obtain a ranked list of candidate enzymes to overexpress. These manipulations increase a specified product-to-input yield to near the theoretical maximum, as given by the elementary modes, and also increase the flux of the desired product reaction.

Various third-party software operating on the MATLAB platform are used in this study, some requiring modification to meet the needs of this investigation. These all will be described in detail in Chapter 2.

1.2 Background

Metabolic engineering is the directed improvement of the biochemical properties of cells by using recombinant DNA technology to alter the chemical reactions occurring within the cells or to add new reactions (Stephanopoulos, 1999). These improvements are often increases in yields with respect to a specified product and input. One of the primary ways this is accomplished is through gene underexpression or overexpression. Therefore, one of the questions metabolic engineers face is which genes should be targeted. Some of the potential target genes control enzyme concentrations within the cell. Enzyme concentrations are directly related to the kinetic properties of reactions in the cell, and by changing enzyme concentrations, one can often change the cellular properties that one wishes to improve. Because of this, it is feasible for one to focus one's attention on genes that control enzyme concentrations. This simplifies the problem to a degree by allowing one to ask which enzymes are most critical to the cellular reactions one wants to improve.

At this point, the problem becomes one of characterizing the various reactions within the cell. Cellular reactions form an interconnected network within the cell, with the products of one reaction serving as the reactants in another, and some cellular products serving as regulators of other reactions by inhibiting or activating them (Jeong et al., 2000). Knowledge of both the topology of a metabolic network and the kinetics of the reactions is useful for choosing target enzymes.

Various methods for analyzing metabolic networks have been developed. *In vivo*, it is difficult to attain kinetic data for chemical reactions (Edwards and Palsson, 2000). Due to this, many approaches to analyzing metabolic networks have been developed that rely heavily on the stoichiometry and topology of the network and avoid the

need for kinetic information. Some of these methods include flux balance analysis (Varma and Palsson, 1994), metabolic control analysis (Fell, 1992), elementary mode analysis (Schuster et al., 1999), and extreme pathway analysis (Schilling et al., 2000). These methods tend to reveal details about the steady state flux distributions of the network and do not describe the network dynamics. For this reason, they are useful for discovering the maximum possible yield of a particular product metabolite with respect to a given input metabolite.

One of the common goals of metabolic engineering is to alter the metabolic network to achieve this maximum yield. To accomplish this, the network's reactions to changes in enzyme concentration and/or activity need to be analyzed. The aforementioned analysis techniques are limited in their usefulness for this pursuit. Flux balance analysis, elementary mode analysis, and extreme pathway analysis can only consider those enzymatic changes that affect the topology of the metabolic network, namely, gene knockouts. Though this is useful, it does not allow for the analysis of gene overexpression. Metabolic control analysis, on the other hand, only allows one to analyze the effects of small changes in enzyme concentrations because it relies on linearization of the system, which is only valid for small perturbations to the network (Schuster, 1999). This limits its usefulness, as the gene underexpressions and overexpressions of interest usually result in large changes in the effected enzyme concentrations.

Another drawback of these methods is that they cannot consider overall production rates (i.e., the overall scale of fluxes across the networks), which is vital to the usefulness of the modified network. If one achieves a high yield for a target product and input, it is also important that the system be outputting the target product at a relatively high rate. Otherwise, the amount of product that can be produced in a reasonable time is too small to be useful, even if it is produced very efficiently.

Therefore, it is desirable that a dynamic modeling technique be developed. This is usually difficult due to the lack of available kinetic data associated with the enzymatic

reactions taking place in cells. Experiments that yield reaction rate data are time-consuming to conduct.

Recently, a method has been developed to generate a variety of dynamic models of metabolic networks without the need for detailed kinetic data. This method, described in detail by Tran et al. (2008), generates a large ensemble of models by randomly sampling model parameters that are constrained such that each model converges to a specified steady-state flux distribution for some initial conditions. The steady-state flux constraint greatly reduces the parameter space being spanned in the sampling of parameters, preventing one from having to generate prohibitively large sets of models to find at least some models that are representative of the actual system. This initial ensemble of models is then screened using readily available data from phenotypic experiments. These data are routinely collected during cellular metabolic engineering experiments (Tan et al., 2011). The screening process involves simulating the perturbed system using each of the ensemble models and comparing the perturbed models' predictions to the corresponding experimental results. Those models that do not exhibit the experimental phenotypes are screened out of the ensemble. This screening process is iterated with the screened models using additional experimental data. After each screening step, the ensemble becomes smaller, but more predictive. After a certain number of screening steps, a small ensemble of predictive models remain that would be useful for guiding further enzyme choices for overexpression and/or underexpression.

This method for generating dynamic models has already been applied successfully toward a number of systems (Contador et al., 2009; Rizk and Liao, 2009). However, it is not necessarily clear that the screened ensemble of models can be used to effectively aid in strain design. In the case where experimental data are scarce, the time and resources spent performing perturbation experiments to screen an ensemble could be used to test hypotheses on which enzyme perturbations are good targets. As a result, requiring these experiments to screen the ensemble restricts its usefulness. This is especially true when considering that there is no guarantee that the screened

ensemble will give accurate predictions. One question of importance is whether the particular perturbation experiments chosen for screening affect which models survive the screening process. Tran et al. (2008) have shown evidence that it does, but this demonstration is based on simulated experimental data rather than actual experiments.

A method such as this would be more useful if it did not rely on having extensive experimental information. If one knows little about which perturbations will produce favorable behavior in the system, it becomes difficult to choose perturbations *a priori* that will serve any purpose other than to allow for the construction of an ensemble of models. Ideally, one could predict those enzymes that are most likely good targets without having an extensively-screened ensemble, thus avoiding using time performing random perturbation experiments.

It is the aim of this investigation to develop whole-ensemble methods that require no screening to predict enzyme targets that increase a target yield and flux. One way of accomplishing this is finding those perturbations that produce favorable behavior in the largest number of ensemble models. "Favorable behavior" can be described as a large flux for a target reaction, coupled with a high yield with respect to an input. Therefore, it is clear that both fluxes and yields will be variables of interest. One approach to identifying the enzyme perturbations that would lead to favorable behavior is sensitivity analysis. Each enzyme's total concentration could be perturbed slightly upward and downward, and the network models' reactions to the perturbations would indicate candidate enzymes that are most likely to optimize network behavior. One concern with this suggestion, however, is that the overexpressions and knockouts imposed on the actual system involve large changes in the respective total enzyme concentrations, and the system's inherent nonlinearity may make extrapolations from small perturbations to large ones invalid. For this reason, a variation on sensitivity analysis is suggested and attempted here that uses large perturbations instead of small ones. This variation will be referred to as "perturbation analysis" to distinguish it from sensitivity analysis.

At this point, a possible framework for analysis becomes apparent. One can generate an ensemble of models from experimental wild-type flux data. This ensemble of models can then be used to simulate the system's response to a series of single-enzyme perturbations. These simulations will give time-dependent concentration and flux data that can be used to calculate target variables of interest, including target fluxes and yields. Because each model will predict a different set of fluxes and yields, one can look at the distribution of fluxes and yields over the ensemble models resulting from a given perturbation and compare these distributions to the wild-type target flux and yield to find the enzyme perturbation that increase these values for the most models. It can then be hypothesized that this enzyme perturbation is the one that is most likely to increase the target flux and yield in the actual system. This is the basic framework of the investigation to be reported here.

Potential problems may be foreseen in the details of this framework, however. One potential issue is that metabolic networks tend to be fairly robust, and changes in just one enzyme may not be enough to elicit a significant response from the system. One must realize that the robustness of a network is often due to redundancy in the network (Stelling et al., 2002). With this in mind, two methods are suggested. First, one can determine subsets of reactions that are structurally limited to having the same flux at steady state, as described by Pfeiffer et al. (1999), and perturb these enzymes in tandem. This is particularly helpful in the case of overexpressions, in which case one or more enzymes may restrict the flux of another overexpressed enzyme's reaction. Another approach is to knock out a minimal cut set that eliminates undesirable elementary modes. By eliminating all but the highest-yield elementary modes, only the maximum yield is theoretically possible at steady state. With this limitation, interest is transferred from yield (which is now restricted to the desirable theoretical maximum) to the ability of the model to reach steady state. Only by never reaching a steady state may a model not achieve the maximum theoretical yield.

Combining these approaches gives the framework which this investigation will follow. A series of perturbation simulations will be conducted using ensemble models

constructed with various specifications. Each model in the ensemble will predict different fluxes, yields, concentrations, and other measures of network behavior resulting from the perturbations. These predictions will be analyzed by examining the distributions of select measures of network behavior, including change in the target flux value, yield, and the rate of accumulation of metabolites. Favorable behavior will at first include maximizing the target flux and yield, and as the investigation reveals problematic bottlenecks in the network, additional conditions for favorable behavior will be considered, such as minimizing the rate of accumulation of metabolites that tend to accumulate and increasing the concentration of metabolites that tend to be scarce. Evaluating which perturbations tend to alleviate which bottlenecks will reveal the important mechanisms behind the functionality of the network. This insight will suggest enzyme targets for target flux and yield optimization.

The above framework is subject to human judgment, which is slow, potentially inaccurate, and impossible to automate. It also may require simulation of a large number of perturbations. To avoid these issues, a systematic approach to enzyme targeting will be developed. To start developing the systematic method, one needs to have quantitative data that defines current state of the system and a target optimal state of the system. For a metabolic network, one way to describe the state of the system quantitatively is with fluxes. As such, two flux vectors are calculated. The first is a representative flux vector that represents the general behavior of the models. The second is an ideal flux vector that has three features: (1) it has the maximum yield (the optimal state), (2) it is at steady state, and (3) of all maximum-yield steady-state vectors, it is closest to the representative vector according to some similarity metric. The purpose of similarity is to reduce the number and severity of perturbations required to reach the optimal fluxes and to maximize the likelihood that the ideal flux vector can be reached by the system.

One must find a way to systematically calculate the ideal flux vector. Recall that a minimal cut set knockout can be found that represses all but the maximum-yield elementary modes, forcing the network to have the maximum yield at steady

state. Also note that the maximum-yield elementary modes allow one to form a basis set of vectors for the maximum-yield steady-state space. This suggests an approach that will serve as the framework for the systematic method. First, simulate the MCS knockouts that eliminate all elementary modes but those with the maximum yield. Next, calculate a flux vector that is representative of the ensemble models' predicted flux distributions. Project this vector onto the maximum-yield steady-state space to obtain an ideal flux vector. Comparing the two vectors quantitatively can simultaneously suggest multiple enzyme targets. This is a significant advantage over perturbation analysis and other single-enzyme or predefined-group targeting methods, since the effects of multiple simultaneous enzyme overexpressions cannot be predicted from the effects of individual overexpressions. Also, attempting to test all possible combinations of enzymes or enzyme groups quickly becomes computationally prohibitive, whereas the systematic method only requires one perturbation to be simulated.

Chapter 2

Methods

A flowchart representing the general method to be presented is given in Figure 2.1. There are two different approaches that will be demonstrated in this study. The first, referred to as the manual approach, is driven by human judgment and focuses on comparing results qualitatively. It is primarily useful for hypothesis testing of network behavior. For example, one can hypothesize that the overexpression of an enzyme E will lead to a larger average flux for reaction R across the models. This approach lacks a predefined routine to guide the user, which is disadvantageous. For this reason, the second approach, called the systematic approach, was developed. This approach is driven by quantitative calculations and has specific instructions for each step. The systematic method outputs a list of the enzymes of the system rank-ordered according to how strongly the method indicates the enzymes to be effective overexpression targets. For this method, "effective" means likely to increase the target yield and flux of the network.

2.1 Choosing a network

Two networks were chosen for examination in this study. The first is a model of DAHP production in *E. coli*, and the second is a toy model used to further test the systematic method.

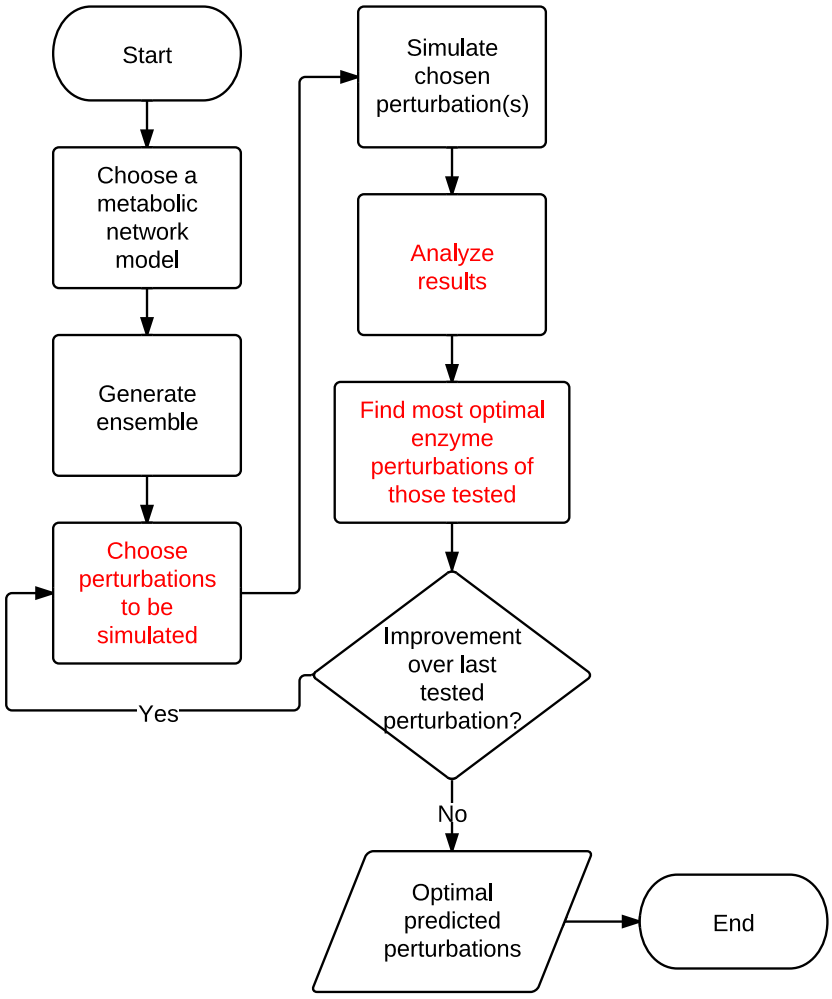

Figure 2.1: Flowchart of the general method of investigation. Steps that differ between the manual and systematic approaches are shown in red text.

2.1.1 DAHP production network

The network that will be studied in this investigation is a model of the production of 3-deoxy-D-arabinoheptulosonate-7-phosphate (DAHP) in *Escherichia coli*. DAHP is a precursor to the production of aromatic amino acids in the cell. Aromatic amino acids have numerous industrial uses, primarily in the food and pharmaceutical industries. For example, L-phenylalanine is used in the production of aspartame, an artificial sweetener, and is used as a flavor enhancer and as an intermediate in pharmaceutical production (Rizk and Liao, 2009).

This particular network was chosen for a few reasons. First, it is well-studied experimentally, allowing for checking of the feasibility and effectiveness of suggested enzyme targets against results reported in the literature. Also, it has previously been studied by Rizk and Liao (2009) using the ensemble modeling method, allowing for one to check one's application of the ensemble modeling method by reproducing similar results. This allows one to attribute any inconsistencies with experimental data to the novel approach presented here and not to incorrect application of any elements of the ensemble modeling method.

A map of the network to be studied is shown in Figure 2.2. This network is the same network studied by Rizk and Liao (2009). The network includes glycolysis, the phosphotransferase system for phosphorylating glucose and initiating glycolysis, the pentose phosphate shunt, part of the tricarboxylic acid cycle for succinate production, and additional pathways for acetate and formate production from pyruvate. In addition to the reactions shown in Figure 2.2, note that there are three artificial cofactor sink reactions for ATP, NADH, and NADPH in the model not shown explicitly.

Table 2.1 lists the full names and abbreviations of the enzymes and metabolites present in the network, as well as the stoichiometry of each reactions. Table 2.2 gives a list of the reactions in the network model and their properties. The properties listed in Table 2.2 include the wild-type steady-state fluxes used in this investigation,

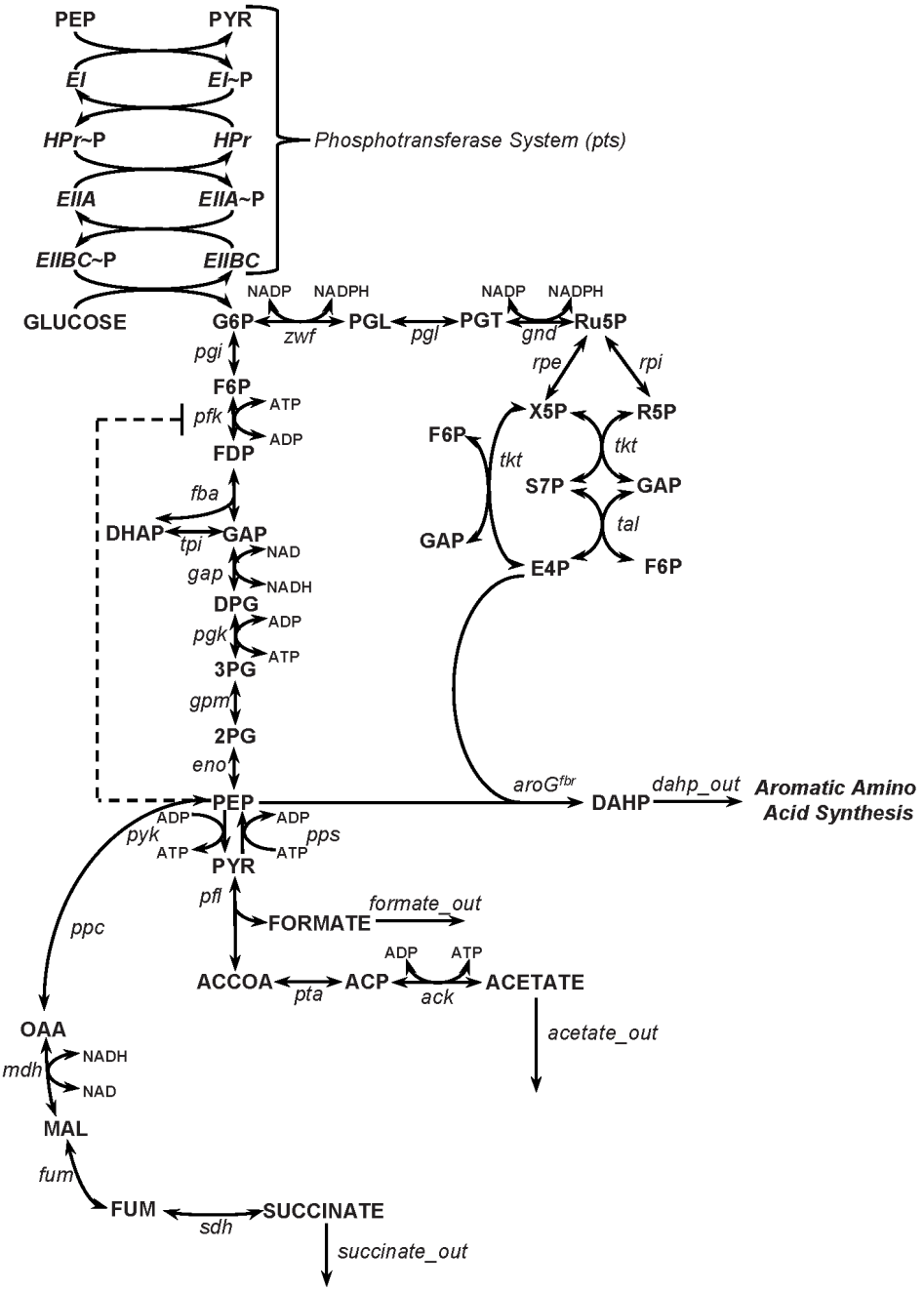

Figure 2.2: A map of the DAHP production network. Note that the artificial cofactor sink reactions are not shown in this map.
(Source: Rizk and Liao (2009))

Table 2.1: The full and abbreviated names of the metabolites and reactions, and the stoichiometry of each reaction in the DAHP network.

