UNIVERSITY OF CALIFORNIA SAN DIEGO

Kinetic Model of Tryptophan Synthesis: Construction and Application

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Bioengineering

by

Kevin Matthew Glass

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Nathan Lewis
Professor Christian Metallo

2021

The thesis of Kevin Matthew Glass is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

Dedication

I would like to thank Professor Palsson for chairing my committee as well as Dr. Zielinski for his mentorship throughout this process. I would also like to thank Professor Metallo and Professor Lewis for being members of my committee.

I would also like to thank my father Dr. Brian Glass for his advice and generous loan of an office chair for the thesis writing process.

Table of Contents

Abbreviations

PEP Phosophoenolpyruvate
PRPP Phosphoribosyl Pyrophosphate
GLCptspp D-glucose transport via PEP:Pyr PTS (periplasm)
PTS Phosphotransferase System
GLCpts Glucose phosphotransferase system
PGI Glucose-6-phosphate isomerase
PFK Phosphofructokinase
FBP Fructose-bisphosphatase
FBA Fructose-bisphosphate aldolase
TPI Triose-phosphate isomerase
GAPD Glyceraldehyde-3-phosphate dehydrogenase
PGK Phosphoglycerate kinase
PGM Phosphoglycerate mutase
ENO Enolase
PYK Pyruvate kinase
PPS Phosphoenolpyruvate synthase
PGCD Phosphoglycerate dehydrogenase
PSERT Phosphoserine transaminase
PSP_L Phosphoserine phosphatase (L-serine)
CHORS Chorismate synthase
DDPA 3-deoxy-D-arabino-heptulosonate 7-phosphate syntheses
DHQS 3-dehydroquinate synthase
DHQTi 3-dehydroquinate dehydratase, irreversible
PSCVT 3-phosphoshikimate 1-carboxyvinyltransferase
SHK3Dr Shikimate dehydrogenase
SHKK Shikimate kinase
TALA Transaldolase
G6PDH2r Glucose 6-phosphate dehydrogenase
GND Phosphogluconate dehydrogenase
RPI Ribose-5-phosphate isomerase
TKT1 Transketolase
TKT2 Transketolase
ANS Anthranilate synthase
ANPRT Anthranilate phosphoribosyltransferase
IGPS Indole-3-glycerol-phosphate synthase
PRAIi Phosphoribosylanthranilate isomerase (irreversible)
TRPS1 Tryptophan synthase (indoleglycerol phosphate)
RPE Ribulose 5-phosphate 3-epimerase
PGL 6-phosphogluconolactonase
SK_g6p_c D-Glucose-6-phosphate sink
SK_glc__D_p D-Glucose sink
SK_pep_c Phosphoenolpyruvate sink
SK_pyr_c Pyruvate sink
SK_f6p_c D-Fructose-6-phosphate sink
SK_adp_c Adenosine diphosphate sink
SK_atp_c Adenosine triphosphate sink
SK_h_c H+ sink
SK_h2o_c H2O H2O sink

SK_pi_c Phosphate sink
SK_dhap_c Dihydroxyacetone-phosphate sink
SK_g3p_c Glyceraldehyde 3-phosphate sink
SK_nad_c Nicotinamide-adenine-dinucleotide sink
SK_nadh_c Nicotinamide-adenine-dinucleotide-reduced sink
SK_amp_c Adenosine monophosphate sink
SK_akg_c 2-Oxoglutarate sink
SK_glu__L_c L-Glutamate sink
SK_ser__L_c L-Serine sink
SK_chor_c Chorismate sink
SK_e4p_c D-Erythrose-4-phosphate sink
SK_nadp_c Nicotinamide-adenine-dinucleotide-phosphate sink
SK_nadph_c Nicotinamide-adenine-dinucleotide-phosphate-reduced sink
SK__6pgc_c 6-Phospho-D-gluconate sink
SK_co2_c Carbon Dioxide sink
SK_ru5p__D_c D-Ribulose-5-phosphate sink
SK_r5p_c alpha-D-Ribose-5-phosphate sink
SK_gln__L_c L-Glutamine sink
SK_ppi_c Diphosphate sink
SK_prpp_c PRPP sink
SK_trp__L_c L-Tryptophan sink

List of Figures

List of Tables

List of Schema

Acknowledgements

This thesis, in part is currently being prepared for submission for publication of the material. Glass, Kevin; Zielinski, Daniel; Palsson, Bernhard. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

Kinetic Model of Tryptophan Synthesis: Construction and Application

by

Kevin Matthew Glass

Master of Science in Bioengineering

University of California San Diego, 2021

Professor Bernhard Ø. Palsson, Chair

Tryptophan is a commodity chemical that poses an attractive target for fermentation, but its production is kept tightly controlled by metabolic and transcriptional control mechanisms. Different strain design interventions attempted to increase tryptophan titer, yield and productivity. In this study, a kinetic model of tryptophan synthesis in *E. coli* is constructed. The model consists of a set of mass action reactions corresponding to the mechanism of each enzyme within glycolysis, pentose phosphate, shikimate, and tryptophan synthesis pathways. This model was then perturbed to recreate the effects of strain design interventions in three case studies. First interventions used in a tryptophan producing strain were simulated, resulting in a

1.89 fold difference between final strain productivity and simulated productivity. Then two different sets of interventions from a study on the inhibitory effect of anthranilate on Indole-3-glycerol-phosphate synthase were simulated. Non-similarity was observed between the strain productivity and the model productivity for the first set of interventions, but the second set of interventions from the same study in a fed-batch set-up led to a final productivity of the model within twice the strain productivity. Finally, the effect of a feed rate profile on a tryptophan producing strain was examined, with overexpression found to dominate the effect of the feed rate profile. A 1.06 fold difference between the final simulated productivity value and the final actual productivity measurement was observed. This work demonstrates the potential of mass action kinetic models to replicate strain design experiments

.

Introduction

With advances in omics technologies coupled with an increased demand for environmentally friendly commodity chemical production, the field of metabolic engineering has been modifying the metabolism of cells to produce products of interest. This has been done via adaptive laboratory evolution (McCloskey et al. 2018), targeted mutagenesis (Santos, Xiao, and Stephanopoulos 2012)), and recombinant DNA technology (Ikeda 2006), among other methods. Fermentation of desirable commodity chemicals, such as amino acids, is chosen for economic and environmental reasons (D'Este, Alvarado-Morales, and Angelidaki 2018)

Tryptophan is a target of great interest for metabolic engineering due to its use in supplements, animal feed, and for the production of pharmaceutical compounds of interest, such as 5-hydroxy-methyl tryptophan, ajmalicine, and melatonin. (Ikeda 2006; Jiang and Zhang 2016). Furthermore, tryptophan is difficult to produce in high quantities due to the metabolic and transcriptional regulation of both the shikimate pathway and the tryptophan synthetic pathway. Advances in omics technologies and synthetic biology have enabled and fueled the development of rational approaches to improve tryptophan yield (Ikeda 2006; L. Liu et al. 2019). Consequently, due to its well characterized genetic and metabolic systems, *E. coli* has been frequently used as a host strain (Rodriguez et al. 2014).

In *E. coli*, tryptophan is synthesized via a branching pathway from the shlkimate pathway, which originates in the pentose phosphate pathway of central metabolism, taking one molecule of PEP and one molecule of erythrose-4-phosphate as reactants through 3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase (DDPA). Different levels of transcriptional and metabolic control regulate tryptophan synthesis. Both the shikimate pathway and tryptophan pathway are transcriptionally regulated by repressors such as *trpR*, which inhibits the

expression of both an isozyme of DDPA and other enzymes of the tryptophan synthesis pathway. Furthermore, tryptophan synthesis is also regulated by metabolic control such as through the allosteric inhibition of anthranilate synthase. Compounding the tight regulation, tryptophan synthesis is metabolically expensive; consuming intermediates such as PRPP, L-serine, PEP, and L-glutamine (Ikeda 2006). To mitigate these challenges, strain design approaches have been utilized to increase the supply of precursors, eliminate regulatory controls (transcriptional and metabolic), and increase flux through the tryptophan synthesis pathway.