Metabolite no.	Abbreviation	Full name	Enzyme no.	Enzyme abbreviation	Enzyme full name	Reaction stoichiometry
1	2PG	2-phosphoglycerate	1	ack	acetate kinase	ACP+ADP ⇌ ACETATE+ATP
2	3PG	3-phosphoglycerate	2	aroG	2-dehydro-3-deoxyphosphoheptonate aldolase	E4P+PEP ⇌ DAHP
3	ACCOA	acetyl-CoA	3	EI	enzyme I	PEP ⇌ P1+PYR
4	ACETATE	acetate	4	EIIA	enzyme IIA	P2 ⇌ P3
5	ACP	acetyl phosphate	5	EIIBC	enzyme IIBC	P3+GLUCOSE ⇌ G6P
6	ADP	adenosine diphosphate	6	eno	enolase	2PG ⇌ PEP
7	ATP	adenosine triphosphate	7	fba	fructose biphosphate aldolase	FDP ⇌ DHAP+GAP
8	DAHP	3-deoxy-D-arabino-heptulosonate-7-phosphate	8	fum	fumarase	MAL ⇌ FUM
9	DHAP	dihydroxy acetone phosphate	9	gap	glyceraldehyde 3-phosphate dehydrogenase	GAP+NAD ⇌ DPG+NADH
10	DPG	1,3-biphosphoglycerate	10	gnd	6-phosphogluconate dehydrogenase	PGT+NADP ⇌ Ru5P+NADPH
11	E4P	erythrose-4-phosphate	11	gpm	phosphoglycerate mutase	3PG ⇌ 2PG
12	P1	phosphate group 1	12	HPr	histidine protein	P1 ⇌ P2
13	P2	phosphate group 2	13	mdh	malate dehydrogenase	NADH+OAA ⇌ MAL+NAD
14	P3	phosphate group 3	14	pfk	phosphofructokinase	ATP+F6P ⇌ ADP+FDP
15	F6P	fructose-6-phosphate	15	pfl	pyruvate formate lyase	PYR ⇌ ACCOA+FORMATE
16	FDP	fructose-1,6-biphosphate	16	pgi	phosphoglucoisomerase	G6P ⇌ F6P
17	FORMATE	formate	17	pgk	phosphoglycerate kinase	ADP+DPG ⇌ 3PG+ATP
18	FUM	fumarate	18	pgl	6-phosphogluconolactonase	PGL ⇌ PGT
19	G6P	glucose-6-phosphate	19	ppc	phosphoenolpyruvate carboxylase	PEP ⇌ OAA
20	GAP	glyceraldehyde-3-phosphate	20	pps	phosphoenolpyruvate synthase	ATP+PYR ⇌ ADP+PEP
21	GLUCOSE	b-D-glucose	21	pta	phosphate acetyltransferase	ACCOA ⇌ ACP
22	MAL	malate	22	pyk	pyruvate kinase	ADP+PEP ⇌ ATP+PYR
23	NAD	nicotinamide adenine dinucleotide	23	recATP	ATP recycle	ATP ⇌ ADP
24	NADH	nicotinamide adenine dinucleotide reduced	24	recNADH	NADH recycle	NADH ⇌ NAD
25	OAA	oxaloacetate	25	recNADPH	NADPH recycle	NADPH ⇌ NADP
26	PEP	phosphoenolpyruvate	26	rpe	ribulose-5-phosphate 3-epimerase	Ru5P ⇌ X5P
27	PGL	6-phosphogluconolactone	27	rpi	ribulose-5-phosphate isomerase	Ru5P ⇌ R5P
28	PGT	6-phosphogluconate	28	sdh	succinate dehydrogenase	FUM ⇌ SUCCINATE
29	PYR	pyruvate	29	tal	transaldolase	GAP+S7P ⇌ E4P+F6P
30	R5P	ribose-5-phosphate	30	tkt1	transketolase (1st reaction)	R5P+X5P ⇌ GAP+S7P
31	Ru5P	ribulose-5-phosphate	31	tkt2	transketolase (2nd reaction)	E4P+X5P ⇌ F6P+GAP
32	S7P	sedoheptulose-7-phosphate	32	tpi	triose phosphate isomerase	DHAP ⇌ GAP
33	SUCCINATE	succinate	33	zwf	glucose-6-phosphate dehydrogenase	G6P+NADP ⇌ PGL+NADPH
34	X5P	xylulose-5-phosphate	34	glucose_in	glucose transport	⇌ GLUCOSE
35	NADP	nicotinamide adenine dinucleotide phosphate	35	acetate_out	acetate transport	ACETATE ⇌
36	NADPH	nicotinamide adenine dinucleotide phosphate reduced	36	dahp_out	DAHP transport	DAHP ⇌
			37	formate_out	formate transport	FORMATE ⇌
			38	succinate_out	succinate transport	SUCCINATE ⇌

Table 2.2: The reactions of the DAHP model and their assumed standard Gibbs free energies (SGFE), inhibitor metabolites, and wild-type steady-state fluxes for a 75:25 glycolysis:pentose phosphate flux ratio.
(Source: Rizk and Liao (2009))

Enzyme no.	Enzymes	SGFE (kcal/mol)	Inhibitors	Steady-state fluxes (mmol gDCW^{-1} hr^{-1})
1	ack	-4.7		1.625
2	aroG	-17.9		0.26
3	ei	-6.45		1.3
4	eiia	-0.1		1.3
5	eiibc	-6.45		1.3
6	eno	-0.2		2.145
7	fba	1.1		1.105
8	fum	1.3		0.26
9	gap	4.2		2.145
10	gnd	-0.8		0.325
11	gpm	-2.2		2.145
12	hpr	-0.1		1.3
13	mdh	-4.8		0.26
14	pfk	-4.5	PEP inhibition	1.105
15	pfl	-2.5		1.625
16	pgi	-2.5		0.975
17	pgk	4.7		2.145
18	pgl	-13.3		0.325
19	ppc	-11.7		0.26
20	pps	-3.6		0.017
21	pta	-3.9		1.625
22	pyk	-8.4		0.342
23	recATP	-0.1		2.99
24	recNADH	-0.1		1.885
25	recNADPH	-0.1		0.65
26	rpe	-0.1		0.13
27	rpi	0.7		0.195
28	sdh	-0.7		0.26
29	tal	-0.6		0.195
30	tkt1	0.9		0.195
31	tkt2	-0.6		-0.065
32	tpi	0.2		1.105
33	zwf	-0.9		0.325
34	glucose_in	-3.5		1.3
35	acetate_out	-3.5		1.625
36	dahp_out	-3.5		0.26
37	formate_out	-3.5		1.625
38	succinate_out	-3.5		0.26

standard Gibbs free energies, and regulating metabolites for each reaction. The network includes 38 reactions, of which five are external transport reactions. One of the transport reactions inputs glucose into the network, while four separate outward transport reactions are responsible for exporting acetate, DAHP, formate, and succinate from the system.

Wild-type steady-state fluxes were determined by Rizk and Liao (2009) from external fluxes measured experimentally. However, it should be noted that the flux ratio between glycolysis and the pentose phosphate pathway at the flux split at glucose-6-phosphate (G6P) is unknown. Rizk and Liao (2009) generated four sets of ensemble models, each using a different glycolysis:pentose phosphate flux ratio (25:75, 50:50, 75:25, and 95:5), and determined that only the 75:25 and 95:5 split ratios lead to predictive ensembles. As such, these ratios are most likely more representative of the actual cellular system. For this investigation, a 75:25 split ratio was assumed.

The reaction governed by the enzyme phosphofructokinase (*Pfk*) is inhibited by phosphoenolpyruvate (PEP). This inhibition is modeled as competitive inhibition. Additionally, the enzyme 2-dehydro-3-deoxyphosphoheptonate aldolase (*AroG*) is assumed to have been modified to be resistant to feedback inhibition from tryptophan (Rizk and Liao, 2009).

2.1.2 Toy network

An additional, smaller network is studied to improve confidence in the generality of the systematic enzyme targeting method presented in Section 2.5.3. This toy network is the same network used by Trinh et al. (2009), though the standard Gibbs free energies of the reactions were contrived in this study. A map of the network is presented in Figure 2.3. For purposes of this study, metabolite A is considered the input of interest, and metabolite P is considered the product of interest. Reaction stoichiometry, standard Gibbs free energies, and wild-type steady-state fluxes are presented in Table 2.3. All reactions but r_{6r} and r_{8r} are irreversible. Metabolites

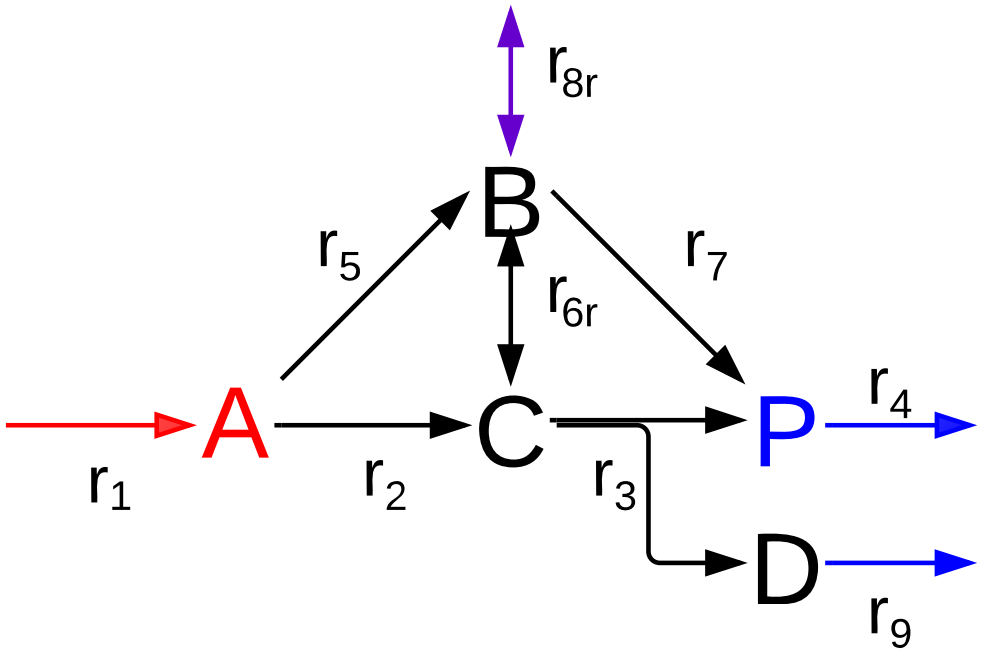

Figure 2.3: A map of the toy network. Inward transport fluxes are colored red, outward transport fluxes are colored blue, and reversible transport fluxes are colored violet. The input metabolite of interest is colored red, and the outward metabolite of interest is colored blue.

Table 2.3: Enzyme names and their corresponding reactions' wild-type steady-state fluxes, standard Gibbs free energies, and stoichiometries for the toy network.

Enzyme no.	Enzyme name	Wild-type steady-state flux (mmol gDCW^{-1} hr^{-1})	SGFE (kcal/mol)	Reaction stoichiometry
1	r_2	0.3	-5	A $\rightarrow$ C
2	r_3	0.75	-5	C $\rightarrow$ D + P
3	r_5	0.7	-5	A $\rightarrow$ B
4	r_{6r}	0.45	-0.1	B $\rightleftharpoons$ C
5	r_7	0.25	-5	B $\rightarrow$ 2 P
6	r_1	1	-5	$\rightarrow$ A
7	r_4	1.25	-5	P $\rightarrow$
8	r_{8r}	0	-0.1	B $\rightleftharpoons$
9	r_9	0.75	-5	D $\rightarrow$

include A, B, C, D, and P and are allowed to vary between 0.01 and 100 times their wild-type steady-state concentrations.

2.2 Generating the ensemble of models

Using the metabolic network information presented in Section 2.1, one can generate an ensemble of dynamic models for the network. The process and theory behind ensemble model generation is described in detail by Tran et al. (2008), with additional details given by Contador et al. (2009). The process as applied in this study will be briefly summarized in this section.

The first step in ensemble model generation is to check the thermodynamic feasibility of the directions of the supplied steady-state fluxes. This determination is based on the specified allowable Gibbs free energy ranges for each of the reactions. The allowable Gibbs free energies may either be specified directly or be calculated from the specified standard Gibbs free energies and allowable metabolite concentration ranges. For this study, the metabolite ranges specified by Tan et al. (2011) are used. Non-cofactor metabolites are allowed to vary between 0.01 and 100 times their steady-state concentrations, and cofactors are restricted to their steady-state concentrations. Cofactor concentrations are restricted in order to simulate the environment of the cell, where cofactor concentrations are tightly regulated. The calculated Gibbs free energies determine the allowable directions for each reaction. Those reactions that are limited to negative free energies may only react in the forward direction. Similarly, those reactions limited to positive free energies may only react in the backward direction. Reactions whose free energy limits span both positive and negative free energies may react in either direction. The forward direction for each reaction is defined by the stoichiometric matrix of the network.

Once the steady-state fluxes are found to be feasible according to the ranges of allowed Gibbs free energies, one must choose a kinetic model type to use to model each reaction individually. For this study, elementary reaction kinetic models are

used for their simplicity and versatility. Elementary reactions have been used in many ensemble modeling studies of metabolic networks to date (Tran et al., 2008; Contador et al., 2009; Rizk and Liao, 2009; Tan et al., 2011).

The generation procedure for a single model is summarized in Figure 2.4. To use elementary reaction modeling, each reaction in the network must first be expanded into a series of elementary reactions. Each elementary reaction j of overall reaction i follows the mass action principle shown in Equation 2.1.

$$v_{i,j} = k_{i,j}[x_1][x_2]\cdots[x_m] \tag{2.1}$$

In Equation 2.1, $v_{i,j}$ is the reaction rate of the jth elementary reaction of overall reaction i, $k_{i,j}$ is the kinetic parameter of elementary reaction i of overall reaction j, and $[x_k]$ represents the concentration of the kth reactant of m total reactants associated with elementary reaction j of overall reaction i. These reactants may be either enzyme complexes or metabolites. This expansion is shown by Contador et al. (2009) for reactions with one or two reactants and one or two products, with or without inhibition.

Note that Equation 2.1 involves many variables that may be unknown, primarily the metabolite and enzyme complex concentrations. To proceed without knowledge of these variables, the concentrations of the metabolites and enzyme complexes are lumped into the kinetic parameter. In the parameter sampling step of the ensemble generation step, the lumped kinetic parameter is sampled, making concentration data unnecessary. As a result of the parameter lumping, each metabolite and enzyme complex concentration is expressed as a fraction of a reference concentration. For metabolites, the reference concentration is the steady-state concentration, and for enzyme complexes, the reference concentration is the total enzyme concentration (the sum of the concentrations of the complexed forms of the enzyme, including the free enzyme). Each concentration is divided by its reference concentration, and the kinetic parameter is multiplied by each of the reference concentrations. This process

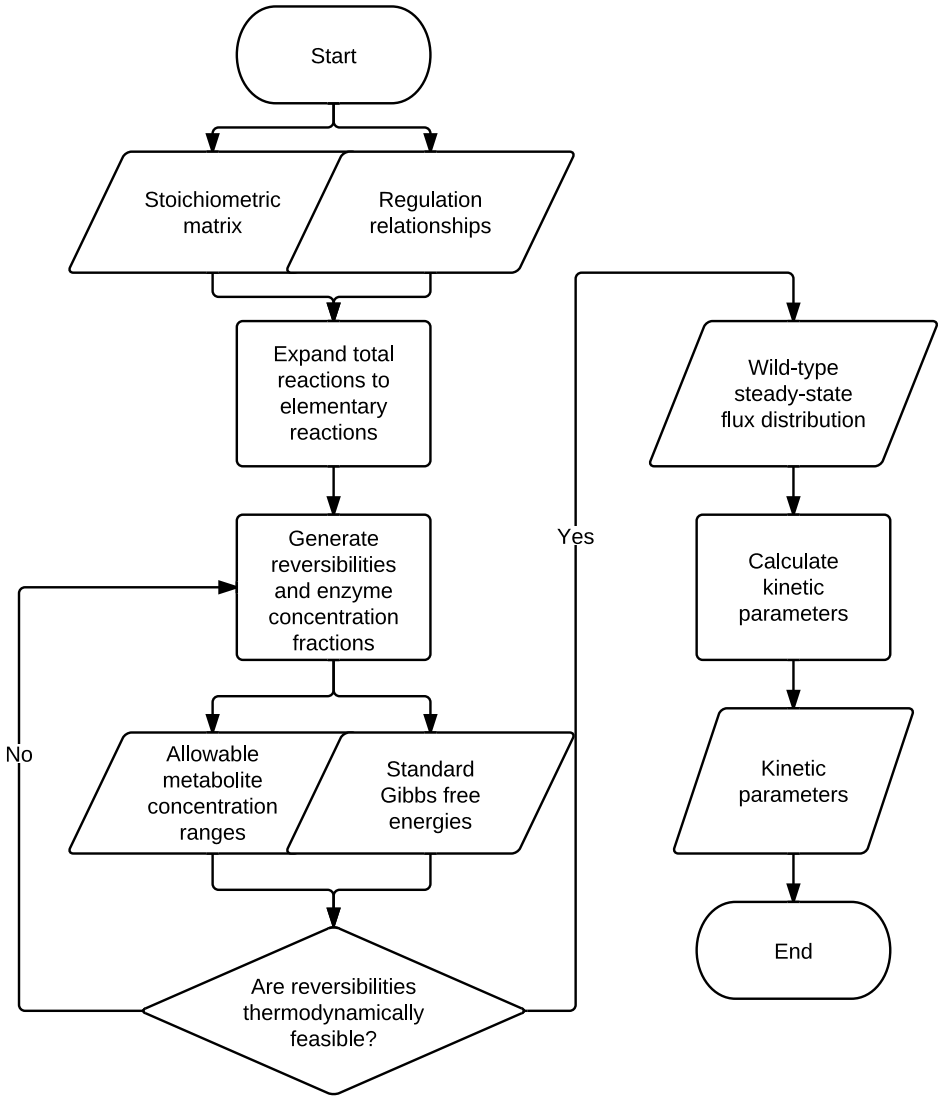

Figure 2.4: The process by which a single model of an ensemble is generated. This process needs to be repeated n times, where n is the specified number of models in the ensemble. Different models are generated each run as a consequence of random sampling of reaction rate reversibilities and enzyme concentration fractions.

is described in much more detail by Tran et al. (2008). As a result, the metabolite and enzyme complex concentration values are expressed as fractions of their reference concentration, as shown in Equation 2.2.

$$\begin{aligned} v_{i,j} &= (k_{i,j}[x_{1,ref}][x_{2,ref}]\cdots[x_{m,ref}])\left(\frac{[x_1]}{[x_{1,ref}]}\right)\left(\frac{[x_2]}{[x_{2,ref}]}\right)\cdots\left(\frac{[x_m]}{[x_{m,ref}]}\right) \\ &= K_{i,j}[\hat{x_1}][\hat{x_2}]\cdots[\hat{x_m}] \end{aligned} \tag{2.2}$$

In Equation 2.2, $[x_{k,ref}]$ is the reference concentration of reactant k, $K_{i,j}$ is the lumped kinetic parameter, and $[\hat{x_k}]$ is the concentration fraction of reactant k.

In this study, the elementary reaction expansion process is done using matrix representations of the system. The system of total reactions is represented as a matrix equation, shown in Equation 2.3.

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{S}\cdot\mathbf{v}(t) \tag{2.3}$$

In Equation 2.3, $\mathbf{x}$ is the vector of metabolite concentrations at time t, $\mathbf{S}$ is the stoichiometric matrix representing the network structure, and $\mathbf{v}(t)$ is the flux vector of the system at time t. The stoichiometric matrix relates reaction fluxes to metabolite concentrations through the stoichiometry of each reaction. Each row of $\mathbf{S}$ represents a metabolite, and each column represents a reaction. Element $\mathbf{S_{i,j}}$ is the stoichiometric coefficient of metabolite i in reaction j. After expanding the total reactions to elementary reactions, Equation 2.4 describes the system,

$$\frac{d\mathbf{x_{exp}}(t)}{dt} = \mathbf{S_{exp}}\cdot\mathbf{v_{exp}}(t) \tag{2.4}$$

where $\mathbf{x_{exp}}(t)$ is the expanded vector of metabolite and enzyme complex concentrations, including both metabolite concentrations and enzyme complex concentrations; $\mathbf{S_{exp}}$ is the expanded stoichiometric matrix of the system that describes the structure of the elementary reaction network; and $\mathbf{v_{exp}}(t)$ is the expanded flux vector that lists the fluxes of each of the elementary reactions at time t.

Once the elementary reactions are constructed, reversibilities are randomly sampled for each of the elementary reaction steps. The definition of reversibility for elementary reaction step j of total reaction i as defined by Tran et al. (2008) is given in Equation 2.5,

$$R_{i,j} = \frac{\min(v_{i,j}^{forward}, v_{i,j}^{reverse})}{\max(v_{i,j}^{forward}, v_{i,j}^{reverse})} \tag{2.5}$$

where $R_{i,j}$ is the reversibility of elementary reaction step j for overall reaction i. "Elementary reaction step" refers to a pair of elementary reactions that governs the forward and reverse reaction rates of the reactions between one set of elementary metabolites and the next within an overall reaction; $v_{i,j}^{forward}$ refers to the forward elementary reaction rate, and $v_{i,j}^{reverse}$ refers to the reverse elementary reaction rate. Reversibility ranges between 0 and 1, with 0 indicating an irreversible reaction and 1 indicating a perfectly reversible reaction.

Additionally, a set of concentration fractions is sampled for each of the enzymes. An enzyme concentration fraction is the fraction of an enzyme's total concentration that is found in a specified complex form of the enzyme. For example, if an enzyme can be found in its free form or bound to either metabolite A or B, then there are three complex forms for the enzyme. The enzyme concentration fraction for the free form specifies the fraction of the total that is present in the free form. The concentration fraction is similarly defined for each of the complexed A and B forms. Concentration fractions are constrained such that all enzyme concentration fractions for a given enzyme must add up to 1.

After the reversibilities and enzyme concentration fractions are chosen via sampling, the reversibilities must be checked for thermodynamic feasibility. Each randomly sampled reversibility is checked against specified allowable ranges of Gibbs free energies. Those that are not consistent are resampled until they meet thermodynamic specifications. The feasible set of reversibilities is used along with the enzyme concentration fractions to solve for kinetic parameters that give the specified steady-state flux distribution.

Once this procedure is finished, one ensemble model has been generated. The procedure is repeated n times, where n is the specified number of models desired in the ensemble. This investigation uses $n = 1500$ models for all simulations for both the DAHP and toy networks, though larger networks will need more models to effectively cover the kinetic parameter space.

In this study, the ensemble generation procedure is performed using a modified version of a MATLAB® script provided by Tan et al. (2011). The program takes as input the stoichiometric matrix of the total reaction system, the standard Gibbs free energies of the reactions, and the allowable concentration ranges The primary modifications made to the program were the separation of the ensemble generation and simulation functionalities into two separate programs and the allowing of multiple enzyme perturbations at any desirable overexpression or underexpression level. More details about the simulation procedure can be found in Section 2.4.

2.3 Choosing the perturbations to be simulated

Many tools are available to help the investigator determine the perturbations to be simulated. The manual method relies on iterating through all possible perturbations in a process called perturbation analysis (see Section 2.3.1). In perturbation analysis, enzymes may be expressed individually or as parts of predefined groups. One can group enzymes into enzyme subsets, which are groups of enzymes that must have the same steady-state fluxes (see Section 2.3.2). Another option is that a minimal cut set may be knocked out prior to perturbation analysis to constrain network functionality and minimize the number of perturbations to be iterated over. This minimal cut set is one of the smallest groups of enzymes that will repress a specified functionality from the metabolic network when knocked out (see Section 2.3.3). Functionalities to be eliminated are described by elementary modes, which are flux distributions that represent the fundamental steady-state flux modes of the network. Elementary modes reveal the maximum theoretical yield of the system and specify how this yield may be

achieved (see Section 2.3.2). The systematic method begins by selecting the minimal cut set knockouts.

2.3.1 Performing perturbation analysis

Perturbation analysis is a modification to sensitivity analysis that can potentially identify enzyme targets for knockout or overexpression. Sensitivity analysis is, in essence, the monitoring of the response of output variables to a small change in a chosen variable in the dynamic equations. This process is exemplified in the determination of control or sensitivity coefficients. A sensitivity coefficient is the fractional changes in one variable in response to an infinitesimal fractional change in another variable (Fell, 1992). These sensitivity coefficients are defined in Equation 2.6.

$$C_P^V = \lim_{\delta P \to 0} \frac{\delta V / V}{\delta P / P} \tag{2.6}$$

In Equation 2.6, C_P^V is the sensitivity coefficient for variable V with respect to variable P.

In metabolic control analysis, flux sensitivity coefficients are used extensively (Fell, 1992). A flux sensitivity coefficient is the control coefficient of a reaction flux with respect to a total enzyme concentration. The problem with sensitivity coefficients with regard to metabolic networks, however, is that enzyme concentration perturbations in perturbation experiments usually involve very large changes in enzyme concentrations. It is not guaranteed that the effect of small changes is at all indicative of the effects of large changes, i.e., the response may not be linear. For this reason, analysis using large changes is likely to be more predictive of the behavior of the actual biological system under perturbation.

Because they are not dependent on linearized approximations of network behavior, ensemble models allow for large changes in enzyme concentrations to be simulated.

Perturbation analysis, then, is the prediction of the response of a target variable to a large change in an enzyme's concentration via ensemble model simulation.

Perturbation analysis is performed through a series of simulations of individual enzyme overexpressions and knockouts. The overexpression and knockout of each individual enzyme or enzyme group is simulated using each model in the ensemble. Either individual enzymes or independent groups of enzymes, such as enzyme subsets (see Section 2.3.2) may be perturbed in turn and iterated over. Each model predicts a different network behavior. This diversity of behaviors is captured in the distributions of the values of the variables of interest predicted by the models. Variables of interest may include fluxes, concentrations, yields, and rates of accumulation of metabolites. Each distribution for each variable of interest resulting from the simulations is represented as a histogram and compared to its respective wild-type value to determine which enzyme perturbations tend to increase the variables of interest the most. More details on how this comparison to wild-type values is performed is presented in Section 2.5.1.

2.3.2 Elementary mode and enzyme subset calculation

Elementary modes describe pathways through a metabolic network that consist of an irreducible set of enzymes required to maintain a steady state (Trinh et al., 2009). Each elementary mode can be considered as a steady-state flux vector that indicates one of these irreducible pathways. Since these vectors are at steady state, where $\frac{dx}{dt} = 0$, they lie within the null space of the stoichiometric matrix $\mathbf{S}$ and are mapped to 0 by $\mathbf{S}$. The entire set of elementary mode flux vectors spans the null space of $\mathbf{S}$. Since all steady-state vectors for a metabolic network lie within the null space of that network's stoichiometric matrix $\mathbf{S}$, any steady-state flux distribution of the network can be represented by some linear combination of elementary modes. As such, elementary modes provide a way to calculate the theoretical maximum yield of a metabolic network with respect to an input and output flux pair of interest. By

finding the elementary mode (or set of modes) with the highest yields, one has found both the highest yield itself and the particular flux space that accomplishes this yield. This knowledge can then be used to determine a minimum cut set of enzymes that, when knocked out, will force the system to manifest the elementary modes at steady state that give the maximum yield. More information on the minimum cut set and its determination is presented in Section 2.3.3.

To derive the elementary modes of the system, the MATLAB® toolbox Metatool 5.1 is used. The toolbox and its availability, use, and functionality are described by Kamp and Schuster (2006). Additional details on the algorithm used by the program to find the elementary modes are presented by Urbanczik and Wagner (2005).

As a part of its calculation routine, Metatool finds subsets of enzymes that are structurally limited to the same flux at steady state (Pfeiffer et al., 1999). These subsets are useful to this study because they reveal enzymes that are parts of a reaction set, such as those involved in a linear chain of reactions, that must be overexpressed as a whole for any significant effect on network behavior to be assessed.

The use of Metatool in this study involves the following inputs:

1. the stoichiometric matrix of the network to be studied and
2. the indication of the irreversibility of each reaction in the network.

Metatool then outputs the following:

1. the flux distributions defining each elementary mode and
2. the enzyme subsets.

For the DAHP production network described in Section 2.1, ten reactions were considered to be irreversible. These reactions were *AroG*, *Pfk*, *Pgl*, *Pyk*, *Pps*, and the five transport reactions. With this input, Metatool indicated twenty-six elementary modes and fifteen enzyme subsets. Of the elementary modes, one mode exhibited the maximum theoretical yield of 0.8571. This mode is shown in Figure 2.5. The enzyme subsets are shown in Figure 2.6.

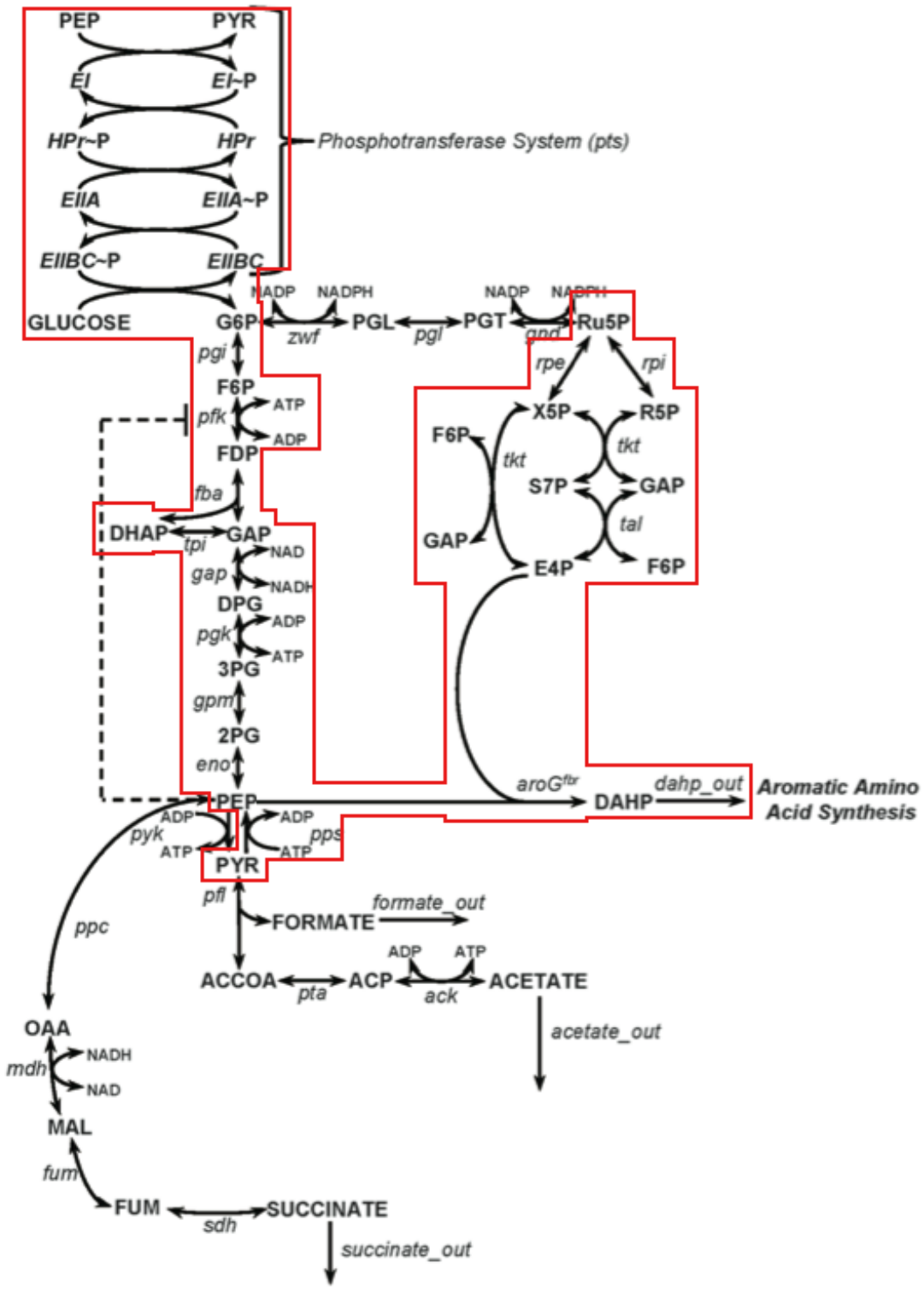

Figure 2.5: Reactions enclosed in the box constitute the maximum-yield elementary mode of the DAHP production network, as determined by Metatool.

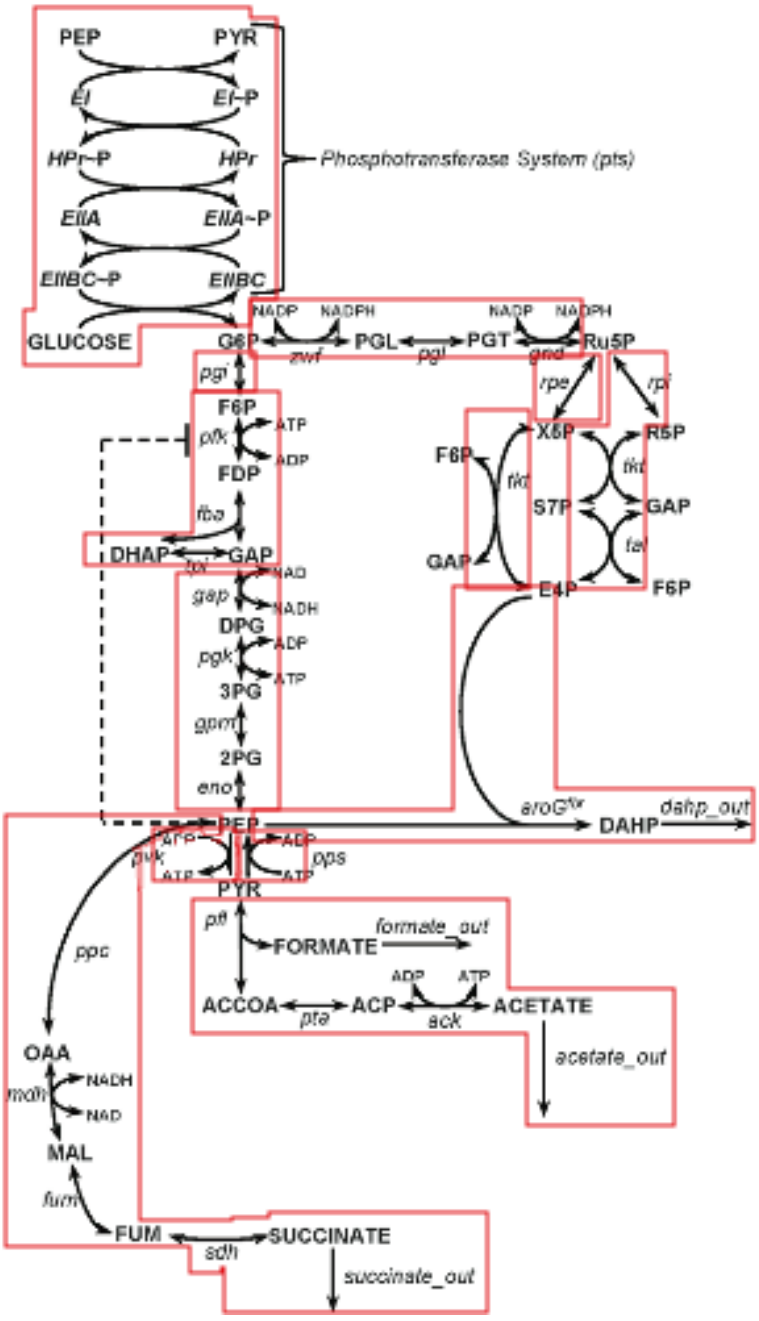

Figure 2.6: Each box encloses an enzyme subset of the DAHP production network, as determined by Metatool.

2.3.3 Minimum cut set determination

For a metabolic network, a minimum cut set is one of the smallest sets of enzymes that, when knocked out, represses a specified functionality (Klamt, 2006). The specified functionality can be a set of elementary modes which, when repressed, will no longer be expressed as steady-state flux distributions. For this study, a minimum cut set was determined to repress all elementary modes other than the maximum-yield elementary modes, while preserving all maximum-yield elementary modes. By knocking out this minimal cut set, the system is guaranteed to have the maximum theoretical yield if it reaches a steady state.

CellNetAnalyzer, a MATLAB® toolbox, was used to calculate the minimum cut set. Klamt et al. (2007) has given a description of the program, its capabilities, and its availability. To obtain minimum cut sets from the program, two inputs were required:

1. the elementary modes that must be repressed, and

2. the elementary modes that must be retained.

Both of these must be specified separately because the modes to be repressed represent the smallest set of modes that will be guaranteed to be repressed by the calculated minimum cut sets, and similarly for the modes to be retained. To illustrate this point, imagine a network with three elementary modes. Suppose that one wants to eliminate the first mode and retain the second and third. If one only specifies that the first mode is to be eliminated, the second and third modes may also be eliminated by the minimum cut sets that are calculated. If one wants to ensure that they are not eliminated, one must explicitly specify that they are to be retained.

For the DAHP production network described in Section 2.1.1, 100 minimal cut sets were found that repress all elementary modes except the maximum-yield mode shown in Figure 2.5. The final minimal cut set chosen for follow-up investigations included *Pfl*, *Ppc*, *Pyk*, and *Zwf*. This minimal cut set was chosen out of the 100 because it was the only minimal cut set for which all metabolites that were products of at least

one intact reaction were reactants of another intact reaction. "Intact reaction" refers to a reaction that is not knocked out. This was determined by visual inspection of the network structure.

2.4 Simulating perturbations using the ensemble models

Simulation using the ensemble models is conducted by setting up and solving a system of ordinary differential equations. Each of the elementary reactions has its own rate equation written in a form derived from the mass action principle shown in Equation 2.1. The inputs of Equation 2.1 are normalized concentrations, and the outputs are reaction rates. In order to set up a solvable system of ODEs, the reaction rate outputs need to be related back to the concentrations. This can be done by substituting Equation 2.1 for each elementary reaction into $\mathbf{v_{exp}}$ of Equation 2.4. This results in $\mathbf{v_{exp}}$ being a time-dependent vector of mass-action principle equations, with metabolite concentrations as both the inputs and time-dependent outputs, and the resulting ODE system is solvable.

The system of ODEs is used to simulate the effect of enzymatic perturbations on the steady-state flux levels and concentrations. To simulate an enzyme perturbation, the total concentration of an enzyme is increased or decreased by some factor. The absolute total enzyme concentration has been lumped into the kinetic parameter during the concentration normalization process. For this reason, the kinetic parameters of the elementary reactions for the enzyme of interest and its complexes are multiplied by the desired perturbation factor. This process is described in more detail by Tran et al. (2008). Note that changing the total enzyme concentration does not affect the enzyme concentration fractions that were sampled as described in Section 2.2. In this study, unless otherwise noted, overexpressions are represented by

a twofold increase in total enzyme concentration, and knockouts are represented by a 99 percent decrease in total enzyme concentration.

In perturbation simulations, initial conditions need to be specified in the form of concentrations of metabolites and enzyme concentration fractions, the latter being sampled as described in Section 2.2. To attempt to use realistic concentration values, the initial concentrations for the metabolites are set to the wild-type steady-state concentrations. In the kinetic parameter lumping procedure described in Section 2.2, concentrations were expressed as percentages (in decimal form) of their respective steady-state concentrations. As a result, the initial condition vector of concentrations is a vector of ones. A second motivation for using the wild-type steady-state concentrations instead of random initial conditions is to avoid possible problems with multiple steady states. Though Tan et al. (2011) mention that multiple steady states are relatively rare, experience gathered from this study indicate otherwise, with as many as 90 percent of generated models exhibiting multiple steady states (unpublished results).

In summary, the dynamic simulation process takes as inputs the following:

1. the kinetic model parameters and enzyme concentration fractions generated from ensemble modeling,

2. the initial conditions in terms of percentages of wild-type steady-state concentrations of metabolites,

3. the length of time to be simulated, and

4. the fold-changes for the total enzyme concentrations of the enzymes to be perturbed,

and outputs the following:

1. the dynamic responses of the fluxes for both total and elementary reactions and

2. the dynamic responses of concentration percentages of metabolites and enzyme complexes relative to the wild-type steady-state concentrations.

The program used to conduct the simulations is a modified version of a MATLAB® script written by Tan et al. (2011). It utilizes the **ode15s** ODE solver. The primary modifications to the script are the removal of a simple check for multiple steady states, the parallelization of the script to simulate using multiple models simultaneously, and the separation of model generation and simulation into separate scripts.

2.5 Analysis of simulation results

To analyze simulation results, two different methods are used. For the manual approach, the distributions are compared according to the model rescuing concept, as described in Section 2.5.1. A particular application of the model rescuing concept that studies which models are closest to reaching a steady state is described in Section 2.5.2. The systematic approach calculates parameters systematically based on the predicted flux values from the models following minimal cut set knockout simulation. The systematic calculation is described in Section 2.5.3.

2.5.1 Model rescuing concept

It is the goal of this study to identify enzymes whose perturbation will increase the target flux and yield of a metabolic system. The "model rescuing concept" is a whole-ensemble simulation analysis method directed toward this end.

When a perturbation is simulated, each ensemble model gives a different prediction for the final values of the variables of interest, which include fluxes, concentrations, and yields. In perturbation studies, one looks for the enzymes which, when overexpressed, lead more models out of the whole set of models in the ensemble to shift toward increased yield and flux of the desired products. In effect, one would plot a histogram of yield or of flux for each perturbation simulated and look for

those enzymes that, when perturbed, would by themselves lead to a larger shift in the histogram distribution curve to the right (toward the higher level of yield or flux value).

Why can this approach be used to predict useful enzyme perturbations? Assume that some subset of models in the ensemble is capable of predicting the phenotype of the system under all perturbations. Those perturbations that lead to more favorable distributions are more likely to change the output variable of interest to a more favorable value for the accurate subsets of models and, therefore, for the actual cellular system.

This line of reasoning is at the heart of the model rescuing concept. The goal is to find those enzyme perturbations that affect target variables in a favorable sense for the largest fraction of models. "Adjusting target variables in a favorable sense" typically refers to increasing transport fluxes for desirable products and/or product-to-substrate yields and decreasing the concentration and rate of accumulation of metabolites that tend to accumulate, though other variable changes may be defined as favorable as seen fit by the investigator. Those perturbations that accomplish the most favorable adjustment as defined by the investigator are proposed to be the best candidates for experimental perturbation studies.

For the DAHP network presented in Section 2.1, the model rescuing method is applied in combination with perturbation analysis to determine whether single-enzyme perturbations can lead to discovery of the optimum set of simultaneous enzyme perturbations. First, single-enzyme perturbation analysis is performed on the network as described in Section 2.3.1 using 1500 models. The dynamic response of each single-enzyme perturbation is simulated for 2000 hours. The distributions for two variables of interest, the outward transport flux of interest and the product-to-substrate yield, are compared to their wild-type values. Only those perturbations that tend to increase the yield are considered to be beneficial. A function that quantifies the optimality of a perturbation based on the statistics of the distribution for each of the variables of interest could be constructed and used to select the most optimal

perturbation. This was not pursued here, however, because the initial distribution result immediately shows from inspection that network rigidity and robustness causes the network to be insensitive to any single-enzyme perturbation. As a result, no useful conclusions may be drawn from this set of single-enzyme perturbation studies (see Section 3.1).

2.5.2 Steady-state analysis method

This section presents an application of the model rescue method described in Section 2.5.1. In this method, the variable of interest is a metric that indicates how far a model is from steady state at the end of 2000 hours of simulation. A minimal cut set knockout is performed on the network first to repress the functionality of all elementary modes except those with the theoretical maximum yield, such that any model reaching a steady state will exhibit that yield. The goal, then, is to find those perturbations that, when performed in addition to the minimal cut set knockouts, enable more models to reach a steady state.

The key to this method is the minimal cut set knockouts because they enable the optimization process. There are usually multiple variables of interest that are sought to be improved simultaneously via enzyme perturbations. The minimal cut set knockout procedure described in Section 2.3.3 allows us to guarantee that the steady-state yield will be at the theoretical maximum. This provides one with a condition that must be met in order to maximize the yield.

A metric is needed to measure how close a model's final predicted flux distribution is to steady state. To derive a metric, we use the fact that any steady-state flux distribution lies in the null space of the stoichiometric matrix of a network. For any steady-state flux distribution column vector $\mathbf{v_{ss}}$, Equation 2.7 must be satisfied,

$$\mathbf{S} \cdot \mathbf{v_{ss}} = \mathbf{0} \tag{2.7}$$

where $\mathbf{S}$ is the stoichiometric matrix of the network, and $\mathbf{v_{ss}}$ lies in the null space of $\mathbf{S}$. Any deviation in the right-hand side of Equation 2.7 from zero indicates that the network is not at steady-state for a given flux distribution vector. Larger deviations in the right-hand side from zero indicate larger deviations of the flux vector from steady-state.

Consider the case where the flux distribution vector v is not a steady-state flux distribution. This case is described in Equation 2.8, with the right-hand side of Equation 2.7 being replaced by a non-zero vector $\mathbf{z}$.

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{z} \tag{2.8}$$

The magnitude of $\mathbf{z}$ represents how far a given flux distribution is from steady-state. As such, $\mathbf{z}$ is used to define s, a scalar steady-state error metric, as shown in Equation 2.9.

$$s \triangleq ||\mathbf{z}||_2 = \mathbf{z}^\top \cdot \mathbf{z} \tag{2.9}$$

Strictly speaking, any value for s other than zero indicates that a flux distribution is not at steady state. However, it can take a very long simulation time for a model to reach a true steady-state, so some leniency is practical for identifying those predicted flux distributions that are "close" to steady state. A threshold value is chosen so that any flux distribution with an s value less than the threshold value is considered to be at steady state. In this study, a threshold value of 0.05 mmol^2 gDCW^{-2} hr^{-2} is used. Appropriate threshold values are to be chosen based on the unit associated with the fluxes and the magnitude of the fluxes in the wild-type steady-state vector.

With the parameter s and its threshold value defined, the general procedure may now be described. An overview of the procedure is shown in Figure 2.7. An ensemble of models is first generated using the wild-type steady-state flux values reported in the literature. The ensemble models are then used to simulate the dynamic response of the network to the knockout of the minimal cut set of enzymes that leaves only the maximum-yield elementary modes intact. The resulting predicted flux vectors

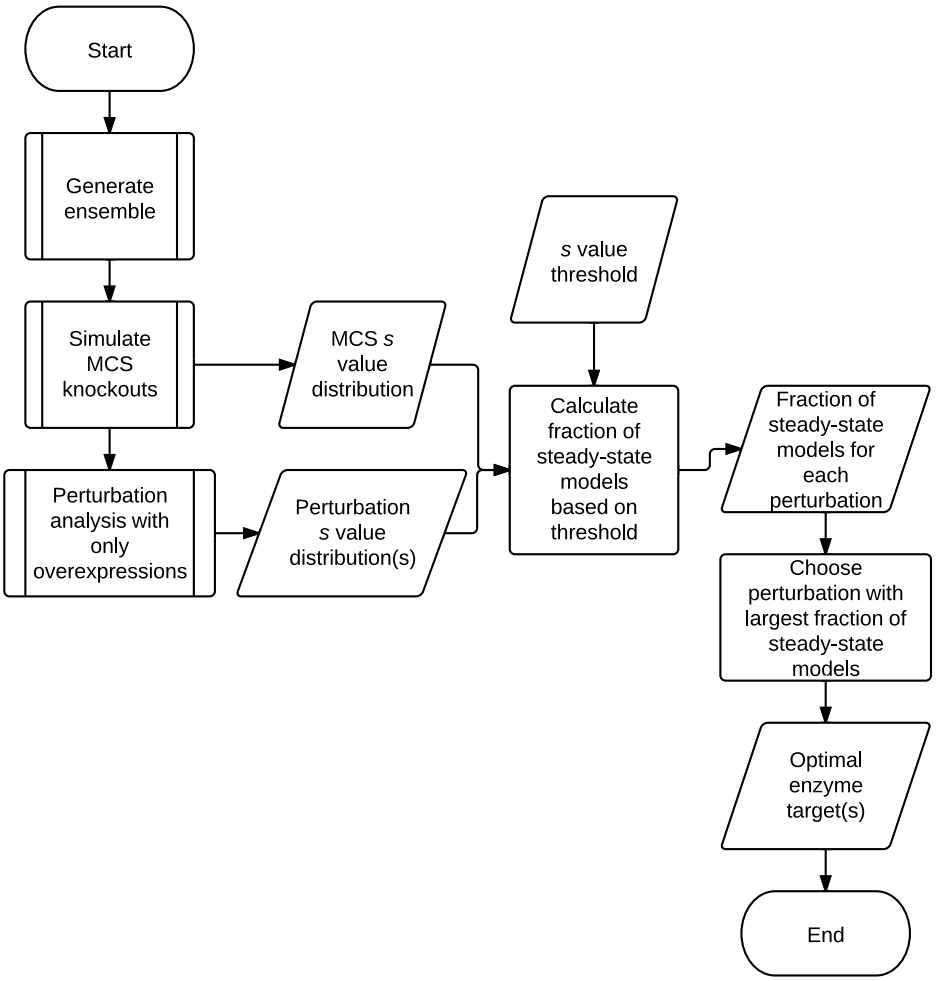

Figure 2.7: The general procedure for the steady-state analysis method.

at the end of simulation time are used to calculate the s values and determine their distribution, thus determining the fraction of models that reach steady state. Perturbation analysis with overexpressions is performed only on the enzymes that remain functional after the minimal cut set knockouts. Additional knockouts are not considered because they tend to shut down the network, reducing all fluxes to near-zero values. The fraction of models that reach a steady state by the end of simulation time is calculated for each perturbation, and the set of simultaneous enzyme perturbations that leads to the largest number of models reaching a steady state is predicted to be the most optimal.