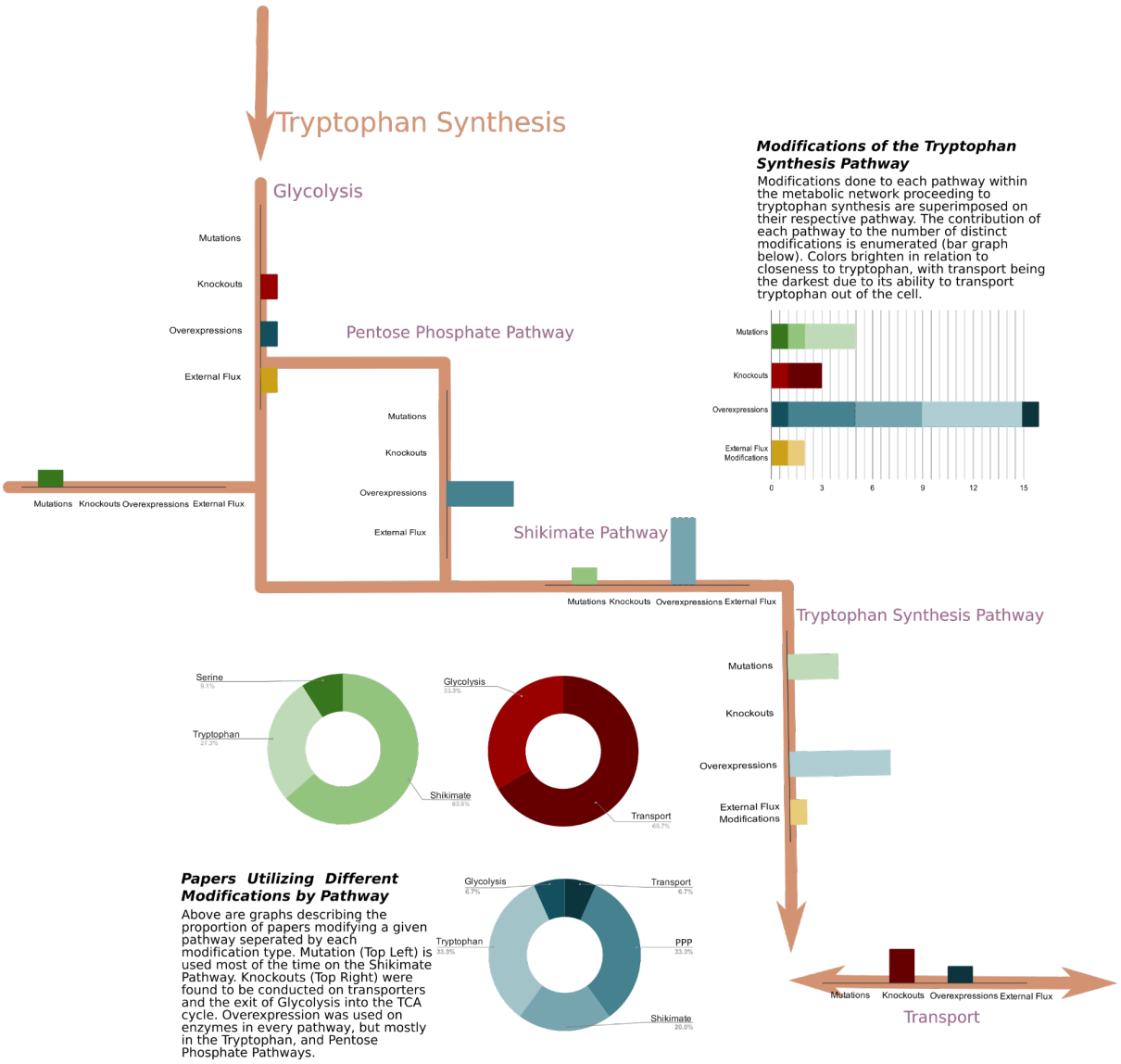

Figure 1.1 A Summary of Historical Modifications to the Tryptophan Synthesis Pathway.
An overview of the different interventions in *E. coli* by pathway. Different colors represent different types of interventions. Red represents knockouts, green represents mutations, blue represents over expression, and yellow represents external flux modifications. Colors brighten in relation to pathway distance from tryptophan synthesis. The bottom left shows the number of papers modifying a given pathway separated by modification type. External flux modifications are omitted due to only two studies using them.

Overall, interventions to increase tryptophan yield and titer are summarized in Figure 1.1

Different interventions have been made to increase tryptophan synthesis at different points in its

synthesis pathway. In the subsequent section, I briefly summarize some of the major experimental interventions made to improve tryptophan production in *E. coli*.

Starting at central metabolism, glucose import through GLCpts has been previously knocked and imported through heterologous importers to increase PEP concentration (Carmona et al. 2015; L. Liu et al. 2019). Other interventions in central metabolism to increase precursor production have also been overexpression of *tktA* (Chen and Zeng 2017) and *ppsA* (Chen and Zeng 2017; L. Liu et al. 2019), as well as knocking out PYK1. Adjacent to central metabolism, mutations of the *pta* and *serA* genes have been done to decrease acetate accumulation (L. Liu, Duan, and Wu 2016), and increase serine production (Chen and Zeng 2017) respectively. To increase flux going into aromatic amino acid synthesis, studies have overexpressed genes encoding enzymes in the shikimate pathway (Lee and Wendisch 2017), such as DDPA, SHK3Dr, and SHKK (Lütke-Eversloh and Stephanopoulos 2008). To relieve allosteric control on the pathway, the first enzyme, DDPA, has been mutated for feedback resistance (Dodge and Gerstner 2002; Jing et al. 2018; Masuo et al. 2016; Jiang and Zhang 2016). To relieve transcriptional regulatory effects on the pathway, repressors such as *trpR* and *tryR* have been knocked out. To increase flux into tryptophan synthesis and overcome regulatory hurdles, the entire operon (*trpABCDE*) has been overexpressed and deregulated in different experiments. Furthermore, enzymes have also been mutated for feedback resistance, with a mutation to *trpE* and *trpC* producing a 20% increase in yield (Chen et al. 2018). Furthermore, *tnaA* is commonly knocked out to prevent the breakdown of tryptophan into indole. Finally, interventions have been made to tryptophan intake and export, such as by knocking out tryptophan permeases encoded by, mtr, *aroP* and *tnaB* (Q. Liu et al. 2012). Knocking these out caused an increase of 51.6% in titer from the parent strain. Furthermore, to increase tryptophan export, the exporter yddG has been overexpressed.

To guide the design of *E. coli*, several different modeling approaches have been used. These include metabolic control analysis (MCA) and gene expression based models where

metabolic control coefficients have been used to couple transcriptional control to enzymes within central metabolism and tryptophan biosynthesis (Schmid et al. 2004). Other approaches have also used enzyme tuning and overexpression data to build models, screening for thermodynamic consistency using an ensemble approach (Rizk and Liao 2009). Furthermore, lumped kinetic models have been deployed to optimize bioreactor setup (Jing et al. 2018; Maria, Mihalachi, and Gijiu 2018). Finally, previous approaches to modeling have used fluxomic, metabolomic, and kinetic data to parameterize a dynamic model(Chassagnole et al. 2002). Each of these approaches have specific limitations. The ensemble modeling approach yields results that both rely upon assumed flux distributions, and are not based upon the underlying mechanisms of the enzymes in the pathway. Schmid et al. 2004 used pooled reactions to encapsulate the entire shikimate and tryptophan pathways (Schmid et al. 2004), and parameterized rate constants based off of fluxes and metabolite concentrations, lacking resolution of individual enzymatic steps. (Maria, Mihalachi, and Gijiu 2018) has also lumped tryptophan synthesis branching off of central metabolism as one pathway. Finally, other kinetic models based off of reactor data and growth rates have been developed, but are even more phenomenological in nature.

To overcome limitations on mechanistic accuracy and model coverage, this study builds a kinetic model utilizing collections of mass action reactions that have been fitted to the known mechanisms of the enzymes in tryptophan synthesis. This study collects enzyme kinetic data from several different databases, curates the information, uses it as information to constrain a non-linear optimization problem to obtain microscopic rate constants within an enzyme mechanism. These rate constants are contained within an enzyme module object and used to represent chemical reactions of enzymes within the model. This approach also accounts for inhibition and activation, and including symmetry model inhibition. To overcome the limitation of metabolite and flux data for building this model, metabolite concentrations and fluxes are optimized based off of experimental flux and metabolite data. The experimental flux data is then

used as inputs to a stoichiometric model of *E. coli*, upon which FBA is run to obtain fluxes for each reaction included within the model. Furthermore, using the fluxes and the metabolite information, in addition to KEQ's, this model produces pseudo-elementary rate constants for form a mass action model of the tryptophan synthesis pathways. Finally, this model is built upon to incorporate enzyme modules to represent each reaction, encapsulating the microscopic rate constants and inhibition steps into one module. Furthermore, this model is applied to replicate the strain design interventions taken by three case studies, the first of which is of a tryptophan over producing strain, the second of which is to simulate the effect of anthranilate inhibition on tryptophan synthesis, and the third of which is to simulate the effect of a modified glucose feed rate in a fed-batch bioreactor. These three case studies were examined to demonstrate this model's utility for strain design.

Methods

2.1. Curation and Network Construction

Curation was initially commenced by reconstructing the reactions present in tryptophan synthesis based upon a literature review of modifications made to central metabolism, shikimate and tryptophan synthetic pathways. Selected reactions were then scrapped via E.C. number to obtain relevant kinetic data from Uni-Prot, Biocyc, and Brenda. This information was then formatted, and validated by manual inspection of the referenced papers. Additional information was obtained from said papers, including but not limited to, buffer information, pH, cofactors and any additional info not initially obtained from the scrape. Furthermore information comprising the enzyme mechanism, inhibition and activation types, and choice of model of inhibition were obtained both from manual inspection of the relevant texts, but also from reasoning about the nature and impact of the data.

Certain enzymes curated required additional considerations not obviously implicated by the source papers, such as some of the finer details of the anthranilate synthase-anthranilate phosphoribosyltransferase enzyme complex, tryptophan synthase A and B, as well as indole-glycerol-3-phosphate synthase(IGPS), and phosphoribosylanthranilate isomerase (PRAIi). Each of these enzymes required separate assumptions about how to model them.

Anthranilate synthase was modeled as a separate enzyme than anthranilate phosphoribosyltransferase, due to technical considerations with the MASSef solver. Furthermore, due to technical considerations, each reaction in the complex was modeled as having competitive inhibition with respect to tryptophan. Similarly, despite IGPS and PRAIi being produced by the same gene, we also modeled them as separate complexes as they seem to have different regulators of activity. Finally, we assumed channeling between *trpA* and *trpB* to produce tryptophan, and thus used a single reaction of TRPS1 to encapsulate the entire reaction.

Furthermore, when kinetic information for a fit was missing, the values were gapfilled using either a gapfilled value of K_m = 0.0001, based on the assumption that Km ~ x (Bennett et al. 2009) or used either (Heckmann et al. 2018) or (Davidi et al. 2016) to gapfill for k_{cat}s

Finally, model input data was obtained from Heinemann, Rabinowitz, and Sauer data sets with additional information obtained from the van Gulik dataset (Taymaz-Nikerel et al. 2009; Schmidt et al. 2016; Bennett et al. 2009; Gerosa et al. 2015). Metabolomics and Equilibrium data was processed through a thermodynamic consistency adjustment to ensure that the metabolite concentrations were thermodynamically consistent with one another. See later section on fitting and preparation for fitting.

2.2. MASSef Construction Workflow

MASSef is a mathematica-python toolkit used to generate mass action rate constants for a given enzyme mechanism based on provided kinetic information. Generally this workflow can

be broken down into several steps: data Import, mechanism specification, generation of equations, fitting to the equations, and finally, evaluating the quality of the fit

Data is first imported into a MASSef workflow from a specified data file (.csv format) and is in turn read and imported into mathematica. Data read includes, Michaelis-Menten constants(K_m), dissociation constants K_d, turnover rate (k_{cat}), kinetic constants (K_i) and Equilibrium constants (K_{eq}). These were then weighed for both inclusion in the fit, and for relative impact on the resulting fit.

After importing relevant data, the mechanism for the enzymes were specified as a series of mass action equations corresponding to microscopic reaction steps (https://github.com/opencobra/MASSef), with microscopic rate constants. Furthermore if small molecule inhibition or activation is present, it is modeled by specifying the inhibitor in question, and adding additional equations to represent the binding reactions. Inhibition and activation equations were added automatically based upon the given type of inhibition or activation specified. The microscopic rate constant regression workflow of (https://github.com/opencobra/MASSef) was followed.