Perturbation analysis, as required in the steady-state analysis method, will be computationally expensive for larger networks due to the large number of possible enzyme combinations whose perturbations must be simulated to find the optimal simultaneous perturbation set. For this reason, a method that yields a set of target enzymes from a smaller number of perturbation simulations is needed. Such a method is presented in Section 2.5.3.

2.5.3 Systematic enzyme targeting (SET) method

A systematic enzyme targeting (SET) approach to identifying target enzymes for overexpression that does not require a large number of perturbation studies is desired. Such an approach would be feasible for large networks, for which the computational intensiveness of iterating over the possible combinations of enzymes is prohibitive. For example, a network with only 40 enzymes has 780 possible two-enzyme combinations and 9,880 possible three-enzyme combinations. Our research has developed an approach that identifies enzyme overexpression targets after only a single perturbation simulation, namely, the minimal cut set knockout perturbation. This is done by first finding a flux distribution that generally represents the end-of-simulation flux distributions exhibited by the models after minimum cut set knockout. This representative flux vector is then projected onto the flux space

spanned by the maximum-yield elementary modes to give an ideal flux vector that is at steady-state and has the maximum theoretical yield. The differences between the representative and ideal flux values are calculated, and the largest element-to-element differences reflect which reactions are furthest from their target steady-state values, indicating that these reactions' fluxes need to be altered via enzyme perturbations. To quantify this deviation, a metric is defined that indicates the error between each representative flux value and its respective target steady-state flux value. It is reasoned that rate-limiting reactions will tend to decrease the flux of all downstream reactions, such that the error metric may appear large for non-bottleneck reactions that are downstream of the truly bottlenecked reactions. To account for this, each reaction's metric is compared to the metrics of immediately-upstream reactions to find those reactions where significant deviations from target steady-state fluxes are first manifested. These reactions are regarded as those that are most likely to be rate-limiting, and their enzymes are suggested as likely perturbation targets. The primary advantage of this method is that it uses only one enzyme perturbation simulation (the MCS knockout simulation) to simultaneously identify multiple enzyme targets for overexpression. This is much more efficient than perturbation analysis, which requires multiple perturbation simulations involving different combinations of enzymes in each perturbation study. When this method is applied to the *E. coli* DAHP production network, the method suggests as overexpression targets exactly the same enzymes reported in the literature to be effective in increasing DAHP flux and yield.

The method involves three main steps:

1. Finding the end-of-simulation flux distribution vector, $\mathbf{v_{rep}}$, that is representative of the whole ensemble of models' flux distributions after MCS knockout simulations are performed,

2. Determining the ideal steady-state flux distribution vector, $\mathbf{v_{ideal}}$, to which the representative flux vector is to be compared, and

3. Comparing the element-by-element differences between $\mathbf{v_{rep}}$ and $\mathbf{v_{ideal}}$ to suggest enzyme overexpression targets.

The first step is accomplished using singular value decomposition of a normalized matrix of end-of-simulation flux vectors for all the models in the ensemble. Let $\mathbf{F}$ be a matrix of fluxes, with rows representing reactions and columns representing models in the ensemble. First, each column vector in this matrix is normalized to a magnitude of 1 to prevent flux vectors with large flux values from falsely biasing $\mathbf{v_{rep}}$. The resulting matrix of normalized column vectors will be referred to as $\mathbf{N}$. Singular value decomposition is performed on this matrix to produce matrices $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$, whose relationship to $\mathbf{N}$ are shown in Equation 2.10,

$$\mathbf{N} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^{\top} \tag{2.10}$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices with orthonormal columns that represent the left and right singular vectors, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix with positive values arranged in decreasing order. These values are referred to the singular values of $\mathbf{N}$ and are representative of the amount of variance, or information, contained in $\mathbf{N}$ that lies along the direction specified by each successive singular vector. The first column of $\mathbf{U}$ is associated with the largest singular value of $\mathbf{\Sigma}$. Therefore, it is the vector that captures the largest amount of the variance in the fluxes of $\mathbf{N}$ and is, therefore, most representative of the set of fluxes present in $\mathbf{N}$. If the $\mathbf{N}$ matrix is effectively rank 1 (i.e., almost all the variance in the system lies along one direction), this first principal component vector is sufficiently representative of the fluxes in $\mathbf{N}$ and is chosen to be $\mathbf{v_{rep}}$. One can check the Scree plot of the system to ensure the effective rank of $\mathbf{N}$ is 1. The Scree plot is derived from the singular values. The cumulative Scree plot shows the fraction of information captured by each successive singular vector's direction. A matrix that is effectively rank 1 will have a very large first singular value relative to all the other singular values. The ease with which one can check the amount of

information contained within the representative flux vector is an advantage of using the SVD method over using an arithmetic mean of the fluxes over the models.

Next, the ideal flux distribution vector $\mathbf{v_{ideal}}$ needs to be obtained. The ideal flux distribution vector needs to be a steady-state vector with the maximum theoretical yield of the network. All possible candidate vectors that meet these criteria reside in the vector space spanned by the maximum-yield elementary mode flux vectors. Of these candidate vectors, the perpendicular projection of $\mathbf{v_{rep}}$ onto the maximum-yield elementary mode space is chosen as $\mathbf{v_{ideal}}$. This is because the nearest ideal vector candidate to $\mathbf{v_{rep}}$ is more likely to require a small number of enzyme perturbations to be attained and is more likely to be a feasible flux distribution for the actual system.

To calculate $\mathbf{v_{ideal}}$, a projection matrix that projects onto the maximum-yield elementary mode space is first constructed. Let $\mathbf{E}$ represent a matrix comprised of the maximum-yield elementary modes, expressed as flux column vectors. Using singular value decomposition, $\mathbf{U_E}$, an orthonormal matrix with the same column space as $\mathbf{E}$, is obtained. The decomposition is shown in Equation 2.11.

$$\mathbf{E} = \mathbf{U_E} \cdot \mathbf{\Sigma_E} \cdot \mathbf{V_E}^\top \tag{2.11}$$

The projection matrix $\mathbf{P_E}$ is then calculated as shown in Equation 2.12.

$$\mathbf{P_E} = \mathbf{U_E} \cdot \mathbf{U_E}^\top \tag{2.12}$$

This allows for $\mathbf{v_{ideal}}$ to be calculated as shown in Equation 2.13.

$$\mathbf{v_{ideal}} = \mathbf{P_E} \cdot \mathbf{v_{rep}} \tag{2.13}$$

The next step is to compare $\mathbf{v_{rep}}$ and $\mathbf{v_{ideal}}$ to see which corresponding elements deviate the most from each other. To do so, a comparison metric $\mathbf{c}$ will be defined. Since the goal of this investigation is to find overexpression targets, it is natural to define the comparison metric such that it is an approximation of the degree of

overexpression required to increase a reaction's flux to its target steady-state value. This required degree of overexpression may be estimated for a reaction by calculating the ratio of the ideal flux of each reaction to its corresponding representative flux. Based on this estimation, $\mathbf{c}$ is defined as shown in Equation 2.14,

$$c_i \triangleq \frac{v_{ideal,i}}{v_{rep,i}} \tag{2.14}$$

where $\mathbf{c_i}$ is the ith element of $\mathbf{c}$, $\mathbf{v_{rep,i}}$ is the ith element of $\mathbf{v_{rep}}$, and $\mathbf{v_{ideal,i}}$ is the ith element of $\mathbf{v_{ideal}}$. Reactions with $\mathbf{c}$ values greater than 1 have smaller-than-ideal representative fluxes, while reactions with $\mathbf{c}$ values less than 1 have larger-than-ideal representative fluxes.

The values of elements in $\mathbf{c}$ show to what degree each reaction deviates from its ideal flux value. However, a reaction that has a significantly smaller-than-ideal flux is not necessarily a bottleneck that prevents the system from reaching a steady state. A true bottleneck reaction will not supply enough flux to downstream reactions to maintain the downstream reactions' ideal flux values. As a result, reactions that are downstream of the true bottleneck will also tend to have smaller-than-ideal fluxes. Because these non-bottlenecked downstream reactions will be rate-limited only by the bottleneck reaction, their deviation from ideal flux will tend to be to the same degree as that of the bottleneck reaction. With this in mind, one may develop a second metric, $\mathbf{l}$, that more effectively indicates truly bottlenecked reactions. To calculate $\mathbf{l}$, one must compare the $\mathbf{c}$ value of each reaction to the $\mathbf{c}$ values of reactions that are immediately upstream of it. "Immediately upstream reactions" refers to those reactions that have at least one product metabolite that is a reactant of the current reaction. To find these reactions, one needs to first identify the reactions immediately upstream of each reaction. This can be done by calculating the upstream reaction matrix, $\mathbf{S_{feed}}$, as shown in Equation 2.15,

$$\mathbf{S_{feed}} = \mathbf{S_r}^{\top} \cdot \mathbf{S_p} \tag{2.15}$$

where $\mathbf{S_r}$ is the stoichiometric matrix of the network's reactants (negative elements) with the products (positive elements) set to zero; $\mathbf{S_p}$ is the stoichiometric matrix of the network's products with the reactants set to zero. All non-zero elements of both matrices are set to 1. Element $\mathbf{S_{feed,ij}}$ indicates the number of products reaction j shares with reaction i. The calculation may be explained as follows. Row i of $\mathbf{S_r^\top}$ indicates whether each metabolite is a reactant for reaction i. If the jth position in row i has a 1, this indicates that the jth metabolite is a reactant for reaction i. Similarly, the jth component of column i of $\mathbf{S_p}$ indicates whether metabolite i is a product of reaction j. By multiplying these two matrices together as in Equation 2.15, the rows of $\mathbf{S_r^\top}$ are multiplied by the columns of $\mathbf{S_p}$ as an inner product. Each row-and-column multiplication gives the number of times a 1 appears in the same element position of both the row and the column. This translates to the number of metabolites shared between the reactants of the reaction represented by the row of $\mathbf{S_r^\top}$ and the products of the reaction represented by the column of $\mathbf{S_p}$. Therefore, element $\mathbf{S_{feed,ij}}$ indicates the number of metabolites that were indicated with a 1 in both row i of $\mathbf{S_r^\top}$ and column j of $\mathbf{S_p}$.

In this investigation, cofactors are not considered to be metabolites when finding immediately upstream reactions for a reaction. This is because cofactors are maintained at near-constant concentrations in the actual cellular system, and the effect of one reaction's varying flux rate on the concentration of a cofactor is likely to be very small. Therefore, reactions are unlikely to influence one another through the reaction linkages provided by cofactors. As such, the rows of $\mathbf{S_r}$ and $\mathbf{S_p}$ that are associated with cofactors are eliminated prior to the calculation of $\mathbf{S_{feed}}$.

With the elements of $\mathbf{S_{feed}}$ indicating which reactions need to be compared, the maximum difference in the value of c between a reaction and all immediately-upstream reactions is calculated for each reaction and stored in the vector l. The reactions along a pathway that exhibit large increases in their $\mathbf{c}$ value relative to the previous reactions are indicated by large values of $\mathbf{l}$. Reactions downstream of bottlenecks will tend to

have similar **c** values and, therefore, low **l** values. The element-wise calculation is shown in Equation 2.16,

$$l_i \triangleq \frac{c_i}{c_{i,min\ c}} \tag{2.16}$$

where l_i is the ith element of **l**, c_i is the ith element of **c**, and $c_{i,min\ c}$ is the smallest value of **c** of any reaction that is immediately upstream of reaction i. The minimum **c** value of all immediately upstream reactions is used so that the highest possible **l** value for each reaction is chosen. This has a particular advantage over averaging the **c** values over all immediately-upstream reactions. In the case where a reaction has a significantly higher **c** value than only some of the immediately upstream reactions, whether the reaction is rate-limiting or not is uncertain. If the reaction is not rate-limiting, including it among the possible enzyme targets would be confusing, but probably not a critical failure. Simulations of various combinations of suggested overexpression targets may reveal that the overexpression target is not an effective one. If the reaction is rate-limiting, however, it is important that it be indicated by the method, whether it impedes all or only some immediately-upstream reactions. Failure to indicate an important enzyme target among all possible targets for overexpression is likely to be a critical error at this point, since there is no way to detect or correct this error later. Therefore, averaging of **c** values over all immediately upstream reactions does not give a small enough value in the denominator of Equation 2.16, leading to a smaller **l** value, which increases the likelihood of failure to indicate critical bottleneck reactions.

Once the **l** vector is calculated, the values in the vector are sorted and ranked, and those enzymes with the largest values of **l** are suggested overexpression targets. The ideal ranking pattern is one where the first few **l** values are significantly larger than the rest that follow, indicating a relatively clear division point for the enzymes that need to be overexpressed.

This method is applied to the *E. coli* DAHP production network from Rizk and Liao (2009). An ensemble of 1500 models is generated and used to simulate the knocking out of a minimum cut set of enzymes that eliminates all but the highest-yield elementary modes. The minimum cut set is calculated using the CellNetAnalyzer package developed for MATLAB® (Klamt et al., 2007). Simulations are continued for 2000 hours, and the initial conditions are set to be the wild-type steady-state metabolite concentrations. Once the simulations are complete, the system is analyzed using the systematic enzyme targeting (SET) method described in this chapter. Section 3.4 reports the simulation and analysis results.

Chapter 3

Results and Discussion

3.1 Individual-Enzyme Perturbation Analysis

Before the SET procedure was developed during the course of this research, an obvious first step to finding effective enzyme targets for perturbation is to try a brute-force approach, iterating one at a time over the single-enzyme overexpressions and/or knockouts and analyzing the resulting target flux and yield distributions of the models in the ensemble. Model rescuing concepts are applied to judge the effectiveness of each of the attempted perturbations. No minimal cut set knockouts are applied before the single-enzyme perturbation studies are carried out. An ensemble of 1500 models is generated.

Despite its simplicity, this method fails to indicate target enzymes for overexpression or knockout. Most of the overexpressions have such a small effect on network behavior that it is difficult to determine whether any perturbation of a single enzyme has a more significant effect than any other. A sample of ten single-enzyme overexpressions that are representative of the general behavior seen among all 38 overexpressions is shown in Figure 3.1, which is a set of histograms showing the distributions of changes in the outward transport flux of DAHP. Figure 3.1 shows that most overexpressions result in little consistent change to the behavior of the

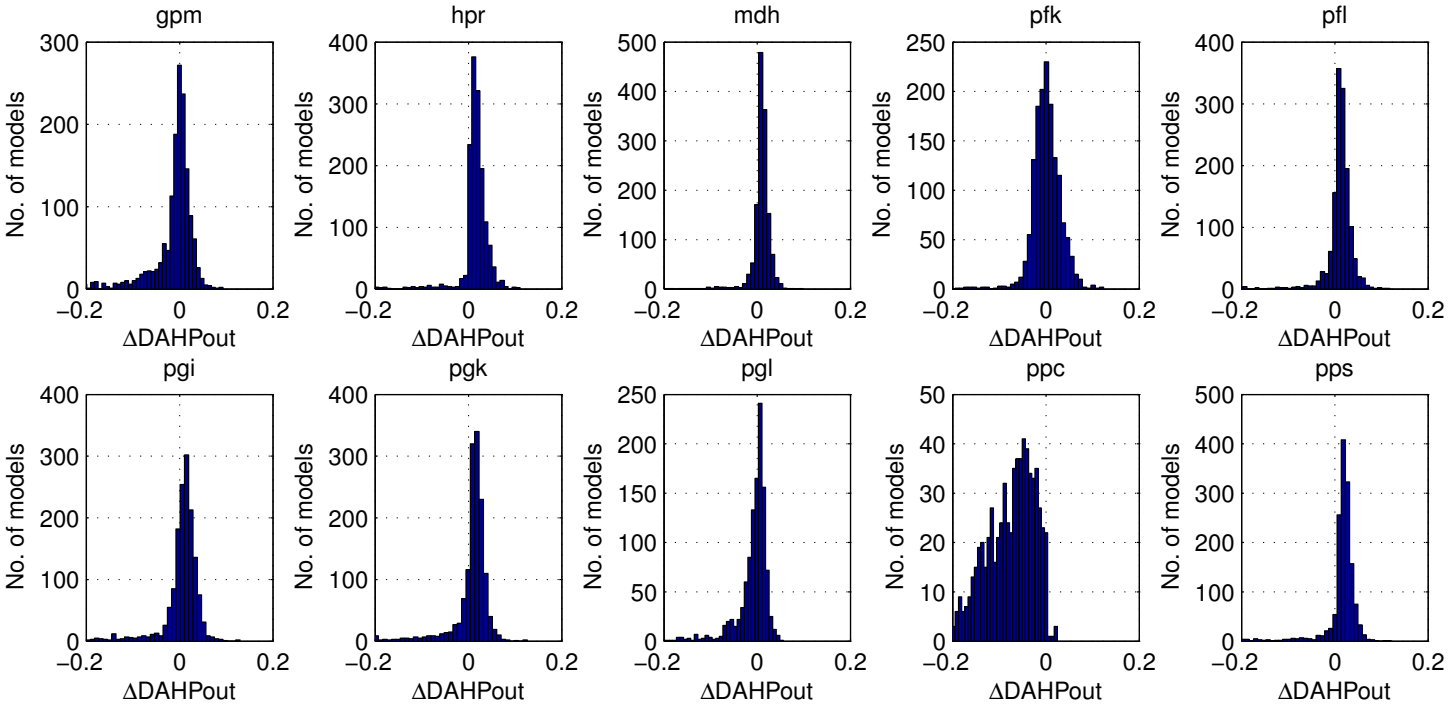

Figure 3.1: The number of models exhibiting various changes in DAHP outward transport flux after overexpression of the indicated single enzyme. Flux changes are in mmol gDCW^{-1} hr^{-1}. A total of 1500 ensemble models were used. Note that in all cases, the majority of the models did not increase the DAHP outward flux. Some single-enzyme overexpressions decreased DAHP flux, most notably *Ppc*.

network. Those that have a larger effect, such as *Ppc*, tend to decrease the DAHP fluxes of most models to near-zero values and shut down the network. What one desires to see in an effective perturbation is the overall distribution curve shifted to the right of the zero point to indicate a general increase in the final DAHP net transport flux. No single-enzyme perturbation distributions give this profile. Similarly, a significant number of single-enzyme knockouts, such as *RecATP* and *Tal*, shut down the network completely, reducing all fluxes of all reactions to near-zero values. Most other perturbations have little effect on the network behavior. A sampling of knockouts that is representative of the different behaviors observed is shown in Figure 3.2.

Some insight may still be gained from this approach, however. For example, as can be seen in Figure 3.2, the knockout of *Sdh* results in some increase in DAHP outward transport flux for most of the models. This is because *Sdh* knockout shuts down the TCA cycle pathway, decreasing outward succinate flux. Shutting down this side-product production pathway increases the production of DAHP, the product of interest, slightly. This type of information, however, is not worth the computational effort required to attain it. Single-enzyme perturbation analysis of the entire network can be computationally expensive when the network involves a large number of reactions.

One feasible way to avoid the problems seen with perturbation analysis of single enzymes is to try enzyme-subset perturbation studies and analysis instead, in which single-subset perturbations are simulated one at a time instead of individual-enzyme perturbations. The rationale is that enzymes in subsets of reactions limit the effects of their member enzymes being overexpressed individually and that all the reactions in the subset must be overexpressed simultaneously to significantly affect the network's behavior. This was attempted with the subsets of enzymes indicated in Figure 2.6, and the results were similar to those shown for the single-enzyme perturbations. Other approaches are obviously needed to effectively determine enzyme targets.

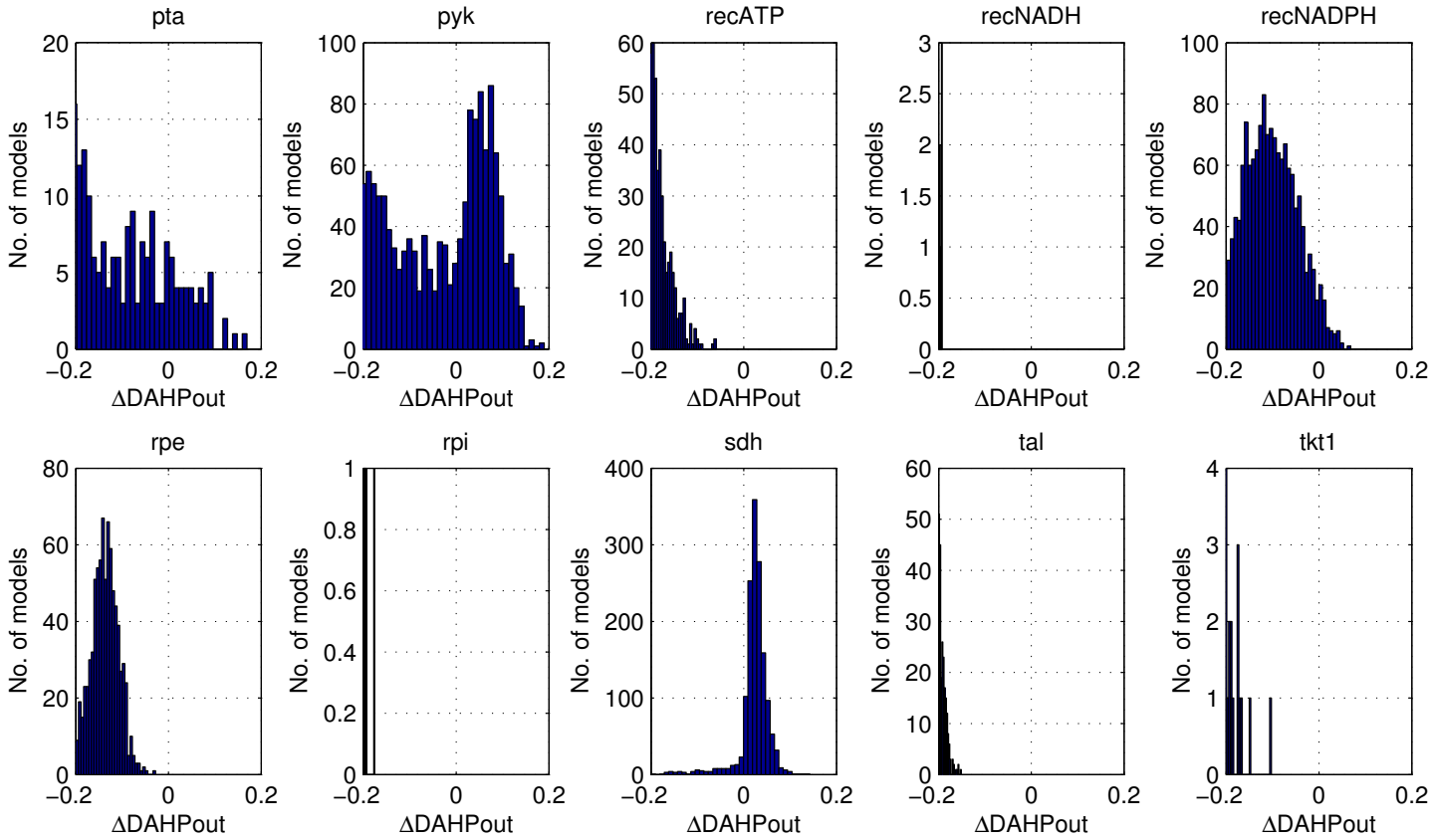

Figure 3.2: Distributions of the number of models exhibiting various changes in DAHP outward transport flux after knocking out the indicated single enzyme. Flux changes are in mmol gDCW^{-1} hr^{-1}. Note that network shutdown is predicted by most models for the knockout of *RecATP*, *RecNADH*, *Rpe*, *Rpi*, *Tal*, or *Tkt1*.

3.2 Minimal cut set knockouts

The above perturbation studies use the wild type as the starting point, with no knockouts being introduced initially. Minimal cut set knockouts should prove to be a more effective way to increase the yield of the network, since they force the network to operate via maximum-yield elementary modes. The minimal cut set knockouts chosen for this network are *Pfl*, *Ppc*, *Pyk*, and *Zwf*, as described in Section 2.3.3. The behavior of the network resulting from minimum cut set knockouts is simulated using the same ensemble of 1500 models used in Section 3.1. Figure 3.3 shows the distribution of changes in outward DAHP flux over the models for the base case that has only the minimum cut set knockouts. Figure 3.4 shows the distribution of yields resulting from the minimum cut set knockouts.

As Figure 3.3 shows, less than ten models predict an increase in DAHP outward flux. This is the expected behavior, since after the minimal cut set knockouts, there are fewer flux paths available through the network. Figure 3.4 shows that most of the models also predict a decrease in yield. Since only one realizable elementary mode remains after knocking out the minimum cut set, this must mean that the models are not reaching steady state by the end of the simulation (at 2000 hours). Figure 3.5 shows the distribution of s values over the models after MCS knockouts are performed. It verifies no model reaches a steady state, at which s would be 0. No models have an s value below the threshold of 0.05. This is an expected result also, since the kinetic parameters and enzyme concentrations of the reactions involved in the maximum-yield elementary modes are tuned to support their wild-type flux values, which differ significantly from some of the fluxes listed in $\mathbf{v_{ideal}}$. Overexpressions are needed to allow the system to reach a steady state.

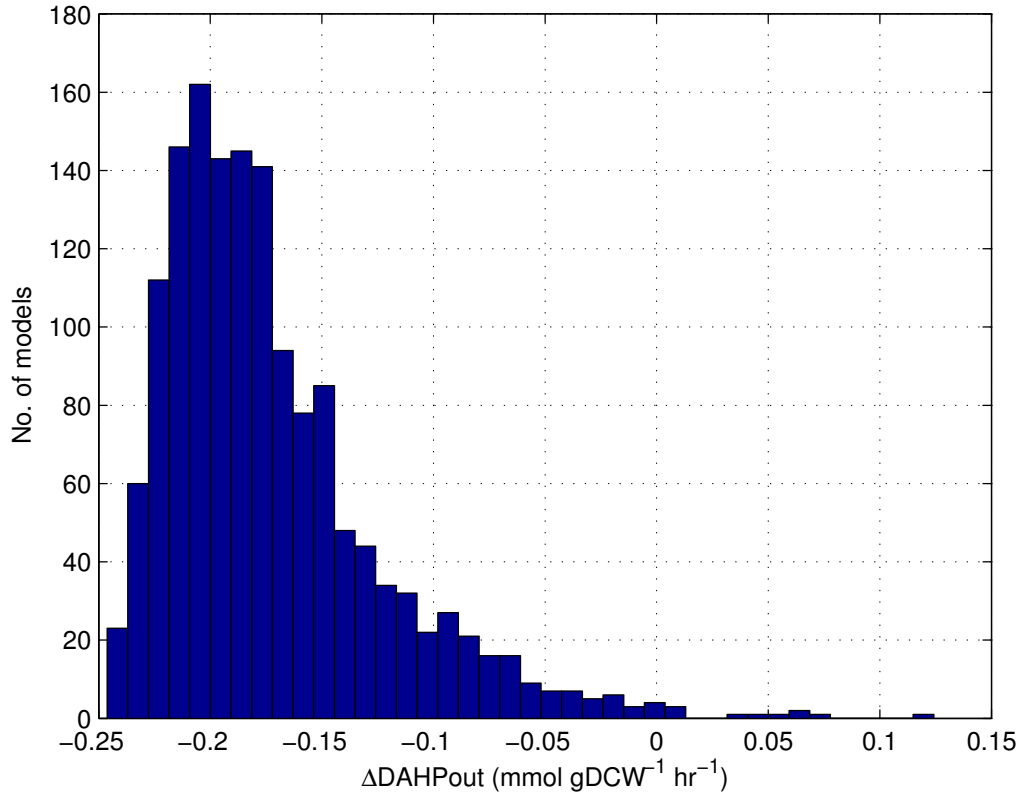

Figure 3.3: The distribution of the number of models exhibiting the indicated changes in DAHP outward transport flux after knockout of the minimum cut set enzymes. The wild-type DAHP outward flux is 0.26 mmol gDCW^{-1} hr^{-1}. Most models predict a marked decrease in DAHP outflux.

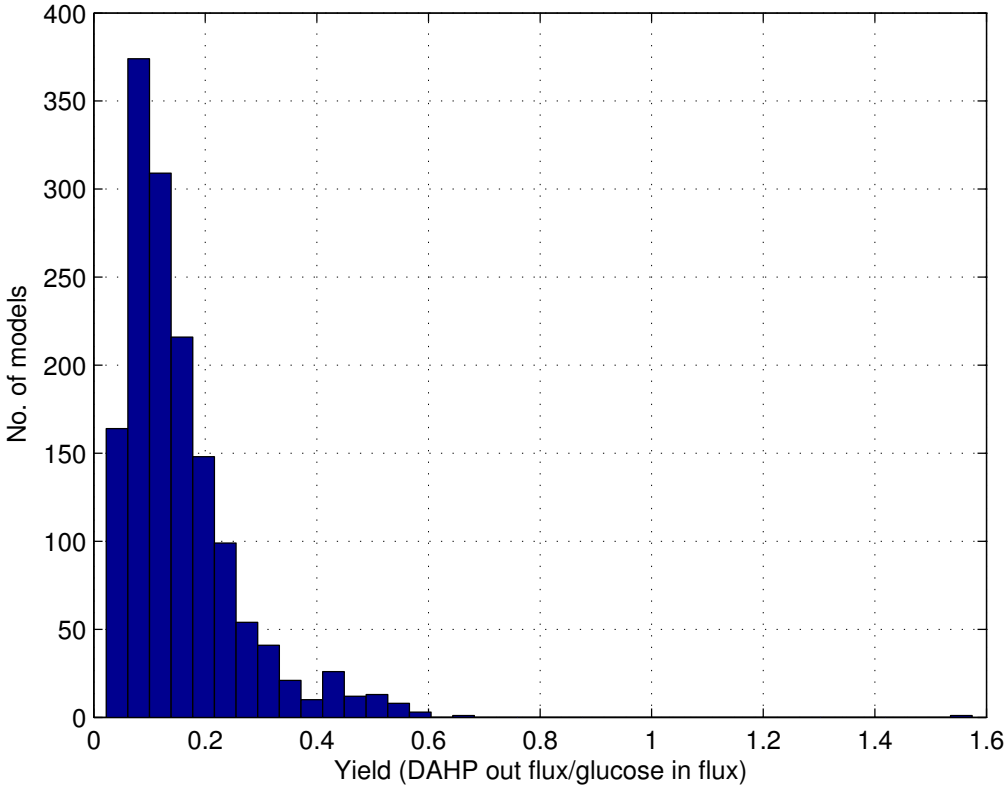

Figure 3.4: The distribution of the number of models exhibiting the indicated DAHP-to-glucose yields after knockout of the minimum cut set enzymes. The wild-type yield is 0.2. Some models show an improved yield beyond 0.2, but most predict a decrease in yield, meaning most of the models are not reaching a steady state.

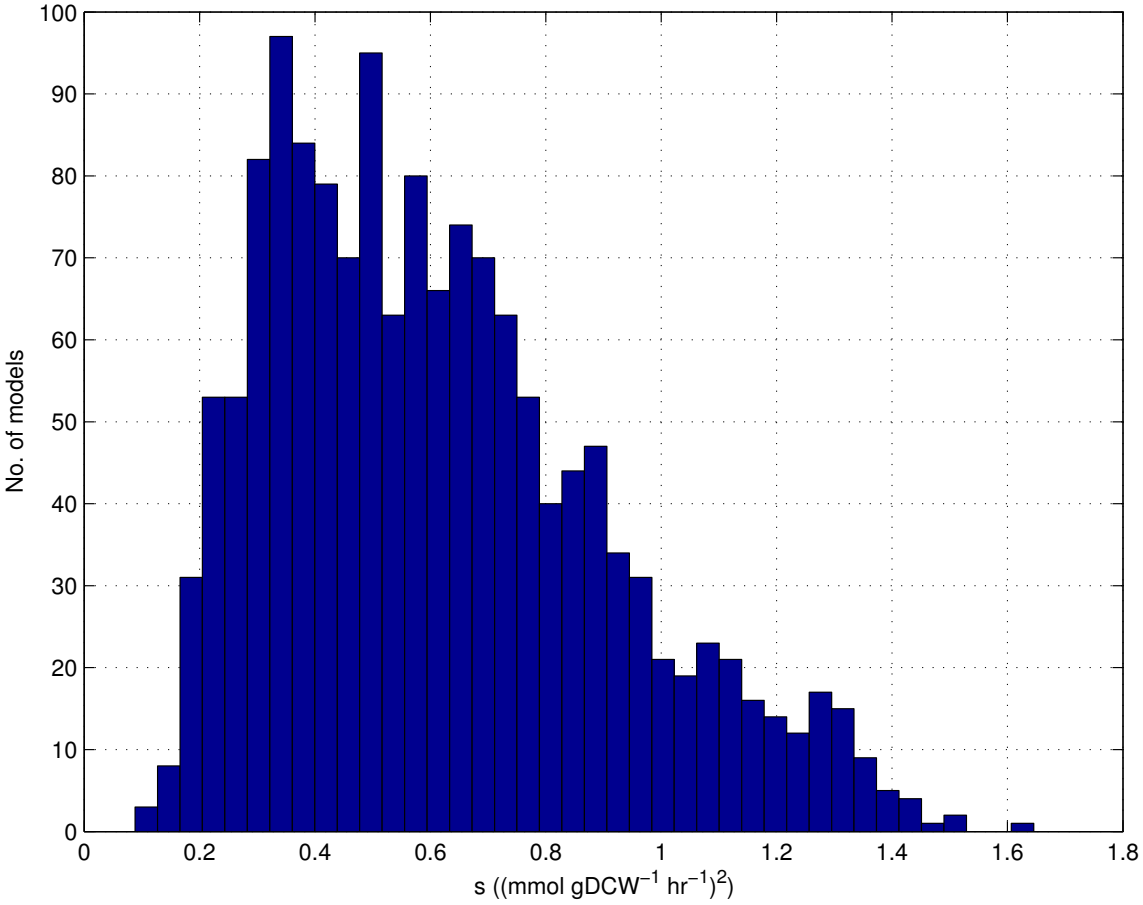

Figure 3.5: The distribution of s values resulting from the MCS knockouts.

Metabolite accumulation rates can help indicate the enzyme targets that will enable the network to reach a steady state, These rates may be attained by calculating the product of the stoichiometric matrix and the flux vector of any given flux distribution. Performing this calculation with the models' predicted flux vectors after 2000 hours reveals that pyruvate accumulates at a rate at least one order of magnitude larger than any other metabolite in the system for more than 95 percent of the models (data not shown). Figure 3.6 shows the distribution of rates of change of pyruvate amount over the models for the minimum cut set knockouts; it shows that almost all models have an accumulation rate greater than 0.2 mmol gDCW^{-1} hr^{-1}. Pyruvate accumulation is preventing the system from reaching a steady state and, therefore, the maximum yield, and reducing pyruvate accumulation should bring the models closer to the desired yield.

Other metabolites have also accumulated during the simulation time, but are no longer accumulating by the time 2000 hours have passed. These metabolites are not indicated by the accumulation rates, but by their concentrations. Concentration data are shown in Table 3.1 for the MCS knockout case after 2000 hours of simulation. Note that PYR, F6P, G6P, PEP, and S7P are accumulating significantly more than other metabolites. Also note the low concentration of E4P, one of the reactants for the production of DAHP. It is not immediately clear what could be causing these accumulations and shortages. Before hypothesizing, it is wise to perform perturbation analysis to gain additional insight.

3.3 Perturbation analysis with enzyme subsets after MCS knockouts

Perturbation analysis is performed with enzyme subset overexpressions on the network after MCS knockouts. The subsets used are shown in Figure 2.6. The resulting DAHP outward transport flux change, yield, and pyruvate accumulation rate distributions

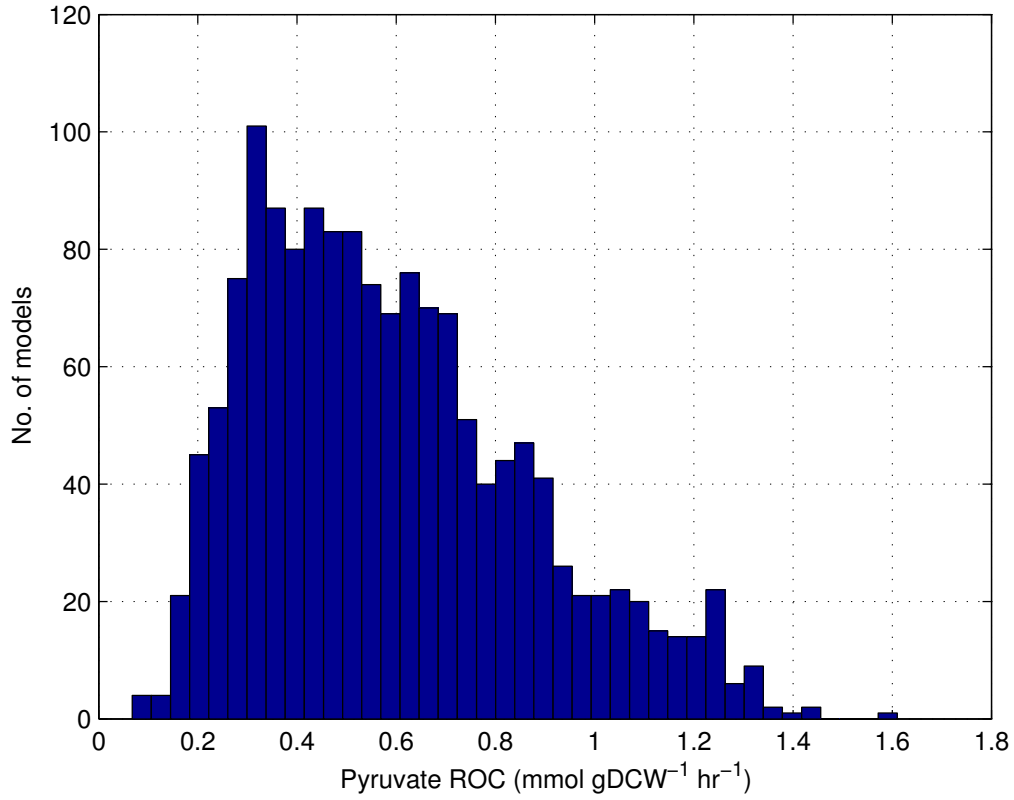

Figure 3.6: The number of models exhibiting various rates of change (ROC) in pyruvate amount after knockout of the minimum cut set enzymes.

Table 3.1: Mean concentration fractions after the minimal cut set knockouts are performed. A concentration fraction is the ratio of a metabolite's current concentration to its steady-state concentration. A metabolite at its steady-state concentration has a concentration fraction of 1. Concentrations higher than the steady-state concentration are represented by concentration fractions greater than 1. Significantly accumulated or scarce metabolites' concentration fractions are shown in red.

Metabolite	Mean conc. fraction
2PG	14.044
3PG	2.2509
ACCOA	0.010101
ACETATE	0.029204
ACP	0.010281
ADP	1.5563
ATP	1.3405
DAHP	0.33509
DHAP	0.63785
DPG	1.8684
E4P	0.12984
P1	1.3295
P2	2.9367
P3	7.2022
F6P	377.77
FDP	2.8495
FORMATE	0.029204
FUM	0.006903
G6P	91.275
GAP	1.9169
GLUCOSE	9.1959
MAL	0.005914
NAD	1.0641
NADH	0.95987
OAA	0.0058363
PEP	62.494
PGL	0.010854
PGT	0.035168
PYR	1505.2
R5P	1.0159
Ru5P	0.28664
S7P	40.073
SUCCINATE	0.016603
X5P	10.563
NADP	1.359
NADPH	1.1535

are monitored and shown in Figures 3.7, 3.8, and 3.9, respectively. The changes in DAHP flux are very similar across all the subset overexpressions, though the *Tkt2* subset has a slightly heavier right-hand tail on its distribution. In the yield distributions, some differences are noted among the subsets. The enzyme *Tkt2* is particularly notable, with a small group of models giving a larger yield around 0.4. This is not seen with any of the other overexpressions. Note that this improvement coincides with a slight increase in the number of models near the zero point in the pyruvate accumulation rates for *Tkt2* relative to the other subset overexpressions. This suggests that more intermediates are forming DAHP instead of accumulating as pyruvate.after *Tkt2* overexpression.

It appears that *Tkt2* is the best single-subset overexpression target for increasing DAHP flux and yield. This is consistent with experimental data, which indicate *Tkt* overexpression results in an increases in DAHP flux (Rizk and Liao, 2009). From the concentration data presented in Table 3.1, *Tkt2* overexpression seems to help increase the concentration of E4P, helping to relieve the shortage. Whether *Tkt2* is the best overexpression target cannot be decided with confidence, as the effects of its overexpression on the network seem to be small. It seems that more significant improvements to DAHP flux and yield will only arise from multiple simultaneous subset overexpressions.

There are a number of factors that could be preventing the system from achieving better yields and DAHP fluxes. The most likely explanation is that multiple subset overexpressions are required before the DAHP flux and yield can begin to increase. Another explanation is that F6P accumulation is limiting the rate of *Tal* by virtue of being one of *Tal*'s products. The accumulation may be due to the feedback inhibition of *Pfk* by the accumulated PEP. In addition, there could be a shortage of ATP, since two of the three ATP-producing reactions in the network were eliminated out by the minimal cut set knockouts.

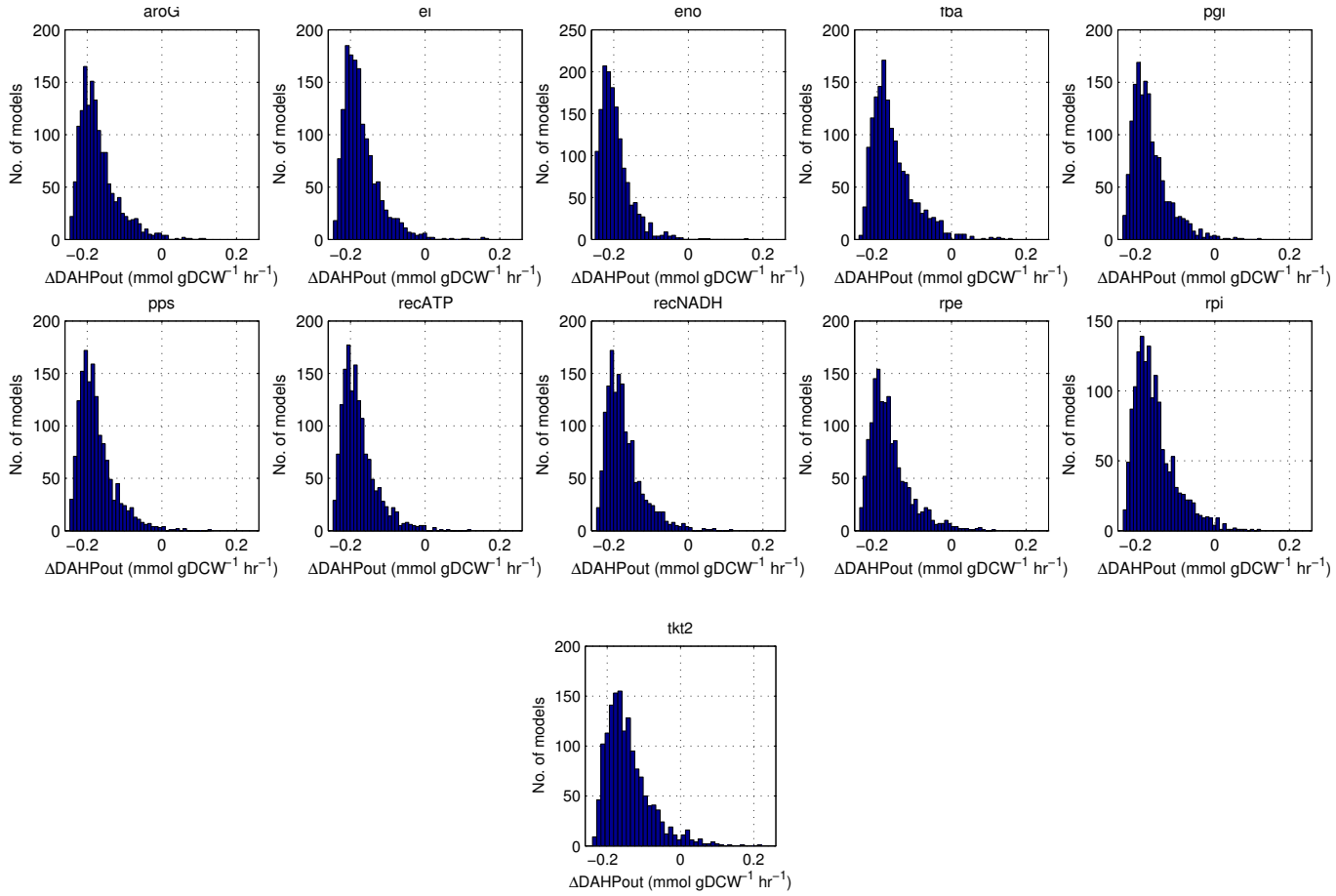

Figure 3.7: The distribution of the number of models exhibiting the indicated changes in DAHP outward transport flux after knockout of the minimum cut set enzymes and overexpression of the subset of enzymes containing the listed enzyme.