Comparison equations for each data type were constructed by solving equations corresponding to the mass balances of each ligand in addition to an expression of the total enzyme concentration. Furthermore, comparison equations were obtained by fitting to equations corresponding to the to a few approximating functions of v_{sat} and v_{rel}. These were used to construct a set of expressions from the K_m, K_i, k_{cat}, and K_{eq} values. The reaction equilibrium data is set by fitting to a Haldane ratio of microscopic rate constants within the enzyme mechanism. Microscopic rate constants were obtained via two step non-linear optimization. The first step entails using a particle swarm optimization (PSO) process to minimize the sum of the squared residuals between the comparison equations and the data values. PSO found rate constants that were successful across different enzyme mechanisms, and were used as the initial values for the derivative based Levenberg-Marquadt, which restrict the rate constants to a low sum of

square deviations to the data. To obtain more realistic fits, several gapfilling approaches were used. These include assigning an “average” K_m in order to eliminate back calculation problems, as well as using gapfilled turnover rates from calculated in-vivo values (Heckmann et al. 2018; Davidi et al. 2016)

2.3. MASSpy Workflow

2.3.1. Calculation of Reaction Gibbs Energies

To estimate reaction Gibbs energies, metabolite data for eight growth conditions for *E. coli* was obtained from the same study as the flux estimates (Gerosa et al. 2015). Reaction equilibrium constants (K_{eq} s) were estimated using the latest group contribution method (Du et al. 2018). Then, a thermodynamic FBA problem (Henry, Broadbelt, and Hatzimanikatis 2007) was solved constraining only high flux reactions (> 0.1 mmol/gDW/hr), subject to uncertainty. Once a feasible set of fluxes, metabolite concentrations (x), and K_{eq} s were identified, convex sampling was used to obtain a distribution of x and K_{eq} values that accounts for measurement gaps and uncertainty. These sampled x and K_{eq} values were used to calculate the reaction Gibbs energies using the definition:

$\Delta G = -RT \log(K_{eq}) + \log(Q)$

$Q = \Pi x S$

where Q is the reaction quotient defined as the product of the metabolite concentrations (or activities) to the power of their stoichiometric coefficient in the reaction (S).

For transport reactions, the chemical potential due to charge transfer is calculated by taking into account the membrane potential, and this additional energy is added to the Gibbs energy of reaction to calculate the final transport K_{eq}. (Jol et al. 2010).

2.3.2. Thermodynamic Consistency Optimization

To calculate a thermodynamically feasible set of metabolite concentrations, fluxes, and equilibrium constants, we used an iterative optimization approach. First, growth was constrained to 90% of the measured growth rate. K_{eq} parameters and metabolite concentrations were set to open bounds, defined by a wide range around the physiological and observed ranges. Then, an initial set of reactions, consisting of core metabolism, were constrained by thermodynamic constraints. An objective was set to be the target metabolite and K_{eq} values, and the distance of the parameters to the target values was minimized in an MIQP optimization.

If a feasible solution was found, the next iteration was set up by identifying any reactions that carried an absolute flux greater than a target value (0.000001 mmol/gDW/hr). These ‘high flux’ reactions were then additionally constrained by thermodynamic constraints for the next round of simulation. This iterative procedure is designed to prevent reactions from ‘escaping’ thermodynamic constraints. This procedure is repeated until the list of constrained reactions no longer changes. The final adjusted metabolite concentrations and equilibrium constants are used as a thermodynamically feasible set of values.

For this experiment, the workflow described in the case study 3 of (Haiman et al. 2021) was followed, with specific modifications made for this study. The workflow breaks down into several different steps: Importing data and defining constants, Setting up a constraint based model and obtaining Flux states, Setting up the mass action model, and Setting up and exporting the Enzyme_module based mass action model from MASSef.

2.3.3 Obtaining Flux states

Initially a constraint based model of *E. coli* (iML1515) was used. This was used in tandem with fluxomics from (Gerosa et al. 2015) corresponding to the medium this model is made for (in our protocol, glucose). In order to obtain relevant steady state fluxes, FBA was used with quadratic programming to obtain fluxes that minimized error between the

experimental data, and the model results. Following the case study mentioned in (Haiman et al. 2021) we set pairs of irreversible fluxes that oppose each other, and added 10% to each to ensure they show up in the flux solution. Furthermore, to ensure that TRPS1 showed up in the flux solution, the non-tunneled forms of TRPS, TRPS2 and TRPS3, were knocked out in tandem with TRPAS2. Furthermore to ensure that the fluxes had enough room to adjust, the bounds of the media carbon source import reaction was set to -1000 and 1000.

Resulting fluxes were then plotted and inspected to ensure no issues occurred with the minimization step.

2.3.4 Setting up the Mass Action Model

To set up the mass action model, reactions specified within the scope of our network were specified. For enzymes with multiple isozymes, the percentage flux through each form was determined by the isozyme_flux split attribute, which for our study was set at 0.5. Consequently, this indicates that 50% of the flux going through a given reaction goes through each isozyme form. In the case of DDPA, only the first two isozymes were included, corresponding to *aroG* and *aroF*. The reason for this being that *aroH* has a lack of available isozyme specific kinetic data, and only accounts for 1% of the total concentration. Furthermore, kinetic data for phenylalanine and tyrosine feedback inhibition was not obtained through our scrape of databases, and thus was not included in our model. Since most of the case studies we examined already had *aroG* be feedback deregulated, this was assumed to be a non-issue.

The next step required making mass action reactions from each required steady state reaction from our minimized FBA model. These were then added to the model, and their steady state fluxes were specified based off of the flux solutions. Furthermore, the flux solutions were converted to a M/s format from the default scaling of the FBA model. Modifications were also made to the underlying workflow code to not convert flux amounts twice.

Next, Equilibrium constants were assigned for each reaction corresponding to equilibrium constants obtained from a thermodynamic adjustment of the model(cite dan's thermodynamic adjustment paper). Following this metabolite names are adjusted to match the correct format of the model. Furthermore, in order to include every metabolite involved in the reactions in the model, the enzyme modules are checked for each metabolite bound to them, and that is checked about the metabolites within the model. The metabolites found in the enzyme module species file, but not within the model are added to the model.

Following this, the metabolite values from the thermodynamically adjusted fit are imported. These are used as initial guesses for their respective metabolite initial concentrations for the model. For the ligands not found in the thermodynamically adjusted fit data, a value of 0.001 is assumed as an average metabolite concentration (citation, dan?). In the case of indole, we assume a concentration of 0.00001 based on similar metabolite concentrations. Other cases use an assumed value of 0.001. Furthermore, if the medium is glucose or pyruvate, their respective concentrations are not defined (external glucose and pyruvate respectively). These concentrations are used as the input variables to the concentration solver package within MASSpy, with a thermodynamic buffer of 0.5, and an assumed lower bound of ten to the negative seventh. This lower bound was chosen, as it represents a value below which it would be unrealistic to expect metabolites to stay or be at for later fitting steps. Furthermore, when the QP problem is set up, the equilibrium constants are fixed in the model, as we assume that they are correct. While it is possible to unfix the equilibrium constants, and try to fit both equilibrium constants and metabolite concentrations, metabolite concentrations were chosen to be adjusted mostly because these values are already thermodynamically consistent, and thus the adjustment isn't that great. Resulting changes to the metabolite values are then examined to see evidence of major shifts before proceeding to the next step.

Following getting a concentration solution for the metabolites in the model, a concentration sampler is then used to sample concentrations for the model using different

samples. The results of this sampling step are then compared with ten different seeds and inspected to ensure that the results are within a reasonable range. Finally, the concentrations are set to one of the seed values.

The next step entailed constructing the initial mass action model using the fluxes, metabolite concentrations, and reactions of the model. The model is constructed by making a copy of the model, and then performing the dot operation of the stoichiometric matrix with the fluxes of the reactions within the model. This determines imbalances corresponding to the metabolites not fully generated and consumed within the model over a threshold of ten to the negative tenth. These imbalances are used for the construction of the appropriate sink reactions. These sink reactions correspond to the other reactions within the cell consuming this metabolite to maintain the cell at steady state. The boundary conditions of these sinks correspond to the concentration of the metabolite, their equilibrium constant is one, their forward rate constant is equal to the imbalance divided by the initial concentration of the metabolite, and their steady state flux is equal to the imbalance. In cases where the imbalance is negative, the stoichiometry of the reaction is flipped. Finally, pseudo-elementary rate constants (PERCs) are calculated for each reaction in the model using the K_{eq}'s metabolite concentrations and steady state fluxes of the model. This step produces a number of different models corresponding to the specified number of models in the ensemble.

Following the construction of the mass action model model, the models are inspected to ensure correct functionality. First each model is simulated to steady state, and then boundary conditions are checked to ensure reasonable values, and a simulation is run to ensure that the model maintains steady state and does not have any kind of longer term dynamic instability. Upon seeing that no issues (due to miss inputs or aberrant model behavior) are present, the model can then proceed to the enzyme module incorporation step.

2.3.5. Enzyme Module Incorporation and Model Export

Before adding enzyme modules to the model, a set of reactions to exclude is defined. Within this step, first gene protein reaction associations are obtained from the model construction file, followed by steps skipping each reaction that is either a sink, a mass action reaction of interest, or an enzyme module with an already extant externally made enzyme module that must be imported.

Enzyme modules were then constructed within MASSpy using the make_enzyme_module_from_dir() function from (Haiman et al. 2021). This function takes the steady state flux and metabolite information from the model it is being added to, as well as the path information for the MASSef enzyme module export. This information is then used to calculate the total enzyme concentration based on the microreaciton rate constants, steady state flux through the enzyme, and the steady state concentrations. Finally, after calculating the total concentration of the enzyme module, each enzyme form's concentration can be calculated using the already available information. The result of this is then merged with the model, with its corresponding mass action reaction being removed. In cases where an enzyme has multiple isozymes, the steady state flux of the enzyme is multiplied by the isozyme_flux_split.

After all of the enzyme modules were added, GLCpts was incorporated as an external enzyme module and merged with each model. The new models are then checked for simulatability prior to export.

Models were then simulated to steady state for a series of time points. Models were exported if a steady state was found. If a steady state was not found, the model was deleted.

After exporting all models, analytics were then performed to get information about the quality of the models. Average total concentrations of enzymes from each enzyme module across all models were compared to known literature values and automatically uploaded to an Google Sheet. Furthermore, mass action over equilibrium information was also obtained for each reaction and exported to a Google Sheet.

2.3.6. Model Debugging and Simulation of Experiments

After export, models were then inspected to obtain IC/K_m, IC/K_i, and IC/K_a information for an average model for each reaction where relevant. This information was used in the debugging of the models and was exported to an external database.

After debugging the models, experiments were simulated by alternate model construction, and perturbation. Alternate model construction entailed constructing an ensemble of models for which known alterations to the model were used in the model construction process, while perturbation encapsulated changes to inhibition, enzyme form concentrations, and external feed rate changes.

2.3.7. Obtaining Productivities from Graphs of Titers:

To obtain the rate of change of metabolites in a bioreactor over time, Webplotdigitizer was used to get points corresponding to concentration measurements over time. These concentration over time measurements were converted from g/L over hours to Molarity over seconds using the molar mass of the respective metabolites. These datasets were then put into the gradient function of Numpy, with the dx being the average time between concentration measurements. Furthermore, for the feed rate profile case study, data points corresponding to grams added per minute were taken using Webplotdigitizer, and were converted to molarity added per second. These were then integrated using a symbolic solver (Wolfram Alpha) to obtain the boundary metabolite equations.