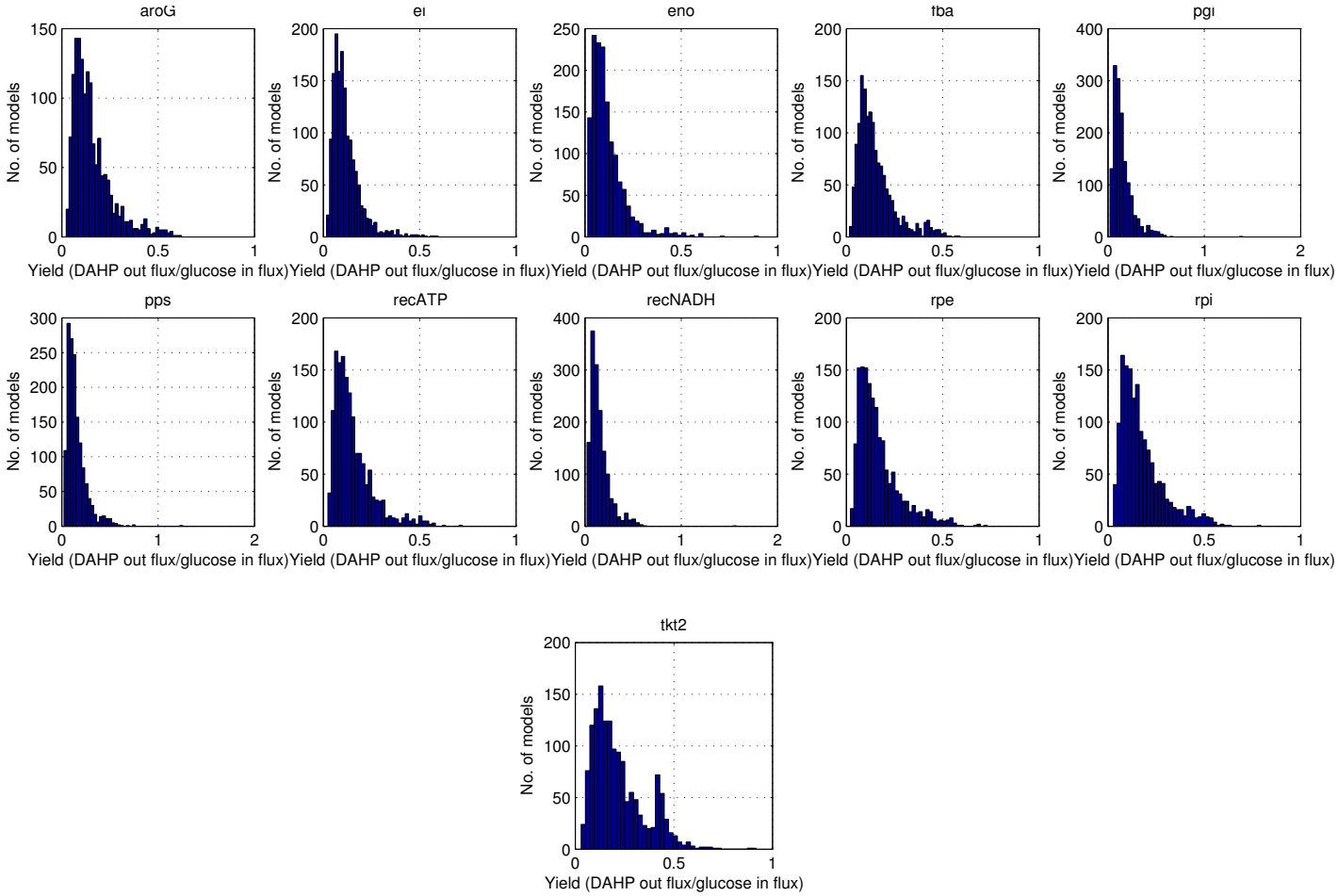

Figure 3.8: The number of models exhibiting various yields after knockout of all minimum cut set enzymes and overexpression of the enzyme subset containing the listed enzyme. The unperturbed yield is 0.2.

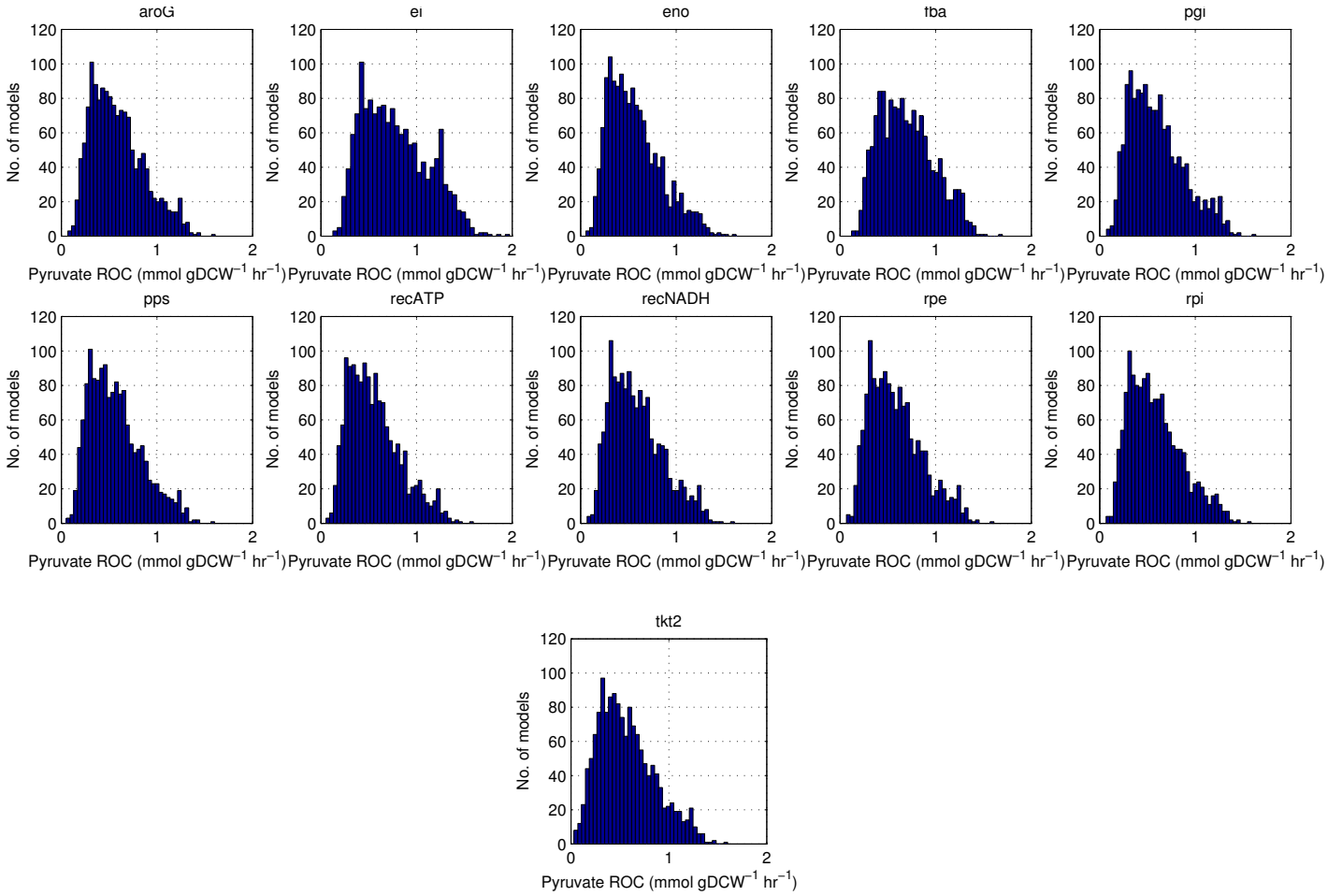

Figure 3.9: The number of models exhibiting various rates of change (ROC) of pyruvate molar amount after knockout of the minimum cut set enzymes and overexpression of the enzyme subset containing the listed enzyme.

All of these potential problems were tested using full-ensemble modeling and model rescue analysis. Removal of PEP's feedback inhibition of *Pfk* is tested by generation of another ensemble of 1500 models without the regulatory inhibition, and little difference is observed in the level of F6P accumulation or the yield and flux levels of the network under the perturbations tested so far. Additionally, restrictions on ATP and ADP concentrations were lifted with expectations that the artificial ATP sink reaction *RecATP* would run in reverse and provide additional ATP (see Section 2.2 for details on the concentration limitations imposed on cofactors). The results showed that the opposite case occurs; ATP shortages are much more common when the concentration limitations are lifted. The reaction *RecATP* has a forward flux in the wild type, and the concentration limitations limit rather than enhance the forward reaction rate. Constraining the concentrations of ADP and ATP to be constant constrains the Gibbs free energy for *RecATP* such that it does not have too large of a forward flux.

These results leave us to suspect that multiple overexpressions are what is primarily required. One possible approach is to test every combination of two or three enzymes or subsets. This would be very computationally expensive and prohibitive, so another approach is sought. An analytical method of determining the most likely enzyme candidates for overexpression is necessary to proceed.

3.4 Systematic enzyme targeting (SET)

3.4.1 Systematic analysis of DAHP network

An analytic method was developed as described in Section 2.5.3 and is used here to attempt to identify effective overexpression targets after the MCS knockouts are performed. Each model predicts a flux distributions after 2000 hours of simulation of the MCS knockouts. The flux distributions predicted by all models are normalized to a length of 1 and listed as columns in the matrix **N**. Singular value decomposition

is performed on $\mathbf{N}$ to determine the first left singular vector, which is a vector that spans the one-dimensional space that contains the largest amount of variance within $\mathbf{N}$'s flux vectors. This vector is referred to as $\mathbf{v_{rep}}$.

It is desirable that $\mathbf{v_{rep}}$ be representative of the fluxes predicted by all the models. For this to be true, the effective rank of $\mathbf{N}$ must be near 1, in which case most of the information contained in the flux vectors in $\mathbf{N}$ lies along one dimension. The effective rank of $\mathbf{N}$ may be determined by comparing the relative values of the singular values in a Scree plot. The cumulative sum of the squares of the singular values of $\mathbf{V_n}$ are plotted as fractions of the sum of squares of the singular values in the Scree plot in Figure 3.10. The plot shows that 99.4 percent of the information contained in $\mathbf{N}$ resides in a one-dimensional space spanned by its first left singular vector. This indicates that $\mathbf{v_{rep}}$ is highly representative of nearly all the models' flux behaviors.

After $\mathbf{v_{rep}}$ is obtained, the rest of the systematic procedure may be carried out. Table 3.2 shows the calculated values of $\mathbf{v_{rep}}$, $\mathbf{v_{ideal}}$, $\mathbf{c}$, and $\mathbf{l}$ for each of the reactions in the network, ordered by decreasing $\mathbf{l}$ value. It also shows the ratio $c_i : c_{input}$, where c_{input} is the $\mathbf{c}$ value of the inward transport reaction of interest. Recall the following details of the systematic method.

1. $\mathbf{v_{ideal}}$ represents the projection of $\mathbf{v_{rep}}$ onto the space of positive-coefficient linear combinations of maximum-yield elementary mode flux vectors. These maximum-yield modes are the only steady-state modes attainable by the system after the MCS knockouts. In essence, $\mathbf{v_{ideal}}$ is the feasible maximum-yield steady-state flux vector that lies nearest to $\mathbf{v_{rep}}$.

2. c_i represents the approximate level of overexpression required to increase reaction i's flux from $\mathbf{v_{rep}}$ to $\mathbf{v_{ideal}}$. Higher values of c indicate reactions that have fluxes significantly smaller than their respective ideal fluxes.

3. A high $\mathbf{l}$ value suggests a reaction that has a significantly larger $\mathbf{c}$ value than at least one immediately-upstream reaction. Therefore, a higher value of $\mathbf{l}$

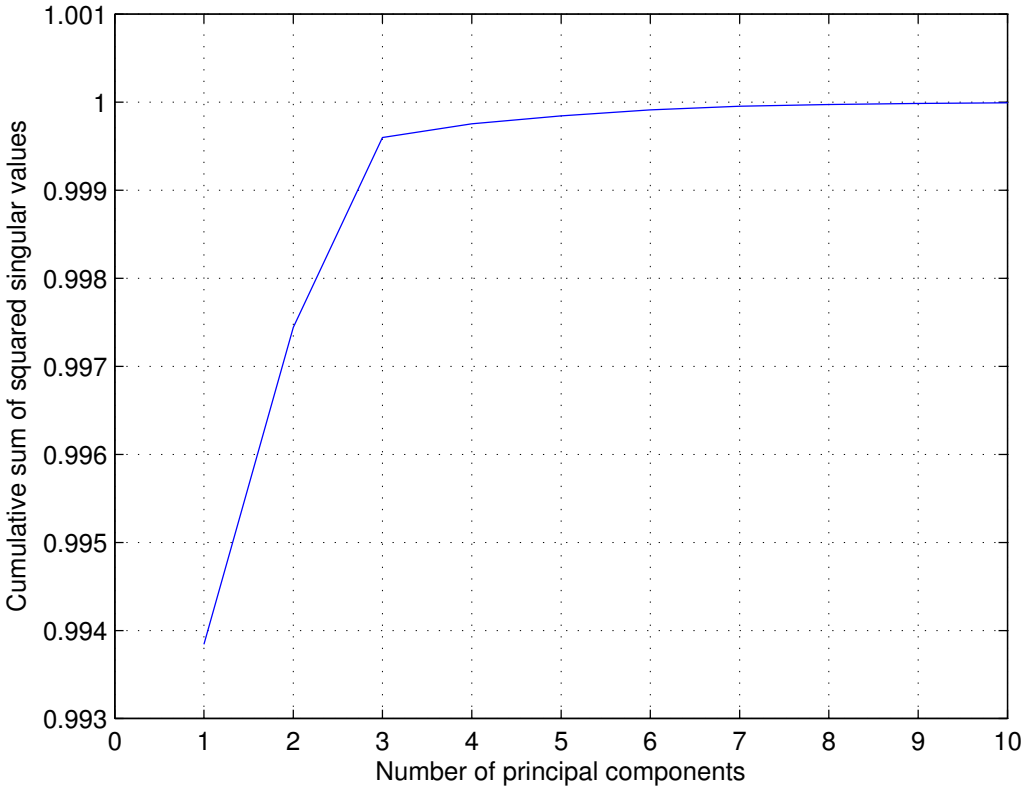

Figure 3.10: The cumulative Scree plot of the squares of the singular values of the normalized flux vector matrix resulting from simulation of the minimum cut set knockouts. Note that the effective rank of the normalized flux vector matrix is 1, as shown by the relatively large value of the first singular value relative to the other singular values. The projection of all columns of $\mathbf{N}$ onto the one-dimensional subspace of $\mathbf{N}$ spanned by the first left singular vector of $\mathbf{N}$ would capture approximately 99.4 percent of the variation in the matrix.

Table 3.2: The representative and ideal fluxes, **c** values, and **l** values for each of the reactions in the DAHP network, as determined by the systematic enzyme targeting method and ordered by decreasing l value. Note that the first four enzymes have **l** values much larger than the other enzymes.

Enzyme	$v_{rep,i}$	$v_{ideal,i}$	c_i	c_i/c_{input}	l_i
pps	0.0041022	0.21321	51.975	64.7650	64.556
tal	0.0071743	0.0609	8.4911	10.581	9.1677
aroG	0.039285	0.1828	4.652	5.7967	7.8121
tkt2	-0.032102	-0.1218	3.7953	4.7292	4.4578
ei	0.26482	0.2132	0.80512	1.0032	1.352
rpe	-0.013414	-0.0609	4.5415	5.659	1.1966
pfk	0.16456	0.1523	0.92548	1.1532	1.087
pgi	0.25043	0.2132	0.85138	1.0609	1.0574
tkt1	0.017966	0.0609	3.3908	4.2251	1.0119
eiibc	0.26482	0.2132	0.80513	1.0032	1.0032
eno	0.3069	0.1828	0.59549	0.74202	1.002
fba	0.16441	0.1523	0.92632	1.1543	1.0009
gpm	0.30751	0.18275	0.5943	0.74054	1.0001
pgk	0.30753	0.1828	0.59427	0.74049	1
recATP	0.16142	-0.1828	-1.1322	-1.4108	1
recNADH	0.30543	0.1828	0.59836	0.74559	1
glucose_in	0.26568	0.2132	0.80253	1	1
tpi	0.16443	0.1523	0.92621	1.1541	0.99987
eiia	0.26492	0.2132	0.80482	1.0029	0.99983
hpr	0.26488	0.2132	0.80496	1.003	0.9998
dahp_out	0.039303	0.1828	4.6499	5.7941	0.99955
rpi	0.01818	0.0609	3.3508	4.1753	0.73782
gap	0.30753	0.1828	0.59427	0.74049	0.64161

indicates a higher likelihood that the reaction's associated enzyme is a good overexpression target.

4. The general concept of the method is that those reactions that deviate downward more significantly from their ideal fluxes than immediately-upstream reactions are likely to be limited by slow enzyme kinetics. Overexpression of these enzymes should increase the yield and possibly the fluxes of interest of the system.

The method works surprisingly well, with the first four suggested enzymes being *Pps*, *Tal*, *Tkt2*, and *AroG*, the very same four enzymes shown by experiment to be good overexpression targets (Patnaik et al., 1995; Rizk and Liao, 2009).

Table 3.2 indicates an important effect of the MCS knockouts on network behavior. Note that *RecATP* has a negative **c** value, indicating that this reaction's direction after the MCS knockouts is reversed from its wild-type direction. The reaction associated with *RecATP* is an artificial sink reaction for ATP, which allows the model of the system to react excess ATP to form ADP to maintain equal concentration of both cofactors. In the wild-type flux distribution, more ATP is produced than ADP, so *RecATP* converts the excess ATP to ADP. This direction is the forward direction of the reaction. However, after the MCS knockouts, *RecATP* is forced to reverse direction, converting ADP to ATP. This indicates that the cellular system may have problems with ATP shortages as a result of the MCS knockouts. Most likely, these shortages arise from the fact that the MCS knockouts disable two of the three ATP-producing reactions in the network, *Pyk* and *Ack*. One option for avoiding ATP shortages is to underexpress *Pyk* instead of knocking it out. The effect on yield would most likely be minimal, since adding *Pyk* would only add one elementary mode to the system. This mode is the futile cycle consisting of *Pyk* and *Pps*. Because it does not lead to carbon flux exiting the system via an undesirable side-product transport reaction, this mode would probably not cost much yield, though it might exacerbate pyruvate accumulation issues.

Simulations affirm the effectiveness of the overexpression of the suggested targets. The first three and four suggested enzyme overexpressions are performed (referred to as Groups I and II, respectively), in addition to the simultaneous overexpression of *Pps*, *Tkt*, and *Tal* (Group III) and *Pps* and *Tkt* (Group IV) studied by Rizk and Liao (2009). Note that since *Tkt1* and *Tkt2* are reactions governed by the same enzyme, they are both overexpressed in Groups III and IV. The purpose of Group III is to test that the addition of the overexpression of *Tkt1* does not interfere with the effect of overexpression of the other suggested enzymes.

Since $\mathbf{c}$ values are representative of approximate overexpression levels required to bring each reaction to its ideal flux, they are used to calculate the optimal overexpression level for each enzyme. Overexpression factors are used for each enzyme i that are equal to the ratio c_i:c_{input}, where c_i is the $\mathbf{c}$ value of enzyme i and c_{input} is the $\mathbf{c}$ value of the inward transport reaction. This ratio is chosen instead of c_i because the goal of the enzyme overexpressions is to redistribute the inward flux more optimally among pathways rather than attempt to reach an optimal flux state with a higher or lower inward transport flux. Decreasing the inward transport flux limits the outward transport flux that may be achieved. Increasing the transport flux may cause additional enzymes that are close to their maximum capacity before increasing the inward transport flux to become bottlenecks. These new bottlenecks would not yet have been revealed by systematic analysis and could adversely affect the target flux and yield of the system, obscuring the effects of the enzyme overexpressions being simulated. Using the raw c_i value for reaction i assumes that one is attempting to reach the state indicated by $\mathbf{v_{ideal}}$, which is likely to have a lower or higher inward transport flux than $\mathbf{v_{rep}}$.

Figure 3.11 shows the distribution of changes in DAHP flux resulting from each of the enzyme group overexpressions. Overexpressions were simulated to the levels indicated by c_i:c_{input} ratios in Table 3.2. As can be seen, most of the models still have a decrease in overall DAHP flux. This is because the first iterations of the systematic analysis process are primarily oriented toward increasing the yield of the network.

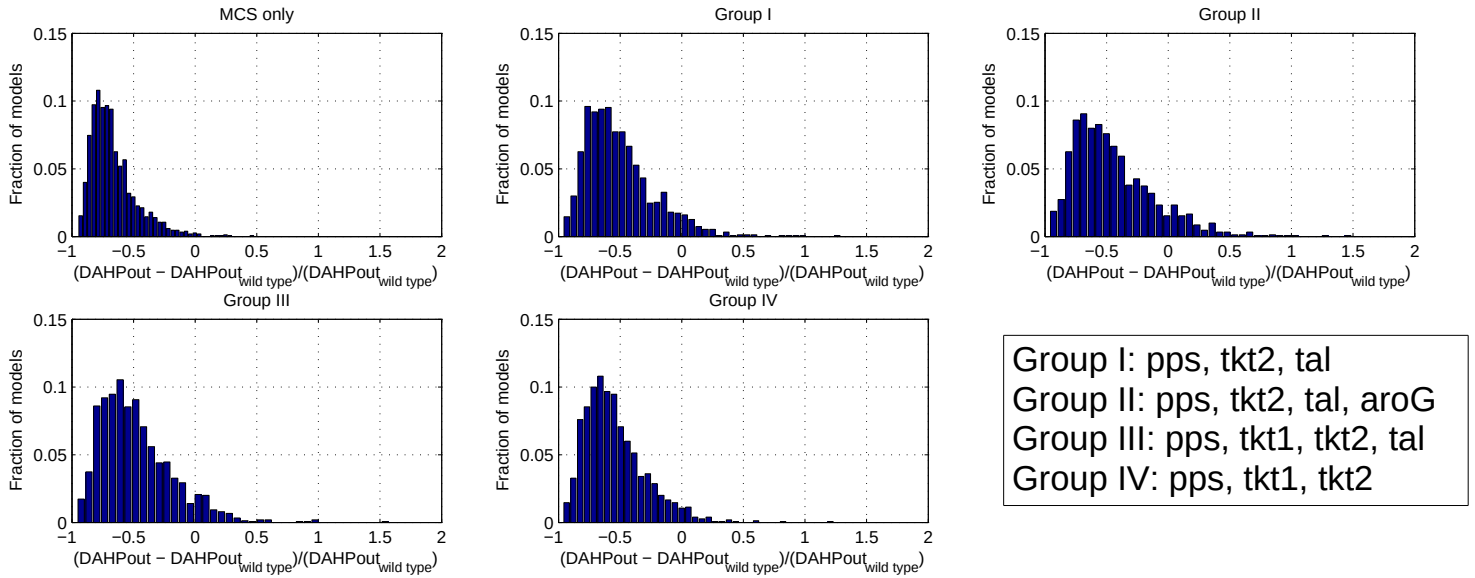

Figure 3.11: Distributions of fractional changes in DAHP transport flux from the wild-type flux after groups of enzymes are overexpressed in conjunction with MCS knockouts. The MCS distribution is also shown for comparison.

Methods for focusing on increasing absolute flux are currently being investigated and are described in Section 4.3. Figure 3.12 shows the distribution of yields after each group overexpression. Group II is especially effective at increasing the yield to near the maximum theoretical yield for a large percentage of the models (about ten percent of the models). The other perturbation groups are notably less successful, but still result in a significant increase in yield compared to the MCS knockouts alone. The pyruvate accumulation rates for the models are shown in Figure 3.13. All four perturbation groups reduce the level of pyruvate accumulation significantly. This is due to the overexpression of *Pps* in all four overexpression groups. Figure 3.14 shows the distributions of s values from each of the enzyme group overexpressions and compares them to the MCS-only distribution. Considering that no models were at steady-state before, the increase in the number is significant, especially for Group II. Group II's larger increases in yield and the number of models reaching steady state indicates that, in addition to *Pps*, *Tal*, and *Tkt2*, *AroG* overexpression results in additional improvement. This result is corroborated by reported experimental results provided by Patnaik et al. (1995) that state that *AroG* overexpressed with *Tkt* tends to increase DAHP flux and yield.