2.4. Simulation of Case Study 1

In this workflow, based on the interventions done in (Chen and Zeng 2017), the changes to the enzyme expression, regulation, and tryptophan exporters were modeled where relevant within the model. To model the relevant over expression, an increase to enzyme concentration

of 3x was used, in line with (de Boer, Comstock, and Vasser 1983), which indicates that the promoter used was 3x more efficient than the default tryptophan promoter. To simulate this fold increase, the forms of each affected enzyme were increased in concentration by three. Changes to model inhibition were modeled by perturbing each model, and setting the forward rate constant of each inhibition reaction to zero. To model the knockout of permeases, with only the exporter yddG being present, the equilibrium constant of the tryptophan sink reaction was altered to better reflect the unidirectional exchange with the environment. In the case of Workflow 4, it used an ensemble that had a higher assumed anthranilate concentration in the course of production, with a concentration put in initially of 0.001 to model the effects of anthranilate inhibition on IGPS as mentioned by (Chen et al. 2018). To model the overexpression of IGPS, the isozyme representing the assumption that IGPS is inhibited by anthranilate, had its enzyme forms multiplied by three. For this model the following enzymes were overexpressed: *aroG*$_{(fbr)}$, *trpE:A*, and *serA*$_{(fbr)}$. Furthermore, knockouts were made to the permiases *mtr*, *aroP* and *tnaB*.

To compare the results of this model with the perturbations representing the interventions, the productivity of tryptophan in the strain used in the study was examined. To obtain a productivity value from the model, the external flux of tryptophan to the environment was modeled as the difference between the flux of the tryptophan sink and the steady state flux of the tryptophan sink reaction (representing all of the tryptophan sinks within the cell not accounted for in the model). This represents what must be exported from the cell as its “productivity”. This new “tryptophan export” flux was then obtained at a time corresponding to the end of the experiment. Furthermore, the response of the model was also plotted against the productivity value from (Chen and Zeng 2017) of S28. The productivity of the model was compared as a ratio between the model’s “tryptophan_export” and the productivity shown in the study. To analyze the distribution of this ratio over different model fits, 63 different models were then perturbed in the manner of this experiment and plotted in a histogram, with bin width, and

therefore bin number being determined from the Freedman-Diaconis rule (Freedman and Diaconis 1981).

2.5. Simulation of Case Study 2

Case study 2 was simulated using the base model as case study 1, however with the addition of data points taken from the graph of the figure shown in (Chen et al. 2018). To model anthranilate inhibition, our model was updated to have an isozyme of IGPS with the inhibition information of anthranilate included. Since this is present in *E. coli*, all models were built with the sensitive form and insensitive form accounting for 50% of the flux through the enzyme module by default.

Furthermore, data from the graph was converted from g/L and was derived to obtain the flux over time, which was used to compare against the perturbed model. Since the strain was taken from the production strain obtained in workflow 4, the baseline modifications are the same. Furthermore, to model the effect of IGPS feedforward resistance, the k_f values for each inhibition reaction of anthranilate were set to zero, and the non-inhibited form of IGPS was perturbed to be zero. To compare the results of the tryptophan export flux to the results of the experiment, the difference between the tryptophan production over time and the tryptophan export flux was taken and averaged across all modeled used for this experiment. In this case, two models were used, since case study 4 established that the range of results for overexpressions lied within a given range.

For this case study, two different tests were replicated. In the first case, the values for tryptophan concentration change over time from a 1.5L batch bioreactor were plotted against the perturbed models corresponding to each strain compared. In this graph, the performance of a strain of the *E. coli* with anthranilate inhibition eliminated was compared to the baseline strain, S28. Each of these models was simulated for the length of time of the experiment, and was

compared to each other, as well as compared to the productivity from the strains from the experiment.

The second test was that of a fed-batch reactor comparing a strain of S28 with a TAC promoter behind the wild type IGPS, with S28 with just the plasmid with the tac promoter. To simulate these perturbations, the S28 interventions were modeled, with IGPSr being overexpressed by 11 times as opposed to the usual 11/3 times. This was compared with the baseline interventions for S28, and the productivities of the two strains via data obtained from the graph.

2.6. Simulation of Case Study 3

The simulation of case study 3 was undertaken in a similar manner to the first. This case study, based on (Dodge and Gerstner 2002), the packages for the set up of case study 1 were imported, followed by importing additional packages of Piecewise() from Sympy for the perturbation. Data points from the graph of the feed rate profile used for the paper were obtained using WebPlotDigitizer (Rohatgi 2017) and converted to Molar/s in the figure setup workflow. Furthermore, the functions used in the set up of case study 1 were also defined for this case study. Differing from workflow 1, this case study used a previous version of the model that did not assume that anthranilate was accumulated. This model was imported as the basis for the analysis.

Furthermore, after setting up copies of this model for simulation, the feed rate profiles were obtained from the paper. Two approaches were used to obtain the relevant equations. The first approach used the equations as they were set up in the paper itself, and solved for the integrals of these feed rates in order to obtain boundary condition concentration changes. The second approach was to plot the concentration over time for each segment of the piecewise function, and use a graphing program (Apple Numbers) to determine the line of best fit for each

section. These equations were then integrated to obtain the change of the boundary metabolite over time.

Since the boundary conditions are assumed to be a fixed variable, the equations used the boundary initial condition of glucose as an initial condition for the first equation of the piecewise function. Initial conditions for the other equations were obtained by continuity between the legs concentrations. Finally, after solving for the equations of the boundary condition, glc__D_b, these equations were implemented as a perturbation by using the Piecewise() Sympy function.

Furthermore, in this experiment, the feed rate was given as grams of feed per minute. To convert this to a flux that could be integrated, a conversion of minutes to seconds was done. Since the paper used a feed that was 60%m/v of glucose, we determined the molarity of the feed by determining the volume added, followed by converting the grams of solute added to mols.

In addition to the feed rate profile, the overexpression of genes in this study was modeled by multiplying all of the individual forms of each enzyme module by the overexpression amount.

Following setup, models were perturbed with the two sets of piecewise functions and were compared to determine whether or not a substantive difference existed between the two sets of equations. Due to its behavior most similar following the feed rate profile of the paper, the equations derived from the graphs were used.

Furthermore, the addition of the piecewise function perturbation was compared against not perturbing the model.

Finally, the values of tryptophan export were graphed against the tryptophan export flux from the model and these results were compared. In addition, the concentration of glutamate was plotted over time as well in order to compare the effects of the feed rate and intervention on glutamate accumulation.

Results

3.1. Curation of Kinetic Data

Curation of data was started initially by determining both the reactions of the tryptophan synthesis pathway, and also reactions from the pentose phosphate pathway and glycolysis before it. Furthermore, the serine synthesis pathway was included as overexpression of PGCD was a prominent intervention, and serine is a precursor for tryptophan synthesis. Following the webscrape of databases of kinetic information, curation of kinetic data obtained K_m, k_{cat}, K_{eq}, K_i, and other data. After data was gathered, the depth of curation and available data was studied, and used to generate a curation level score. A score of 1 indicates only gapfilled information being available for a given enzyme, such as PGL, which has gapfilled values for its k_{cat} and K_m values. One the other hand, a score of 3 represents a fully curated and detailed set of kinetic data for an enzyme; including additional metadata such as buffer information, co-substrates, and Km coverage for both reaction directions. Curation level 3 enzymes include PSCVT (the enzyme inhibited by glyphosate, with over 50 different kinetic data values) and GAPD (with detailed information available, but not in as great a quantity). A score of 2 represents an enzyme with a suitable amount of curation to construct an enzyme module on its own, but lacking some metadata (buffer info) or Km coverage. The results of a review of the automated scape are shown in Figure 2.1 on the right. Out of 44 curated enzymes, the average curation level was 2.2. While curation level 3 was obtained for some enzymes, most enzymes were curation level 2. Some enzymes were curation level 1, being mostly comprised of gapfilled information. From this one can see that on average enzyme modules of the model have sufficient curation quality, which indicates that it was likely not an obstacle to model quality. However, the presence of curation level 1 modules does mean that the mechanism of the enzymes can have rate constants with a broader range due to a lack of information available to constrain the calculated K_{eq} values and micro rate constants.

Figure 2.1 on the left shows the number of enzymes with given curation needs. Looking at specific aspects of curation, not many enzymes curated had K_m information for both reaction directions, at least, as curated. This can be important, as having two sets of K_m values allows for more constrained K_{eq}'s and k_f's to be calculated by the model, and limits the range of possible values. Furthermore, while the number of enzyme modules requiring K_m gapfilling is not a large proportion of all of the enzyme modules, K_m's serve to constrain the range of possible Keq's for each reaction step. Lacking a K_m could cause a wider range of values, and therefore a less accurate module. Finally, about half of the enzyme modules required k_{cat} gapfilling. Since the values used for gapfilling are derived from either (Heckmann et al. 2018) or from (Davidi et al. 2016), within both an assumption of steady state flux of the enzyme is already included, which could cause the enzyme modules to have more realistic concentrations that they might otherwise have.

In addition to the kinetic data points themselves, additional metadata was collected, such as buffer, ionic strength, strain, and co-substrate concentrations, when it was available and processable by the model. Other information, such as types of inhibition, and Hill coefficients for mechanisms not able to be fit in MASSef were collected but not utilized in the module construction workflow.

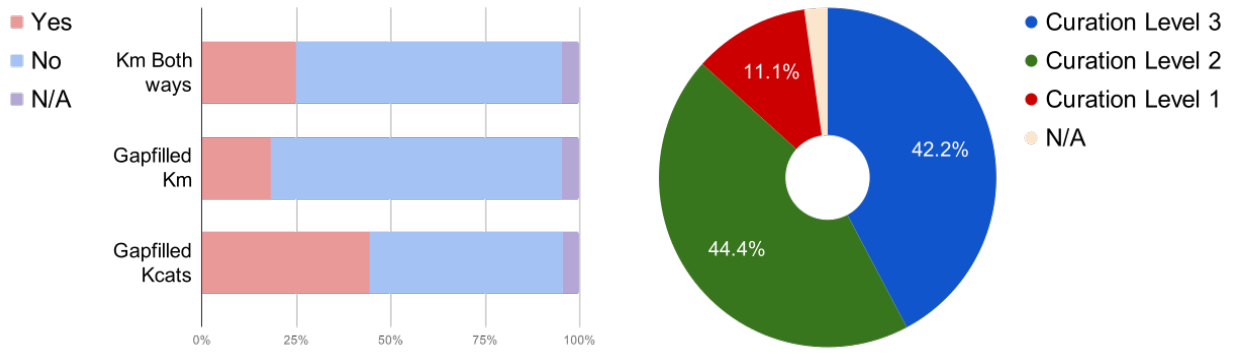

Figure 2.1 Enzyme Module Curation Details
The Bar chart on the left depicts the number of enzymes that have K_m's for each ligand going in both reaction directions (Top), the number of enzyme modules that required K_m gapfilling (Middle), and the number of enzyme modules that required k_{cat} gapfilling (Bottom). Red indicates a Yes, Blue indicates No, and purple represents GLCpts, which is constructed outside of MASSef. The Pie chart on the right depicts the percentage of enzymes in the final curation with various curation levels. Blue represents a curation level of three, green represents a level of two, and red represents a level of one. Beige corresponds to GLCpts, which is built using a different protocol

3.2. Enzyme Kinetic Module Construction with the MASSef Package

Looking at the overall module status as shown in Figure 3.1, the quality of fit is shown for all of the enzyme modules produced from the original set of modules (left) and the final set of modules (right). Among these, most, 66.7% had a non-perfect fit, ranging from marginal with some calculated values not matching up, to completely different, with most values not matching up. The 29.2% of enzyme modules with a good flux fit were mostly concentrated in the pentose phosphate pathway. Previous iterations of fits had a higher percentage of good fits with 50% of the previous fits being good, 47.1% being not good, and 2.9% corresponding to GLCpts which is fit using a different method.. The lower good fit percentage in the flux fit likely owes to the addition of the flux fit step, which conflicts with other data points, leading to a non-perfect fit.

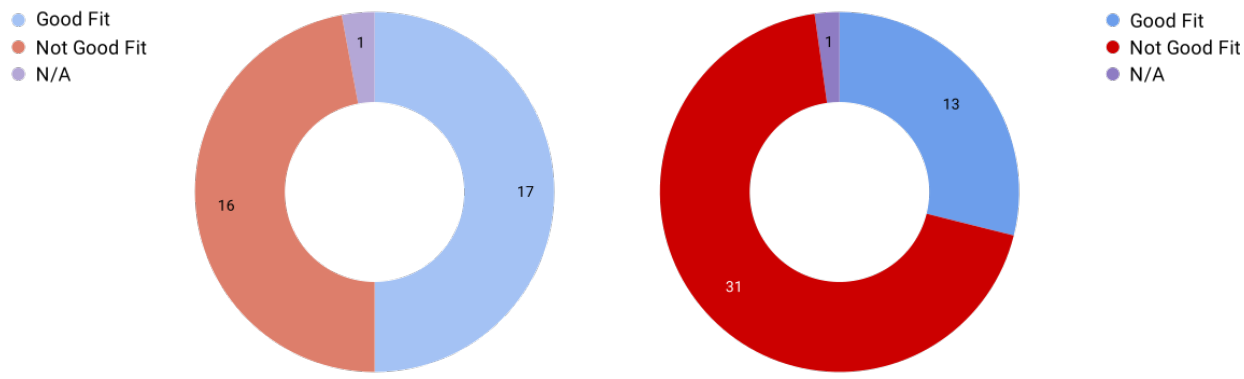

Fig 3.1 Total Module Status.
A comparison of the number of good fits between the original set of enzyme modules (left), and the final set (right). Good fits are represented in blue, while non-good fits are in red. In both sets of data, one enzyme module, GLCpts, was not made using the MASSef workflow.

Moving through the workflow, a mechanism was assembled based on the information obtained from scrapped literature. An example of a mechanism is shown in Scheme 1.

$$\left\{\mathrm{GAPD^c} + \mathrm{nad^c} \overset{\text{GAPD1}}{\rightleftharpoons} \mathrm{GAPD^c\,\&\,nad^c},\ \mathrm{GAPD^c\,\&\,nadh^c} \overset{\text{GAPD2}}{\rightleftharpoons} \mathrm{GAPD^c} + \mathrm{nadh^c},\ \mathrm{GAPD^c\,\&\,nad^c} + \mathrm{g3p^c} \overset{\text{GAPD3}}{\rightleftharpoons} \mathrm{GAPD^c\,\&\,nad^c\,\&\,g3p^c},\right.$$
$$\left.\mathrm{GAPD^c\,\&\,nadh^c\,\&\,13dpg^c} \overset{\text{GAPD4}}{\rightleftharpoons} \mathrm{GAPD^c\,\&\,nadh^c} + \mathrm{13dpg^c},\ \mathrm{GAPD^c\,\&\,nad^c\,\&\,g3p^c} + \mathrm{pi^c} \overset{\text{GAPD5}}{\rightleftharpoons} \mathrm{GAPD^c\,\&\,nad^c\,\&\,g3p^c\,\&\,pi^c},\ \mathrm{GAPD^c\,\&\,nad^c\,\&\,g3p^c\,\&\,pi^c} \overset{\text{GAPD6}}{\rightleftharpoons} \mathrm{GAPD^c\,\&\,nadh^c\,\&\,13dpg^c}\right\}$$

Scheme 1.1: An Example Reaction Mechanism.
The equations above describe the individual steps of the mechanism for GAPD, from initial binding steps, to final release.

For the preliminary version of the model, the figures of MASSef had a mixed quality of fit, in part owing to the different factors involved in the process. Some enzyme modules before the addition of additional gap filled data (K_m's, k_{cat}s from (Davidi et al. 2016), had poor fits, and the overall distribution of fits can be seen in Figure 3.1. With the addition of gapfilled data, there was an additional increase in the number of enzymes with a close fit. Upon gapfilling and using the flux fit technique, the number of good fits decreased, however, this is likely due to the inclusion of the flux fit information, which conflicts with the other information used to setup the model,

however it also constrains the values of rate constants to be more in line with the steady state flux of the model. Furthermore, in order to implement the flux fit, the concentrations of products had to be substituted to zero, which could cause conflicts with the kinetic data used for the fit.

Furthermore, for the case of PFK, we needed to use an alternative approach to fitting it. Due to technical limitations simulating substrate allosteric inhibition, the mechanism was fit first using gdp as the inhibitor, which had a similar K_i to adp. After the construction of the rate law, the mechanism was modified to replace all instances of gdp with "@modadp" which acted as a separately bound form of adp.

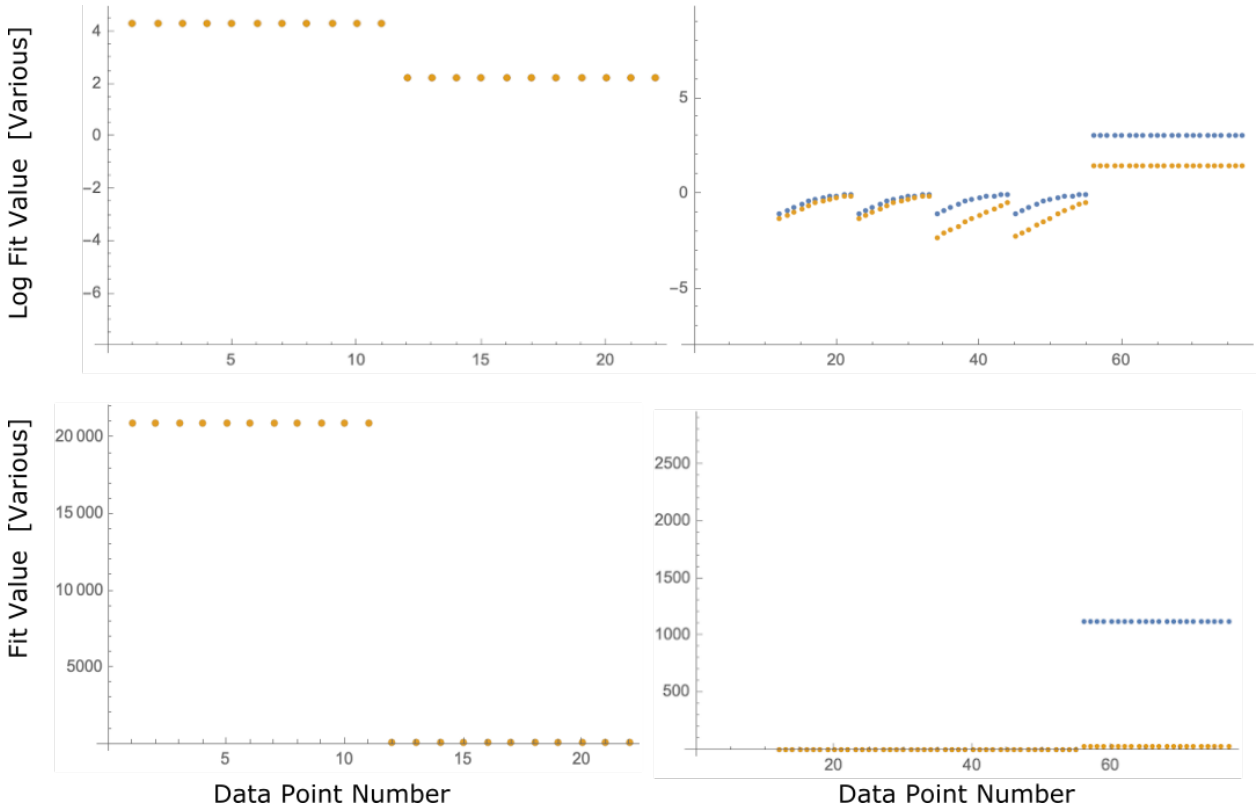

Figure 3.2: Simulated Data and best fit for two enzymes PGL (left) and SHK3Dr_iso_2 (Right). Orange points are the simulated data points, while the best fits are the blue points. The top graphs are logarithmic, while the bottom graphs are in terms of the constants themselves.

After equations are set up, a rate constant optimization can be run ,with the enzyme modules producing graphs of simulated data versus best fits. These graphs compare the products of the optimization step with the results of the simulated data step as seen in Figure 3.2. Figure 3.2 shows two different sets of graphs from PGL and SHK3Dr respectively. PGL on the left shows a “good” fit, where no blue dots are visible, while SHK3Dr_iso_2 shows a “non-good” or poor fit. In this case, the best fit results of SHK3DR_iso_2 are above the simulated results by multiple orders of magnitude, both the logarithmic (top) and the linear (bottom) graphs showing such. Figure 3.3 shows that quality of fit is not necessarily indicative of the quality of the module. SHK3Dr_iso_2, shown in the left, has tighter ranges of micro rate constants, and micro equilibrium constants, compared to the broader ranges of the micro rate constants and K_{eq}’s for PGL. In the case of PGL, its “good” fit might be attributable to a lack of available data, causing overfitting, however, this is compensated for by running a large number of iterations of the optimization procedure, in the process sampling alternative optimal solutions. SHK3Dr_iso_2 has a larger number of available data points, and yet has a more constrained ranges of constants, at the cost of increased fitting error.