Perturbation of Group II's enzymes leads to the most favorable network behavior overall. Therefore, this perturbation is selected as the suggested perturbation for the next round of systematic analysis. Figure 3.15 shows the Scree plot of $\mathbf{V_n}$ after MCS knockouts and Group II overexpression. The models' flux predictions begin to diverge, leading to more scatter in the flux vectors. As such, $\mathbf{v_{ref}}$ now only accounts for 92 percent of the variance for this perturbation, as compared to 99.4 percent of the variance with only the MCS knockouts, as shown in Figure 3.10. Table 3.3 shows the $\mathbf{v_{rep}}$ and $\mathbf{v_{ideal}}$ fluxes, the $\mathbf{c}$ values, and the $\mathbf{l}$ values for each reaction after MCS knockouts and Group II overexpressions. The $\mathbf{l}$ values of the top-ranked enzymes have decreased significantly from the MCS knockout case (compare to Table 3.2), and no obvious outliers are evident indicating which additional enzymes are to be overexpressed. Since the top four suggested enzymes are the same four as were

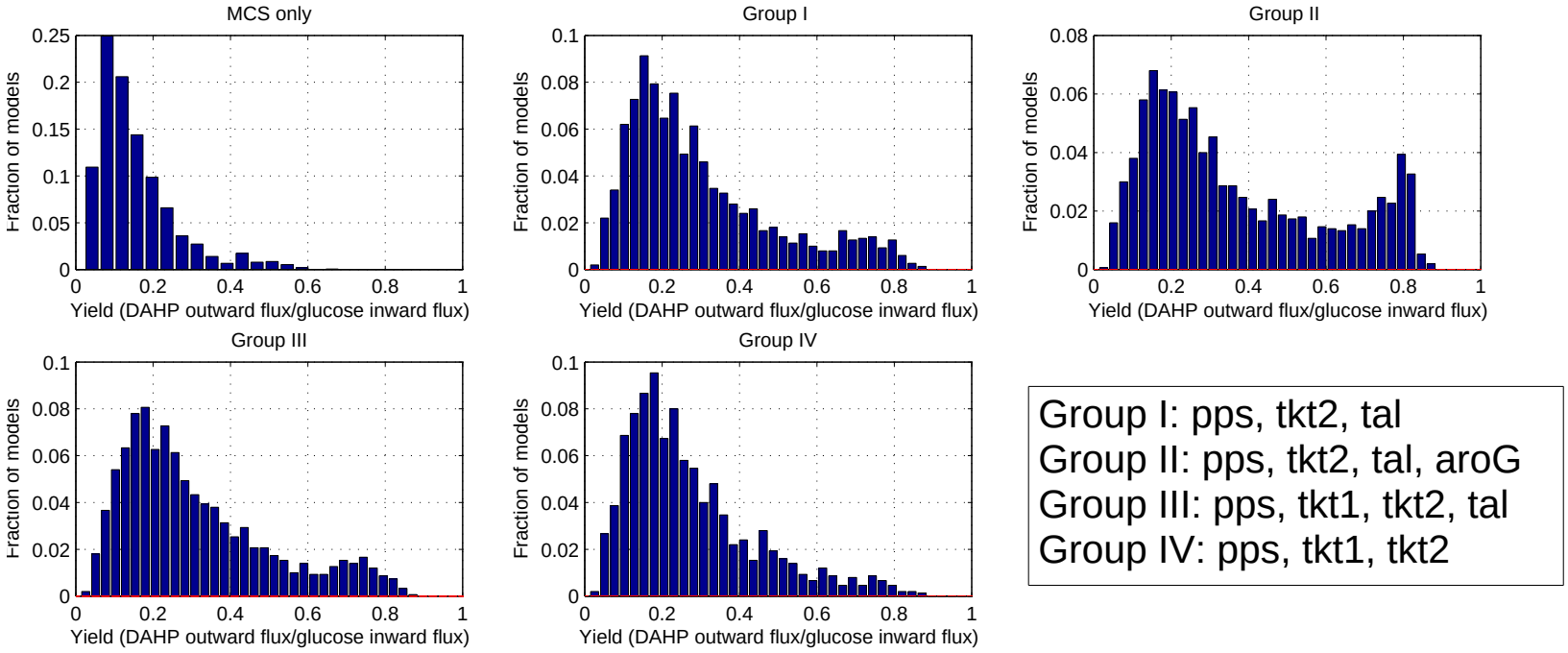

Figure 3.12: Distributions of yields after groups of enzymes are overexpressed in tandem with MCS knockouts. The MCS distribution is also shown for comparison. The wild-type yield is 0.2.

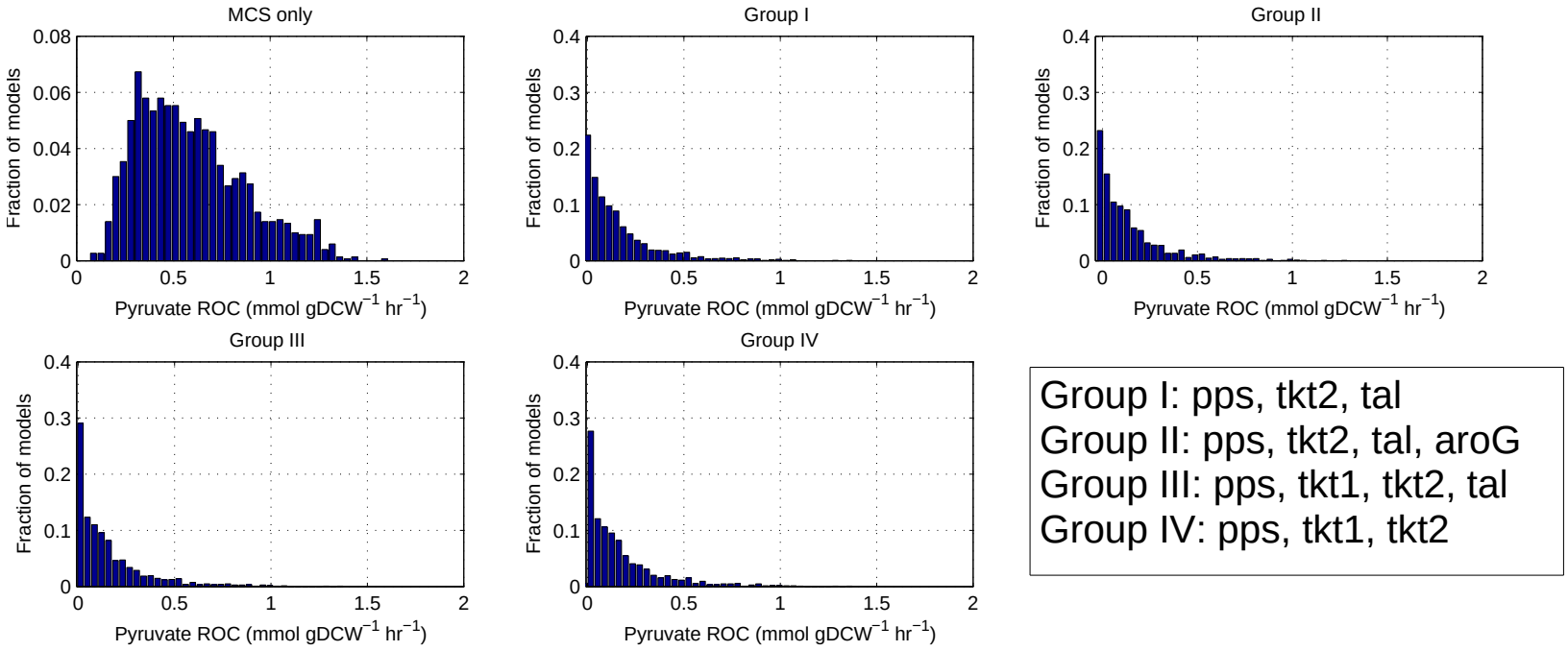

Figure 3.13: Distributions of pyruvate rates of accumulation (ROCs) after groups of enzymes are overexpressed in tandem with MCS knockouts. The MCS distribution is also shown for comparison.

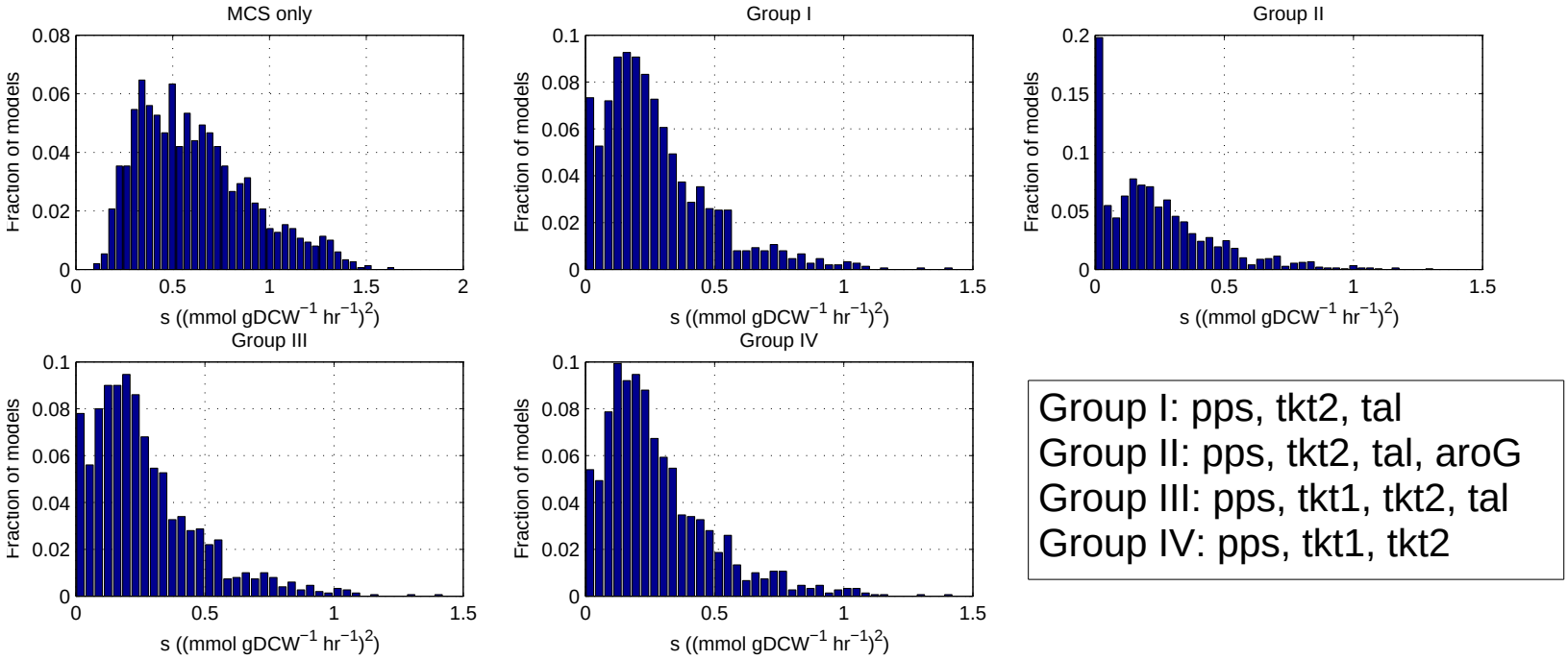

Figure 3.14: Distributions of s values after groups of enzymes are overexpressed in tandem with MCS knockouts. The MCS distribution is also shown for comparison. Note that approximately 20 percent of the 1500 models reach a steady-state after the overexpression of Group II.

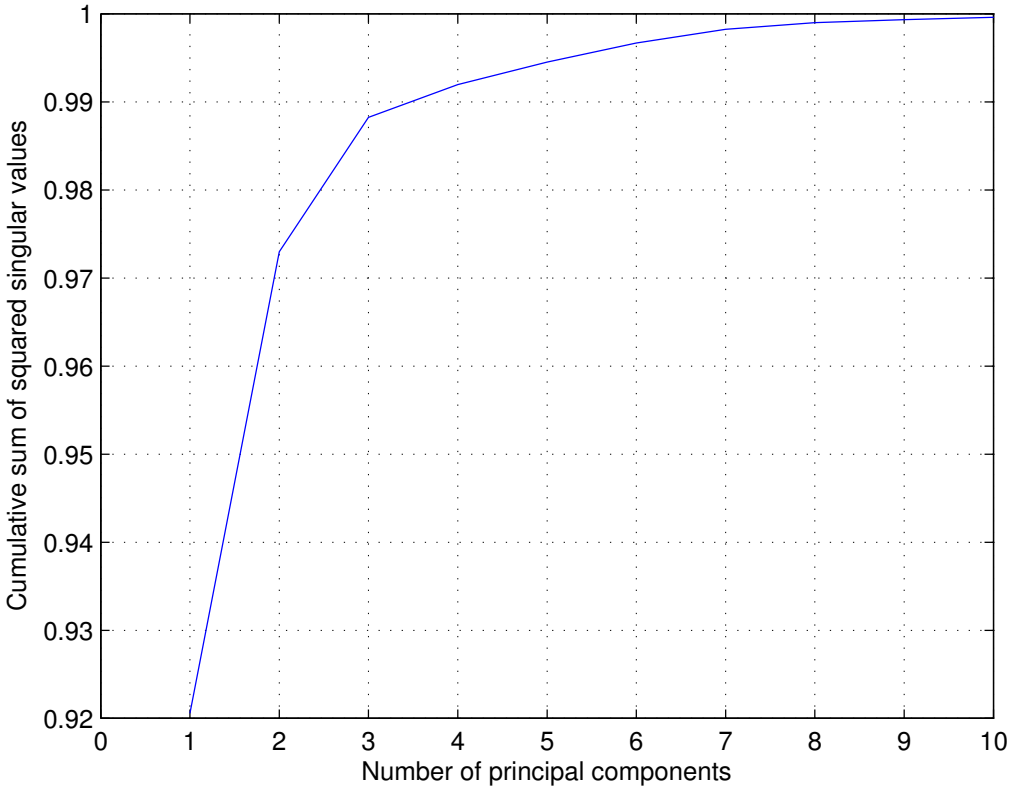

Figure 3.15: The cumulative Scree plot of the singular values of the normalized flux vector matrix resulting from simulation of the minimum cut set knockouts and the suggested perturbations from the first round of systematic analysis for the DAHP production network.

Table 3.3: Systematic fluxes and parameters resulting from the second round of perturbations for the DAHP network.

Enzyme	$v_{rep,i}$	$v_{ideal,i}$	c_i	l_i
tal	0.02921	0.070808	2.4241	2.0471
aroG	0.11571	0.21242	1.8359	1.9885
tkt2	-0.085519	-0.14162	1.656	1.9442
pps	0.20381	0.24783	1.216	1.5292
pfk	0.14948	0.17702	1.1843	1.3904
rpe	-0.036984	-0.070808	1.9145	1.1562
pgi	0.29096	0.24783	0.85175	1.0699
eiibc	0.31131	0.24783	0.79609	1.016
tkt1	0.044262	0.070808	1.5997	1.004
eno	0.23008	0.21242	0.92326	1.0007
eiia	0.31163	0.24783	0.79526	1.0002
gpm	0.23024	0.21242	0.92264	1
recATP	-0.089941	-0.21242	2.3618	1
recNADH	0.22785	0.21242	0.93229	1
glucose_in	0.31628	0.24783	0.78357	1
fba	0.14948	0.17702	1.1843	0.99998
hpr	0.31168	0.24783	0.79512	0.99998
tpi	0.14949	0.17702	1.1842	0.99992
pgk	0.23024	0.21242	0.92262	0.99976
dahp_out	0.11786	0.21242	1.8024	0.98176
ei	0.31168	0.24783	0.79514	0.86124
rpi	0.044441	0.070808	1.5933	0.83221
gap	0.23018	0.21242	0.92284	0.77932

previously overexpressed, it can be concluded that the optimal enzyme overexpression set has probably been found.

It is not clear that the overexpression levels indicated by the c_i:c_{input} ratios are the optimal overexpression levels for the system. Various scalar multiples of the Group II overexpressions are simulated, and the resulting DAHP outward flux and yield distributions are shown in Figures 3.16 and 3.17, respectively. For a doubling of the overexpression factors, network shutdown affects nearly 30 percent of the models, and higher levels of overexpression drive even more models to near-zero fluxes. As the DAHP and glucose fluxes near zero for many models, yield calculation becomes unstable. This results in a wide variety of predicted yields, even some orders of magnitude the theoretical maximum. From these data, one can conclude that the levels of overexpression indicated by the systematic method are scaled correctly.

3.4.2 Systematic analysis of toy network

To test the approach's generality, the toy network presented in Section 2.1.2 was analyzed using the same systematic approach. Figure 3.18 presents the Scree plot of the singular value decomposition of $\mathbf{V_n}$ for the toy network following the minimal cut set knockout of r_3. This shows that $\mathbf{v_{rep}}$ (the first left singular vector) contains 92 percent of the variance of the 1500 ensemble fluxes and is representative of the entire ensemble.

On calculating $\mathbf{v_{ideal}}$ for the toy network, it is notable that the projection of $\mathbf{v_{rep}}$ onto the maximum-yield elementary mode space reduces one of the elementary mode components to nearly zero. The two elementary modes for the toy network are shown in Figure 3.19, in which mode 7 is the zero-component mode. This may be a significant result because it suggests that this mode is difficult for the network to reach. In this case, this is explained by the fact that mode 7 calls for the reversal of direction of r_{6r} compared to its wild-type direction. From Table 2.3, r_{6r} has a wild-type steady-state flux of 0.45, which is significant compared to the mean flux of

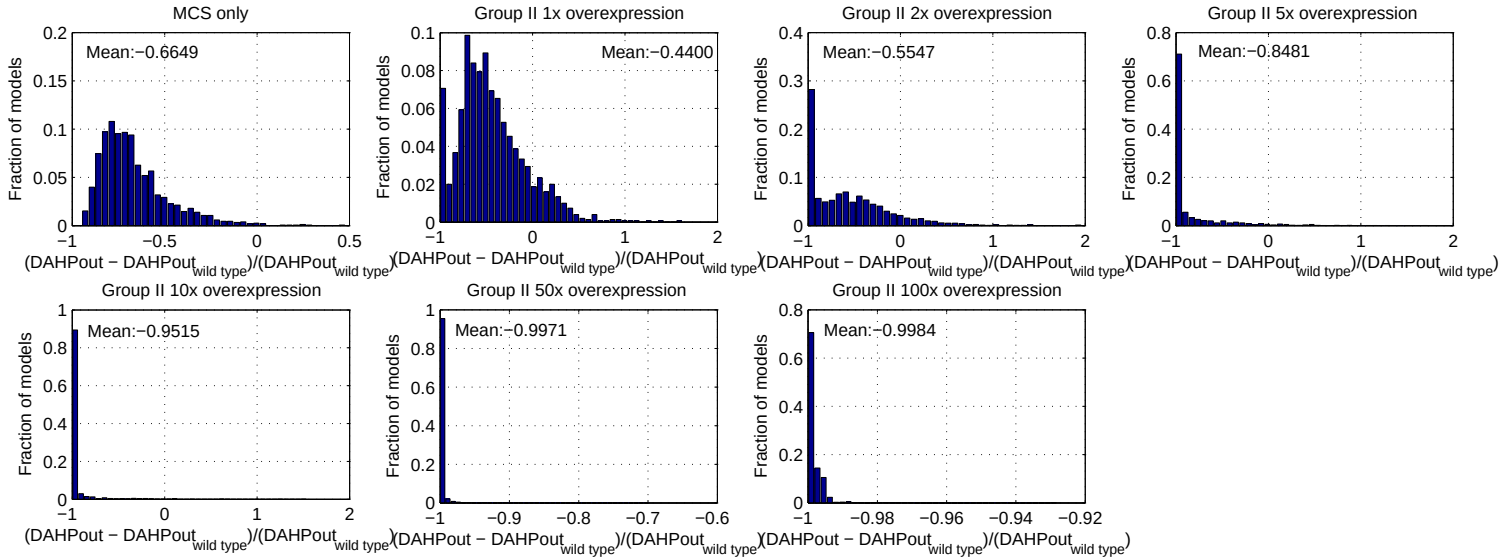

Figure 3.16: The distributions of fractional changes in DAHP flux relative to its wild-type flux for the indicated scalar multiples of the systematically-indicated overexpression factors for the Group II overexpression targets. The systematically-indicated overexpression levels were 65-fold for *Pps*, 11-fold for *Tal*, 5-fold for *Tkt2*, and 6-fold for *AroG*. Network shutdown begins to affect many models with doubled expression levels and worsens as overexpression levels increase.

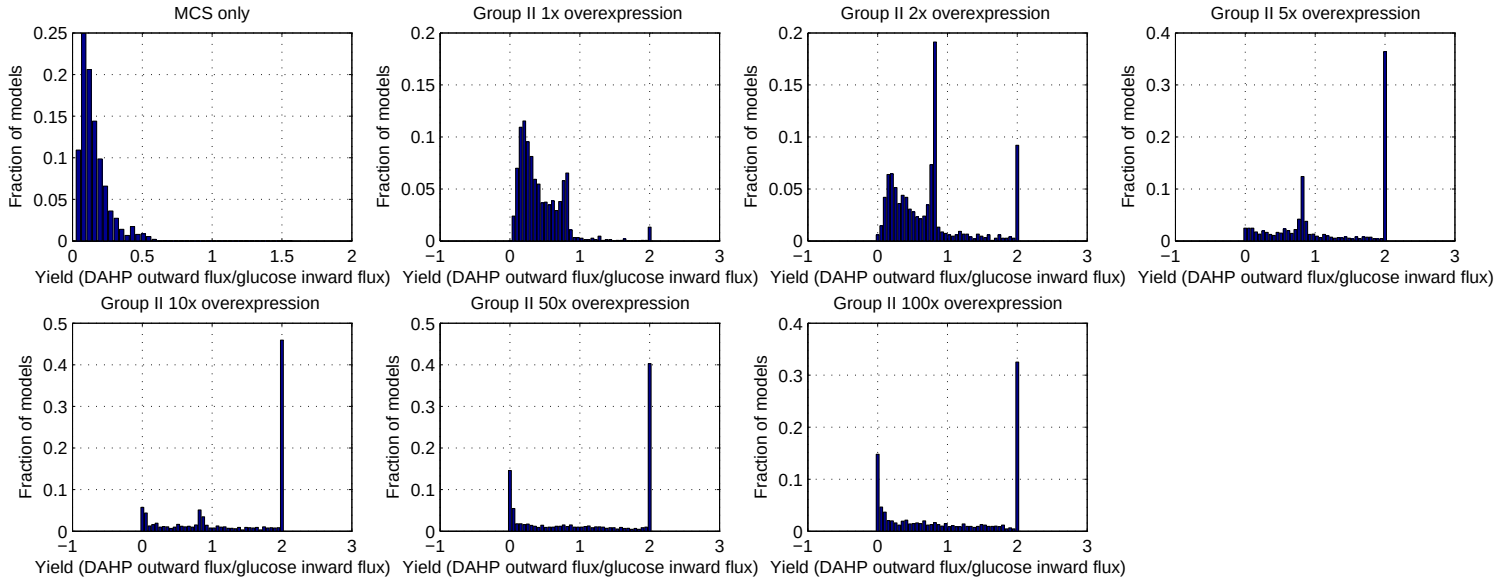

Figure 3.17: The distributions of DAHP-to-glucose yields for the indicated scalar multiples of the systematically-indicated overexpression factors for the Group II overexpression targets. The systematically-indicated overexpression levels were 65-fold for *Pps*, 11-fold for *Tal*, 5-fold for *Tkt2*, and 6-fold for *AroG*. Yields higher than the theoretical maximum are observed because of the numerical error resulting from dividing the near-zero fluxes predicted by models that exhibit network shutdown.