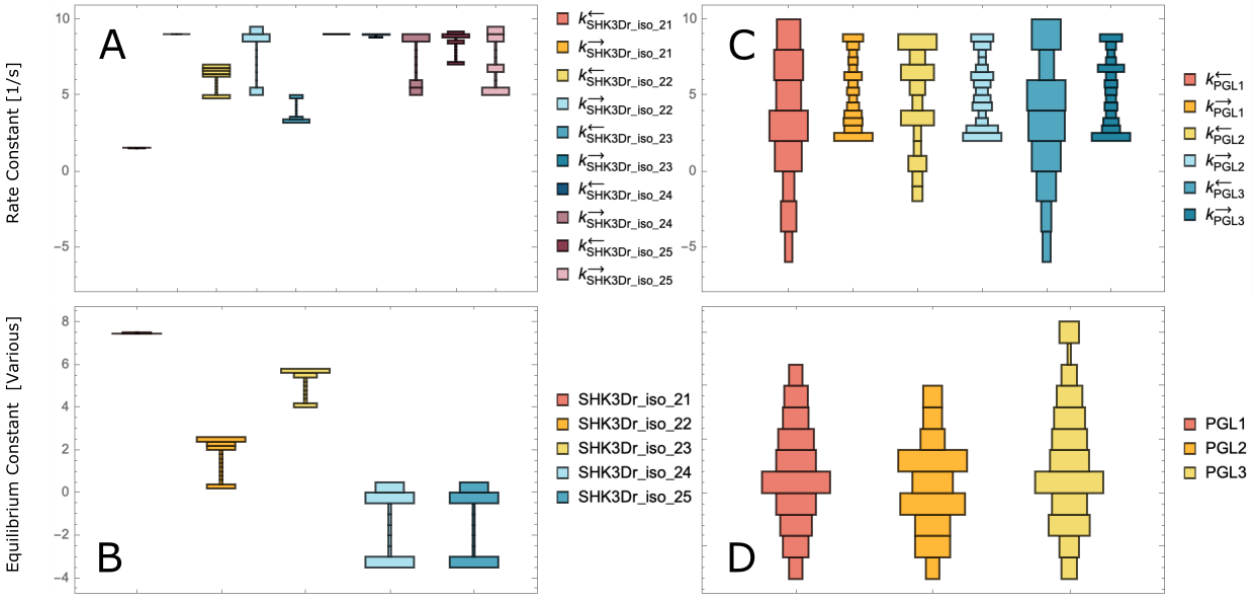

Figure 3.3 Comparison of Ranges for Rate Constants and Equilibrium Constants across PGL and SHK3Dr_iso_2.
At first the rate constants are compared between SHK3Dr_iso_2 (A) and PGL (C), which show the ranges for the rate constants obtained from the fit. SHK3Dr_iso_2 and PGL are then compared based on their K_{eq}'s (B) and (D) respectively.

Figure 3.4 shows the distribution plot of the data error taken from GAPD. The graph shows the error from substituting the range of K_{eq} and k_f values into the Haldane ratio, absolute forward and other ratios. As one can see, the error of recalculating these ratios is varies across orders of magnitude. In this instance, certain parameters, such as the relative rate for g3p had a wider range of error whereas, the error for phosphate was much higher but has a much more narrow range. This plot is reflective of GAPD's worse fit for the flux fit. This is confirmed in Figure 3.5, a box plot of recalculated enzyme kinetics parameters from the microscopic rate constants generated by the workflow. As one an see, there is a wider range for the K_m of pi and g3p, which is consistent with the wider range seen in 3.4. Overall, adding in the flux fit causes a more error between the back calculated kinetic values and the actual values. This overall reflects a degree of divergence between the curated kinetic information and the elementary rate constants due to to the flux fit. This example is within the pattern of the flux fit causing a greater portion of the

enzyme modules having a poor fit. In this case, due to the flux fit, and available data used, there is more inconsistency between values, reflected as higher error. After looking at the back calculation error graphs, enzyme modules were saved and exported into a file format compatible with import into MASSpy.

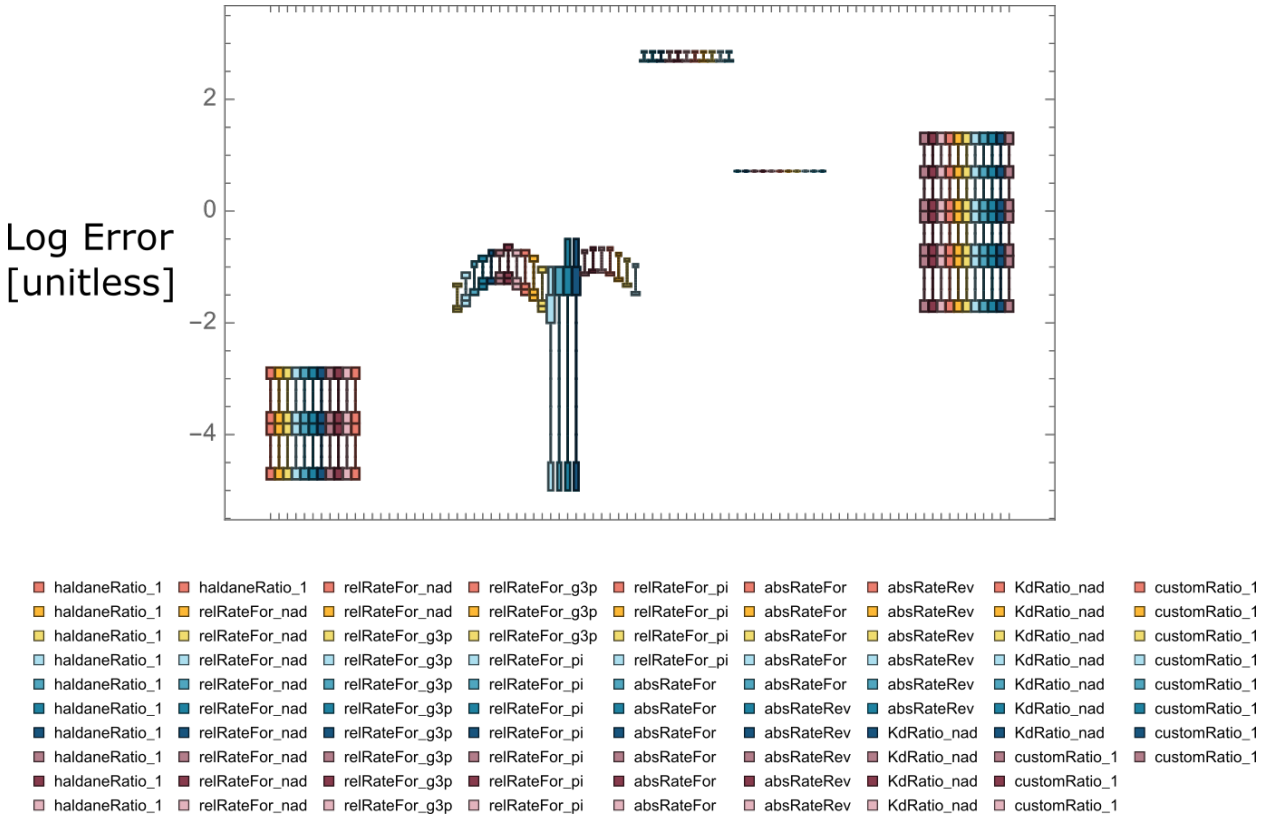

Figure 3.4: An Example of a Distribution Plot of the Data Error Taken from GAPD.
This plot compares error of the substituted in K_{eq} and k_f terms to obtain the distribution of error for the Haldane ratio, absolute forward and reverse rates, as well as relative forward and reverse rates for specific ligands.

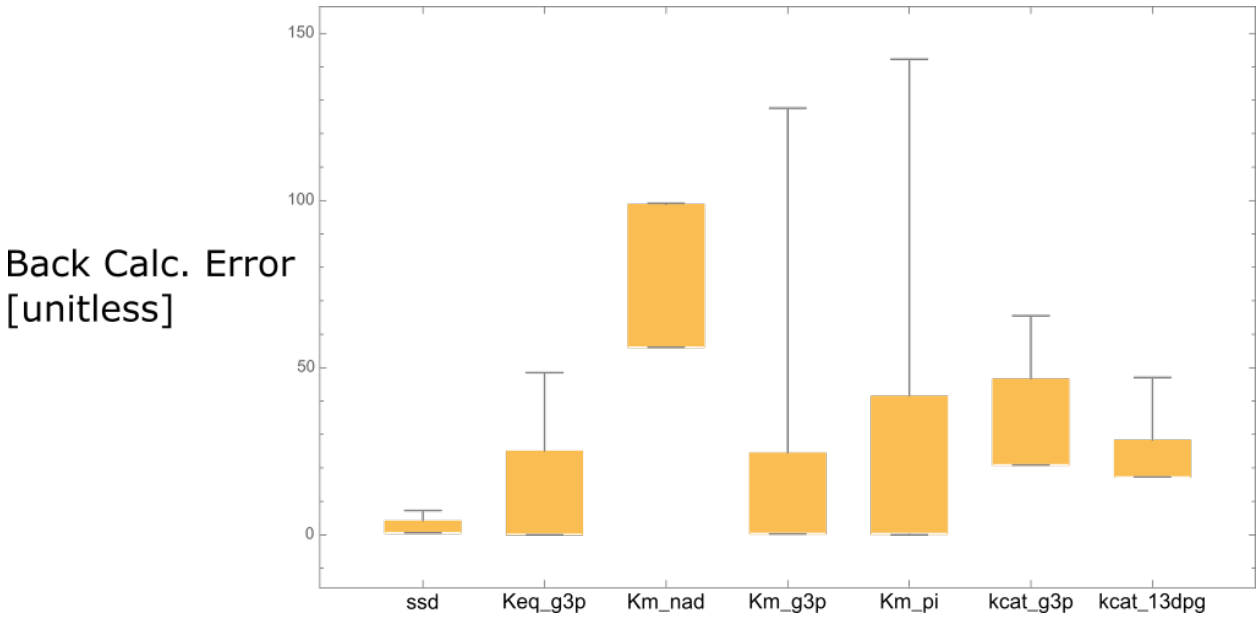

Figure 3.5: A Box Plot of the Error between GAPD's Calculated and Actual Values.
In this plot, the errors of each of the recalculated kinetic parameters are shown.

3.3. Construction of a Tryptophan Kinetic Model using a MASSpy-based Workflow

3.3.1. QP Optimization and Thermodynamic Adjustment

The sampled concentrations of metabolites are plotted against the initial concentrations of the metabolites. In the top graph, these values are taken from different literature sources from experiments of *E. coli* grown on glucose. The bottom graph uses values taken from the thermodynamically consistent values from shifted values from Figure 4.1. The largest metabolite shift is anthranilate, which in this diagram shifts due to it being set at a high for the model used for two of the case studies. While the final shift in metabolites seems to be less, this is a result of a fed adjustments. First, the thermodynamic buffer used in the QP optimization step was set to a low value. Secondly, values were obtained for a greater number of metabolites. Finally, these values were adjusted to be thermodynamically consistent with the rest of the entire stoichiometric model using thermodynamic FBA. As a result, the values shown are consistent with the equilibrium constants, fluxes and other metabolite concentrations. This difference in the

results suggests that utilizing metabolite concentrations that have been adjusted in accordance with a larger model likely is a better method, and likely means that the concentrations are consistent with each other.

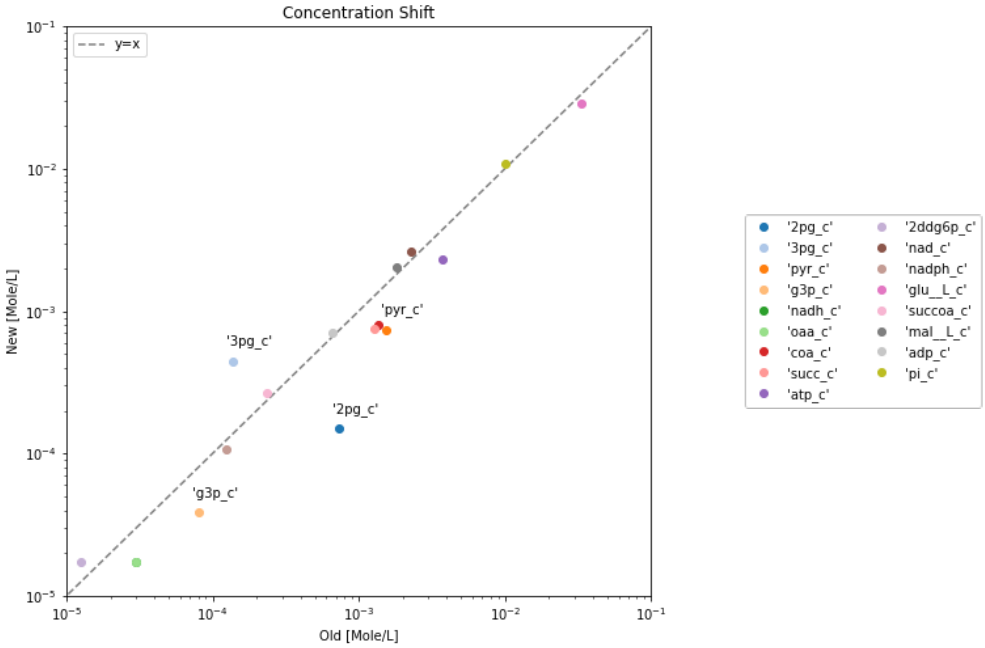

Figure 4.1: Concentrations of Metabolites Before and After Thermodynamic Adjustment. Metabolites are labeled if the fold change is larger than $10^{0.3}$. Concentrations after thermodynamic adjustment are plotted against their values before the thermodynamic adjustment. The center line is equal to y = x.

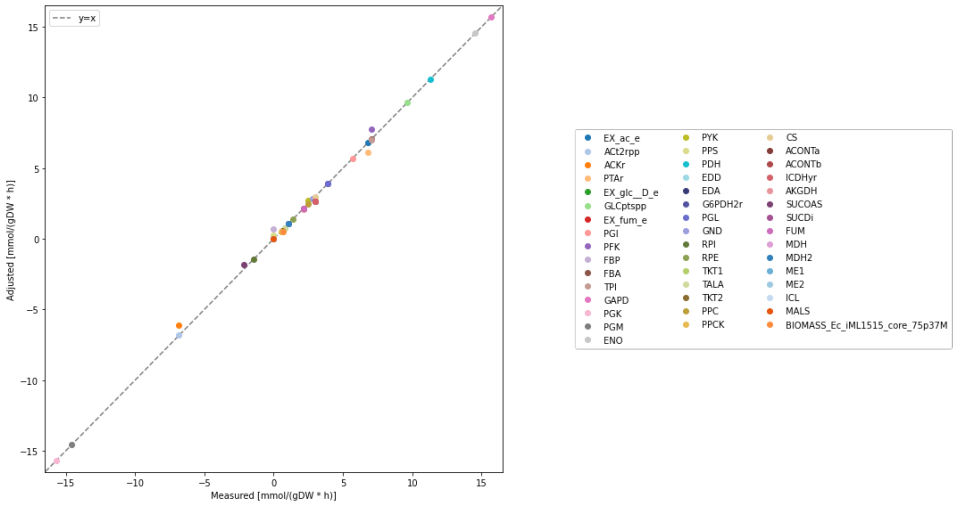

Figure 4.2: Optimized versus Measured Fluxes of Reactions in the Model.
Above is pictured the FBA optimized fluxes in the model. Adjusted fluxes correspond to the fluxes of the optimized, whereas the measured fluxes correspond to the original literature flux values.

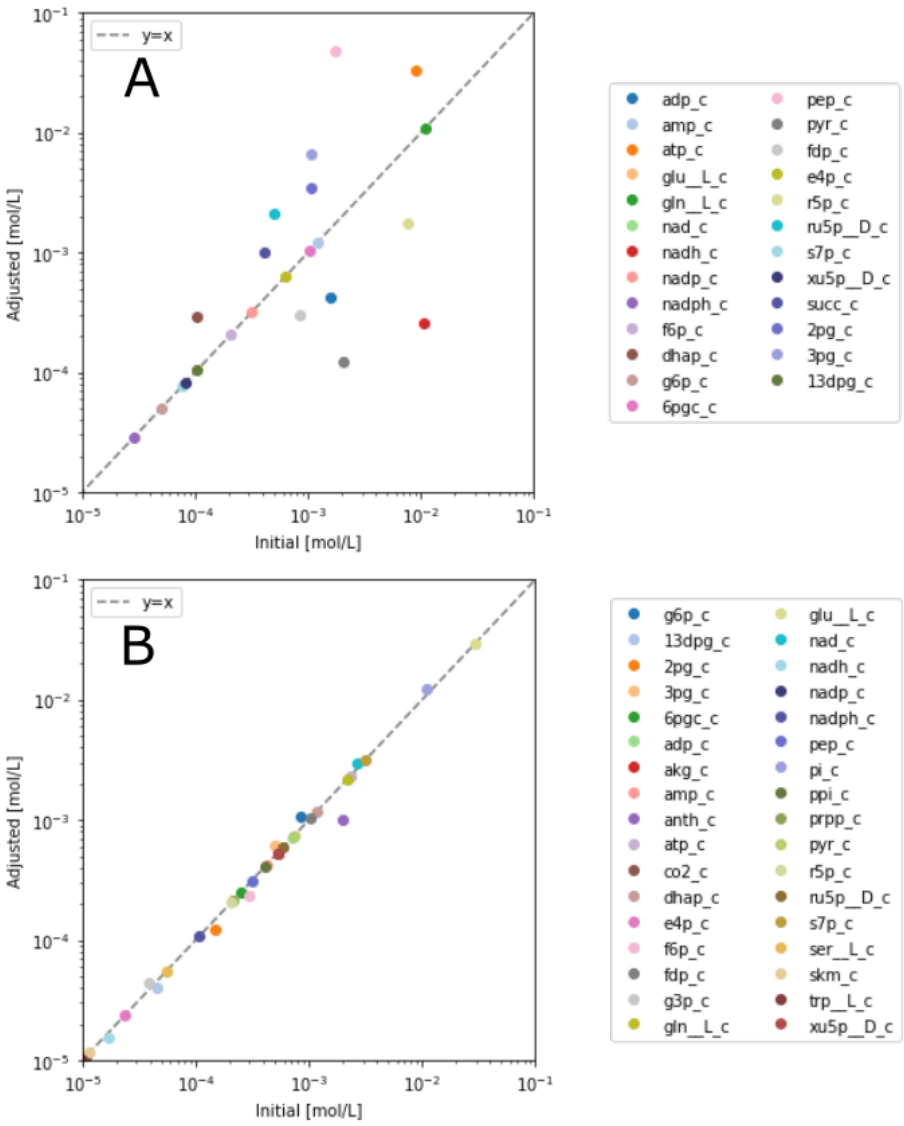

Figure 4.3: Concentrations of Metabolites in the Model Before and After QP Optimization and Sampling. Metabolite values are represented by different colors in each scatterplot. Graph A shows the shift from a model using data from different sources is graphed. The dotted line represents the line of the adjusted value of the metabolite concentration being equal to the literature value. B shows the shift of metabolites from the thermodynamically adjusted metabolite concentration dataset. In this graph the dotted line represents unity. Different colors correspond to different metabolites between graphs due to different numbers of metabolites having literature values present. The initial optimization was conducted using a smaller set of literature concentrations.

After assembling the enzyme modules, an additional thermodynamic adjustment was performed in order to ensure the metabolites of the model were consistent thermodynamically.

The result of this shift can be seen in Figure 4.1. Labeled metabolites are ones with a fold change greater than $10^{0.3}$. The largest fold change in metabolite concentration was in 2pg, followed by 3pg. The largest fold change of the equilibrium constants for reactions was PGMd, with a fold change of 0.2.

Comparing the shifts between the metabolite concentrations before and after the QP optimization step between thermodynamically adjusted values and non-adjusted values shown in Figure 4.3, one sees less of a shift between the adjusted values and the initial values. This demonstrates the usefulness of obtaining a consistent set of metabolite initial conditions.

Comparing the shifts between the fluxes in Figure 4.2 , one can see that when using the thermodynamically adjusted values, there is not much flux adjustment.

3.3.2. Mass Action Model Setup

After adjusting the metabolite concentrations and reaction fluxes, the concentrations, fluxes, and equilibrium constants, are used to generate a mass action model using pseudo-elementary rate constants, (PERCs). Boundary reactions for the model are constructed based off of whether all of a metabolite is sufficiently created and destroyed within the model, according to its fluxes. After the construction step, among other checks, boundary conditions were checked to make sure that they had concentrations that are within the range of their literature or calculated value, and are shown in Figure 4.5. Of note is the fact that there are 30 boundary conditions and therefore exchange reactions for the metabolites shown in the figure. This suggests that, due the the boundary conditions being present there are other reactions in the whole stoichiometric model that utilize these metabolites, but are not present within the model. Furthermore, due to the amount of exchanges within the model, there might be potential issues with the ability of concentrations in the model to increase outside of the range of the boundary conditions.