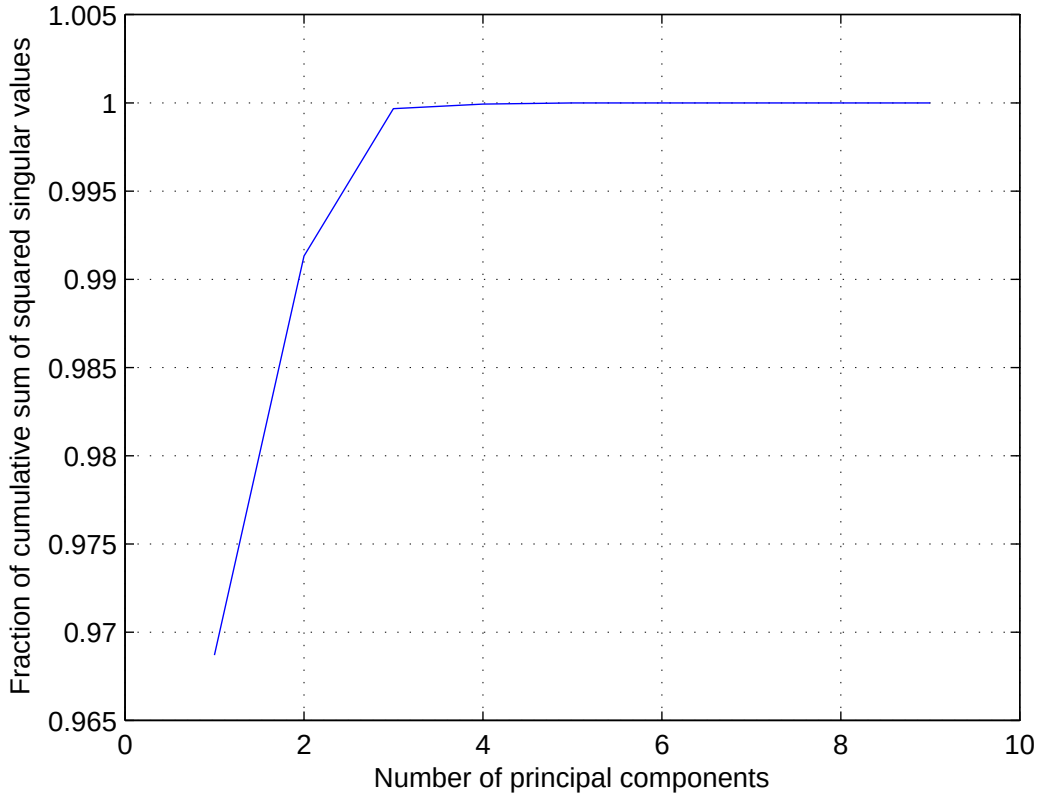

Figure 3.18: The cumulative Scree plot of the singular values of the normalized flux vector matrix resulting from simulation of the minimum cut set knockout for the toy network.

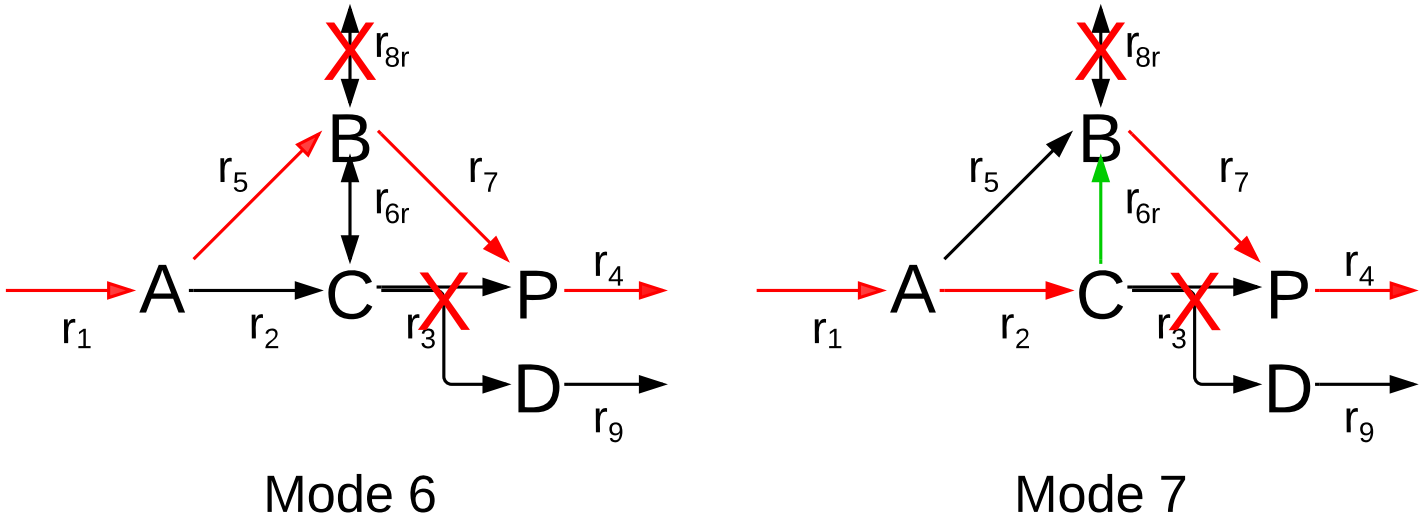

Figure 3.19: The maximum-yield elementary modes for the toy network are colored red. Minimal cut set knockouts to eliminate all other modes but these are shown. Note that in mode 7, r_{6r} (the reaction shown in green) must have a net flux from C to B, while the wild-type net flux direction is from B to C.

0.61. This wild-type flux is in the direction from B to C, while mode 7's direction is from C to B. This significant forward wild-type flux discourages the reaction from running backward toward B without significant concentration buildups of C.

The representative and ideal fluxes and the **c** and **l** values are calculated and presented in Table 3.4. Note that only the four reactions in mode 6 are listed. The only clearly suggested overexpression target following MCS knockout is r_7, since its **l** value is significantly larger than those of the other reactions. Reaction r_7 has an **l** value of approximately 1.47, while the other reactions' **l** values are near 1.

To test whether r_7 is indeed the ideal overexpression target, single-enzyme perturbation analysis is performed on the network after MCS knockout. Since the MCS knockout has been performed, only overexpressions will be analyzed. The distributions of changes in r_4 flux for each of the perturbations are shown in Figure 3.20, while the distributions of yields are shown in Figure 3.21. Flux and yield distributions for the MCS knockouts are also included in these figures for comparison.

It is clear that both the flux and yield distributions suggest the overexpression of r_7 after MCS knockout, since the distributions clearly lie further to the right for r_7 than for any other overexpression. These results confirm that the prediction made by the systematic method is effective in increasing the target flux and yield of the system, and confidence in the generality of the systematic enzyme targeting method is increased.

Table 3.4: Systematic method fluxes and parameters as calculated for the toy network with only minimal cut set knockouts. Note that one of the high-yield modes had a near-zero component, so its reactions are not included.

Enzyme	$v_{rep,i}$	$v_{ideal,i}$	c_i	l_i
r7	0.32067	0.36864	1.1496	1.4716
r5	0.47189	0.36864	0.78121	1.017
r1	0.47991	0.36864	0.76814	1
r4	0.65401	0.73728	1.1273	0.98061

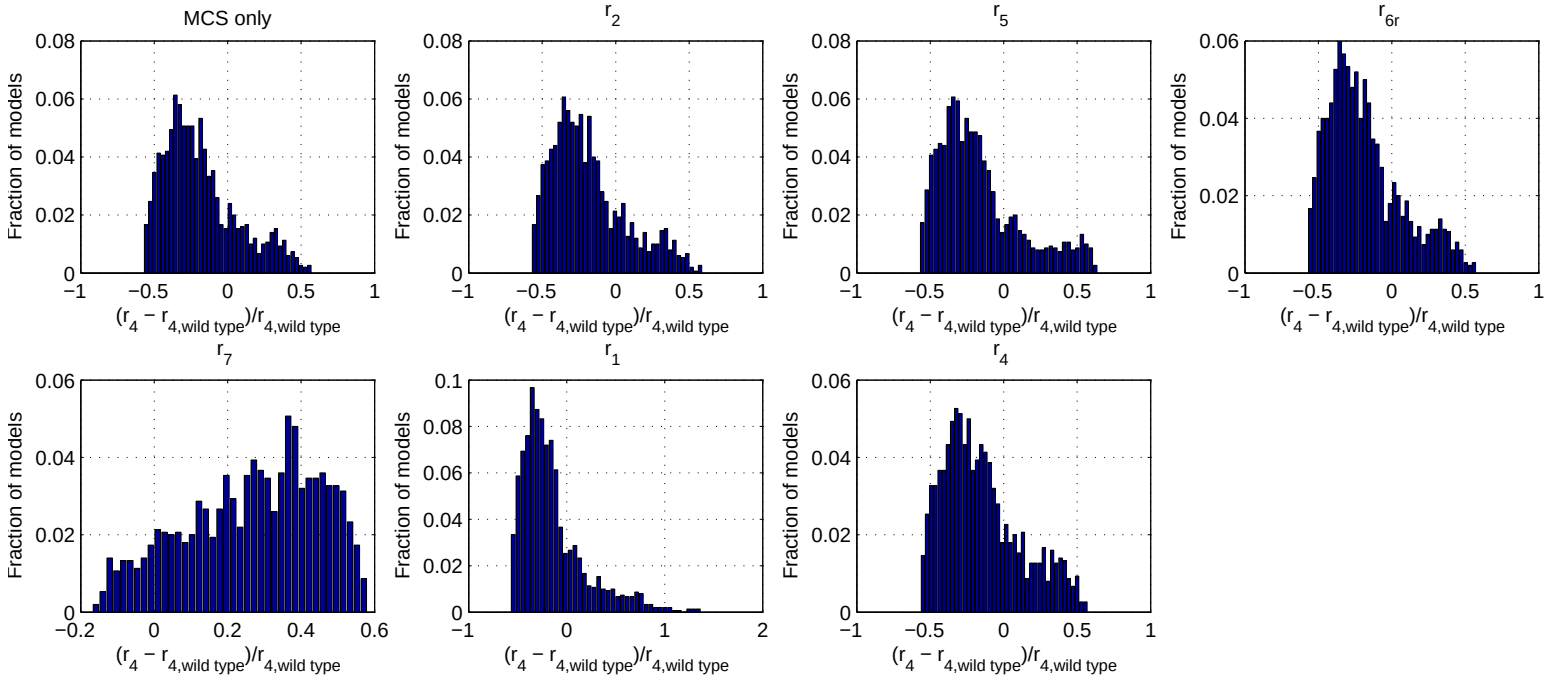

Figure 3.20: Distribution of changes in r_4 outward transport flux in the toy network resulting from the minimal cut set knockout and an additional overexpression. The wild-type flux is 1.25.

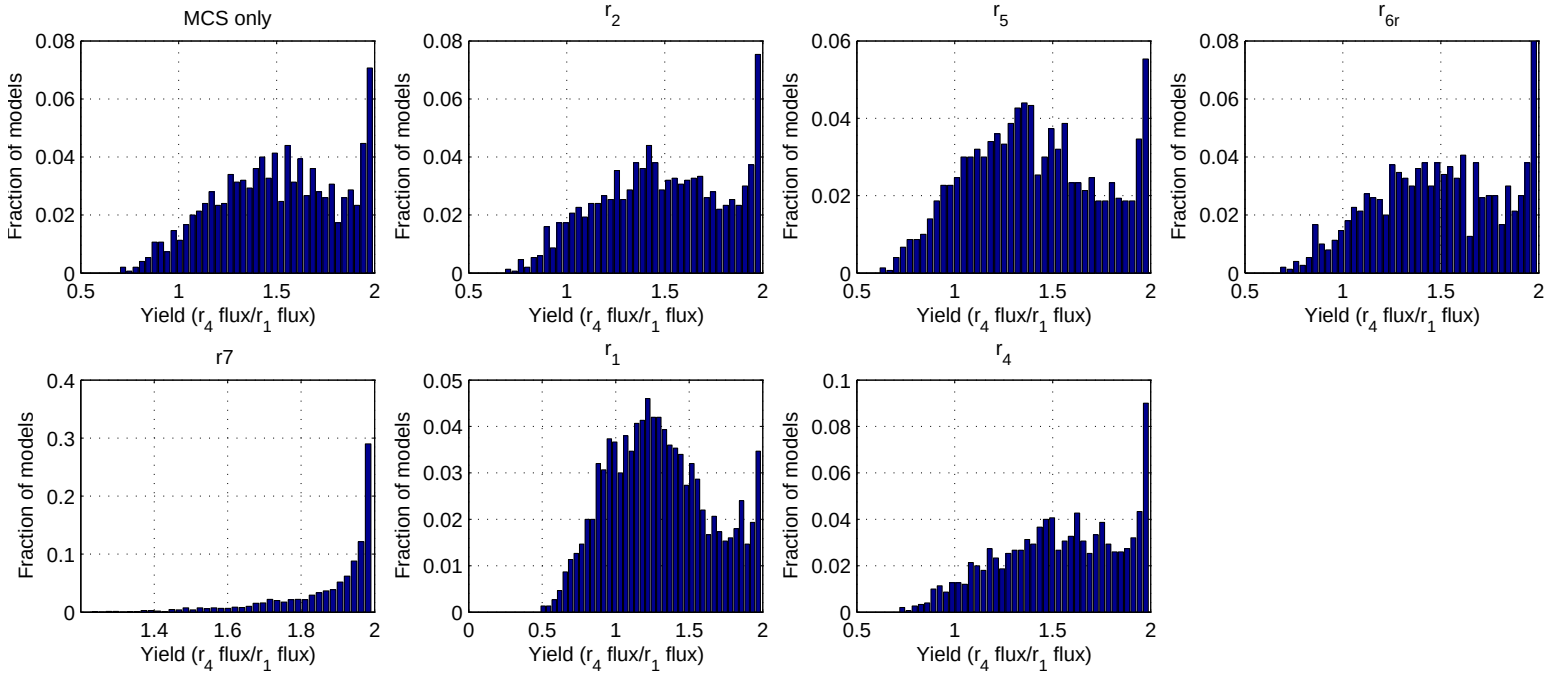

Figure 3.21: Distribution of r_4-to-r_1 yields in the toy network resulting from the minimal cut set knockout and an additional overexpression. The wild-type yield is 1.25.

Chapter 4

Conclusion

In summary, a systematic enzyme targeting (SET) method has been developed to identify enzyme overexpression targets and to estimate their respective levels of overexpression required to increase a target flux and yield of a metabolic network. The steady-state wild-type flux distribution of the system is the only experimental data required to perform the SET procedure. The method employs ensemble modeling to simulate the knocking out of a minimal cut set of enzymes that eliminates all but the maximum-yield elementary modes of the network. The SET method simultaneously identifies those enzymes whose overexpression brings the MCS-knockout network to steady state for the largest fraction of ensemble models, thereby increasing the yield of the network to the theoretical maximum for these models. Two network systems were analyzed using the SET method: (1) a toy network that demonstrated the concepts and feasibility of the SET method, and (2) a network representing DAHP production in *E. coli*. Results demonstrate the effectivenes of the methods. Enzyme targets provided by the SET method for the DAHP network exactly matched the overexpression targets found in the literature. Upon simulation of the SET method's suggested enzyme overexpressions, both networks were predicted to reach near-theoretical-maximum yields, and outward transport flux was increased for many models.

4.1 Evaluating the manual and systematic methods

Two primary approach philosophies were attempted in this study to predict enzyme targets for knockout and overexpression. First, a manual, human-judgment-driven approach was used that relied on subjective judgment and gaining an understanding of the system through the simulation data collected. Though this approach did work to some degree, it was slow and did not reveal the most effective solutions to the problem at hand. It does have a use in leading the investigator to a better understanding of the underpinnings of the network, such as how reactions tend to interact when perturbed and which metabolites accumulate. This sort of approach is especially useful in allowing the investigator to probe any desired concentration or flux value at any simulated time for any model. However, the large amount of data generated can be daunting, so concrete solutions may be difficult to ascertain. Additionally, this method relies heavily on perturbation analysis, which is typically computationally limited to single-enzyme or single-group perturbations. Attempting combinations of enzymes or groups of enzymes increases the number of required perturbations significantly, making these kinds of studies prohibitively expensive in terms of computational resources. This is particularly problematic for networks consisting of large numbers of reactions.

The systematic approach was highly effective and much more computationally efficient. Rather than requiring large numbers of simulations for perturbation analysis and hypothesis testing, the systematic method requires just a few simulations. Solutions are suggested with almost no human input involved, and the solutions seem to be very accurate and consistent with reported experimental results. Unlike the manual method, the systematic method can find combinations of enzymes simultaneously. Like any automated calculation method, however, one needs to be able to interpret the results. It is possible that some enzyme suggestions will not be feasible. For example, if the method were to suggest overexpression of *recATP* in the

DAHP network, one would need to realize that this is an artificial ATP sink reaction, and the suggestion of this enzyme as a target may indicate that the system needs additional ATP to maintain favorable yields and fluxes.

4.2 Problems with the methods

Some problems can be foreseen with the analysis of metabolic networks using whole ensembles of models. The primary problem is that it will not scale well into larger, whole-cell metabolic networks. The increase in the number of reactions and metabolites involved will increase the complexity of the simulations in three ways. First, the number of model kinetic parameters will increase dramatically. Second, the number of models will need to be increased in order to adequately sample the kinetic space. Third, the simulation time will need to be increased to allow the network to reach a point close enough to steady state to allow for analysis of long-term behavior of the network. There are no clear methods for improving the efficiency of this method. One helpful computational aspect of the method is that it adapts well to parallel computation, due to its number of repetitious and independent calculations.

A second potential issue is that the systematic enzyme targeting method seems to focus on increasing yields only, while absolute flux values are allowed to fall. This is to be expected, considering the method's procedure. The direction of a flux vector determines its yield entirely, and magnitude of the vector has no effect on yield. Magnitude does scale the absolute flux values, however. The normalization of the flux vectors in the calculation of $\mathbf{V_n}$ removes the magnitude component, biasing the method toward finding those overexpressions that increase yield instead of absolute flux. This normalization step is necessary, however, to prevent biasing of the relative sizes of the singular values in the singular value decomposition of the final matrix of flux vectors, which would lead to a false determination of the effective rank of the matrix.

4.3 Improvement opportunities

There are many opportunities for further development of the methods presented in this study. Perhaps the most important work that needs to be continued is testing the systematic enzyme targeting method on other systems. Its reliability has not been universally established, since it has only been attempted on two different metabolic networks. Additional networks may point out weaknesses and oversights in the method. In particular, attempting the method on a network with a biomass production component could yield interesting results. Other possible network elements to be considered are additional regulation and larger network sizes.

One possible improvement to the systematic enzyme targeting method involves adding concentration terms to the calculation of $\mathbf{c}$ in an attempt to find perturbation targets that increase flux values. A second stage of the systematic procedure could be introduced to attempt to maximize flux values without negatively affecting the yield. To guide developments toward this end, consider that a network at steady state will only exhibit metabolic buildups and inefficiencies through metabolite concentration imbalances. For a network does eventually reach a steady state, metabolites will tend to build up to some level at the slowest point in the network to force slow reactions up to the required rates for steady state. By finding these accumulation points and the reactions primarily responsible for them, targets may be suggested that are oriented toward increasing flux values. This consideration may be adapted to the systematic enzyme targeting method by adding concentration terms to the calculation of $\mathbf{c}$ such that larger values of $\mathbf{c_i}$ reflect larger metabolite concentration buildups for the reactants of reaction i.

One could also attempt solving the system ODEs algebraically for their steady-states instead of relying on dynamic simulation to reach a steady state. Such an approach has been attempted by Tan et al. (2011) and has the potential to reduce the computation time significantly, since dynamic simulation is the primary time bottleneck. The procedures used in this study, particularly the systematic method,

could be made into a MATLAB® toolbox and made easier to use and more user-friendly.

In summary, the development of the SET approach in targeting enzymes for overexpression in order to increase network performance, if shown to be applicable universally, would represent a significant advancement in metabolic network engineering. The method allows for investigators to avoid investing significant amounts of resources in performing a series of experimental studies directed toward achieving the same objective.

Bibliography

Bibliography

Contador, C. A., Rizk, M. L., Asenjo, J. A., and Liao, J. C. (2009). Ensemble modeling for strain development of l-lysine-producing *Escherichia coli*. *Metabolic Engineering*, 11(4–5):221 – 233. 5, 19, 20

Edwards, J. S. and Palsson, B. O. (2000). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology Progress*, 16(6):927–939. 3

Fell, D. A. (1992). Metabolic control analysis: a survey of its theoretical and experimental development. *Biochemical Journal*, 286:313–330. 4, 25

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407:651–654. 3

Kamp, A. v. and Schuster, S. (2006). Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, 22(15):1930–1931. 27

Klamt, S. (2006). Generalized concept of minimal cut sets in biochemical networks. *Biosystems*, 83(2–3):233 – 247. 5th International Conference on Systems Biology (ICSB) 2004. 30

Klamt, S., Saez-Rodriguez, J., and Gilles, E. (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, 1(1):2. 30, 45

Patnaik, R., Spitzer, R. G., and Liao, J. C. (1995). Pathway engineering for production of aromatics in *Escherichia coli*: Confirmation of stoichiometric analysis

by independent modulation of Arog, Tkta, and Pps activities. *Biotechnology and Bioengineering*, 46(4):361–370. 65, 68

Pfeiffer, T., Nu, J., Montero, F., Schuster, S., et al. (1999). Metatool: for studying metabolic networks. *Bioinformatics*, 15(3):251–257. 7, 27

Rizk, M. L. and Liao, J. C. (2009). Ensemble modeling for aromatic production in *Escherichia coli*. *PLoS ONE*, 4(9):e6903. 5, 12, 13, 15, 16, 20, 45, 57, 65, 66

Schilling, C. H., Letscher, D., and Palsson, B. O. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203(3):229 – 248. 4

Schuster, S. (1999). Use and limitations of modular metabolic control analysis in medicine and biotechnology. *Metabolic Engineering*, 1(3):232 – 242. 4

Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, 17(2):53 – 60. 4

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., Gilles, E., et al. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193. 7

Stephanopoulos, G. (1999). Metabolic fluxes and metabolic engineering. *Metabolic Engineering*, 1(1):1–11. 3

Tan, Y., Rivera, J. G. L., Contador, C. A., Asenjo, J. A., and Liao, J. C. (2011). Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. *Metabolic Engineering*, 13(1):60 – 75. 5, 19, 20, 24, 32, 33, 86

Tran, L. M., Rizk, M. L., and Liao, J. C. (2008). Ensemble modeling of metabolic networks. *Biophysical Journal*, 95(12):5606 – 5617. 5, 6, 19, 20, 22, 23, 31

Trinh, C., Wlaschin, A., and Srienc, F. (2009). Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology*, 81:813–826. 10.1007/s00253-008-1770-1. 16, 26

Urbanczik, R. and Wagner, C. (2005). An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, 21(7):1203–1210. 27

Varma, A. and Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology*, 12:994–998. 4

Vita

David Flowers was born to Michael and Diane Flowers on June 9, 1988 in Nashville, Tennessee. He is the second of three sons, Michael and Brian Flowers. David attended Gladeville Elementary School and Rutland Elementary School, then proceeded to West Wilson Middle School and Wilson Central High School, graduating as valedictorian in 2007. He attended the University of Tennessee in Knoxville, earning a bachelor's degree in Chemical Engineering in 2011 and graduating summa cum laude. During this time, he worked as an undergraduate research assistant under Dr. Tsewei Wang as a part of the Laboratory for Information Technologies, performing research in forensic DNA statistics. He accepted a graduate fellowship at the University of Tennessee, earning a master's degree in Chemical Engineering in August of 2012. He will continue his education at the Massachusetts Institute of Technology as a Biological Engineering doctorate student.