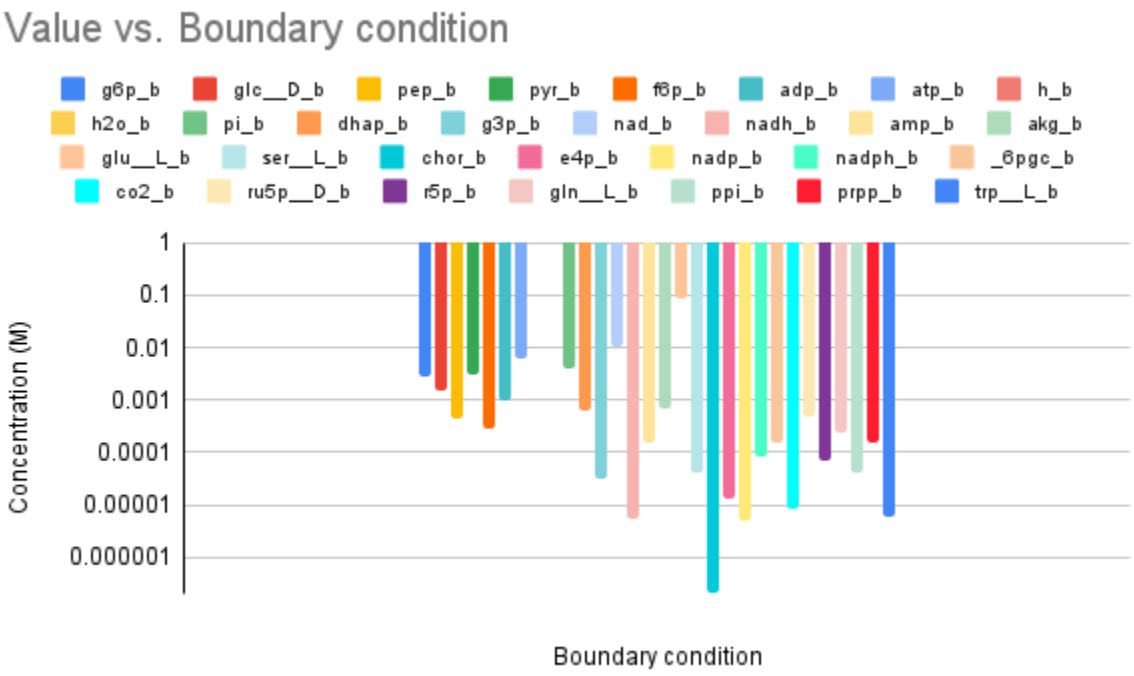

Figure 4.5: Boundary Conditions of the Model.
Plotted above are the boundary conditions for metabolites of the mass action model. Each boundary condition is plotted on the bar chart with a specific color.

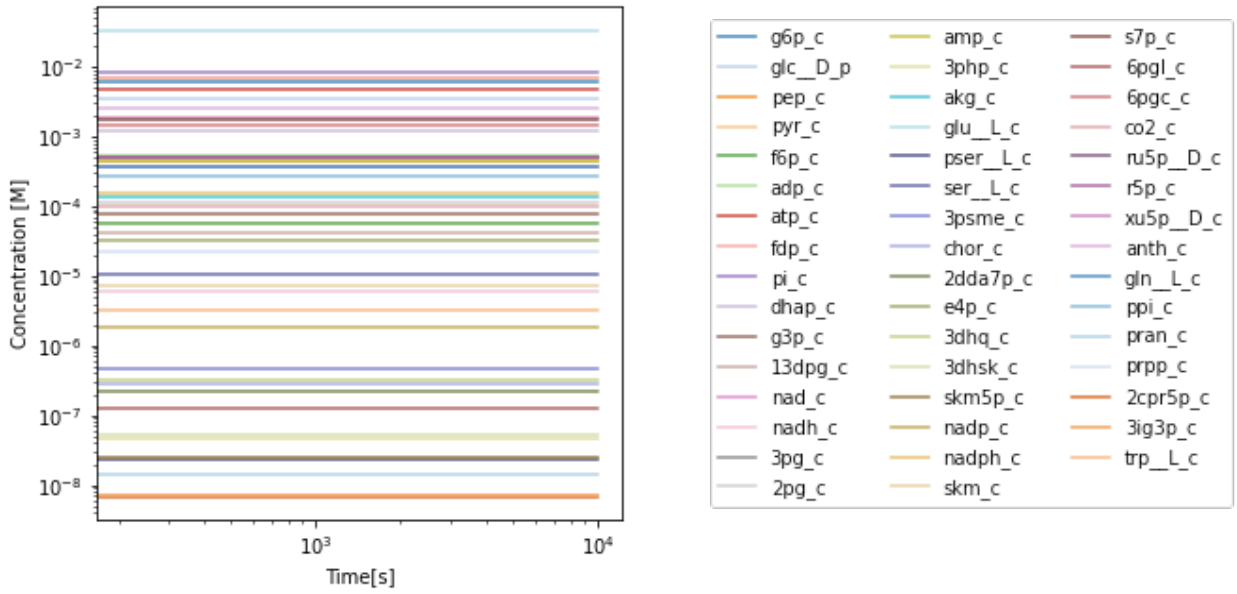

Figure 4.6: Verification of Steady State of the Mass Action Model.
The graph above shows the concentration of each metabolite in the mass action model over time, and in this case verifies that the model is stable and at steady state.

After simulating the model to steady state, the model was checked to ensure the model was at steady state, as seen in Figure 4.6. Each line in the figure represents a metabolite concentration at a given point in time. Looking at the graph, it seems that the model is indeed at steady state prior to enzyme module incorporation, indicating that the model tested is stable prior to incorporation of enzyme modules. In cases where the enzyme module isn't stable, this intermediate step can be a non-steady state.

3.3.3. Enzyme Module Export Results:

After simulating models to steady state, the values of the average of total concentrations of each enzyme were compared against the Sauer data set to determine quality of fit (data not shown). In 31 out of 42 cases, enzyme concentrations were within two orders of magnitude of their literature values. Figure 4.7 describes the comparison between two sets of enzyme concentration data, non-flux fit and flux fit. These results were compared for the sake of

determining if there were marginal improvements as a result of Flux fitting. From this graph, it shows that the flux fit has more calculated over literature values that are within two orders of magnitude of 1 with 31 reactions within that range opposed to 27 reactions within the range in the non-flux fitted data set. Furthermore, the graph also illustrates that the maximum Calc/Lit ratio decreased between the two sets of models, where it was previously approximately 10^9 for PSP_L in the baseline model, whereas it was at most on the order of 10^5 for the flux fit model. However the flux fit model had different reactions that were elevated compared to the baseline. The largest difference between the calculated concentration and the literature concentration in the flux fitted is found in the enzyme TRPS1, which is a 881000 times more concentrated than its literature value as seen in Table 1.1 Contributing factors to these concentration differences can be non-thermodynamically optimized metabolite concentrations, closeness to equilibrium, choice of reverse flux ratio, and quality of MASSef fit.

Table 1.1 Top Five Highest Enzyme Concentration Ratios.
This table shows the top five highest calculated over literature ratios for the most recent iteration of the model. Literature concentration refers to the concentration of the enzyme obtained from Gerosa et al. 2015 while the calculated concentration refers to the total concentration of the enzyme calculated during enzyme module incorporation.

Reaction Name	**Literature Concentration [M]**	**Calculated Concentration [M]**	**Calc/Lit**
TRPS1	1.04E-05	8.81E+00	8.47E+05
FBA1	4.76E-05	2.77E+01	5.83E+05
PFK1	3.44E-06	3.47E-02	1.01E+04
PGCD	2.69E-05	1.49E-02	5.53E+02
PYK1	4.51E-06	2.38E-03	5.28E+02

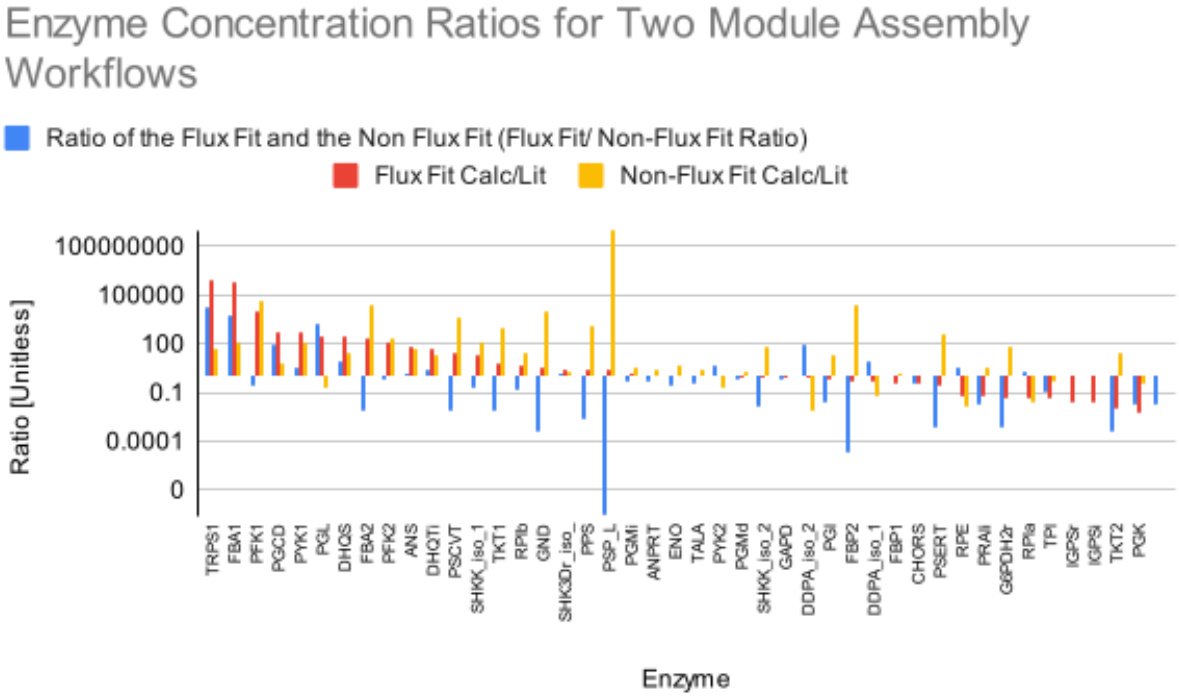

Figure 4.7: Comparison of Calculated Enzyme Concentration/Literature Concentration Ratios between the Flux Fitted model and the Non-Flux Fitted Model.
Red bars indicate the ratio of the calculated enzyme concentration of the current model. Yellow bars indicate the previous ratio of the model prior to flux fitting, and the blue bars represent the difference between the yellow and red bars.

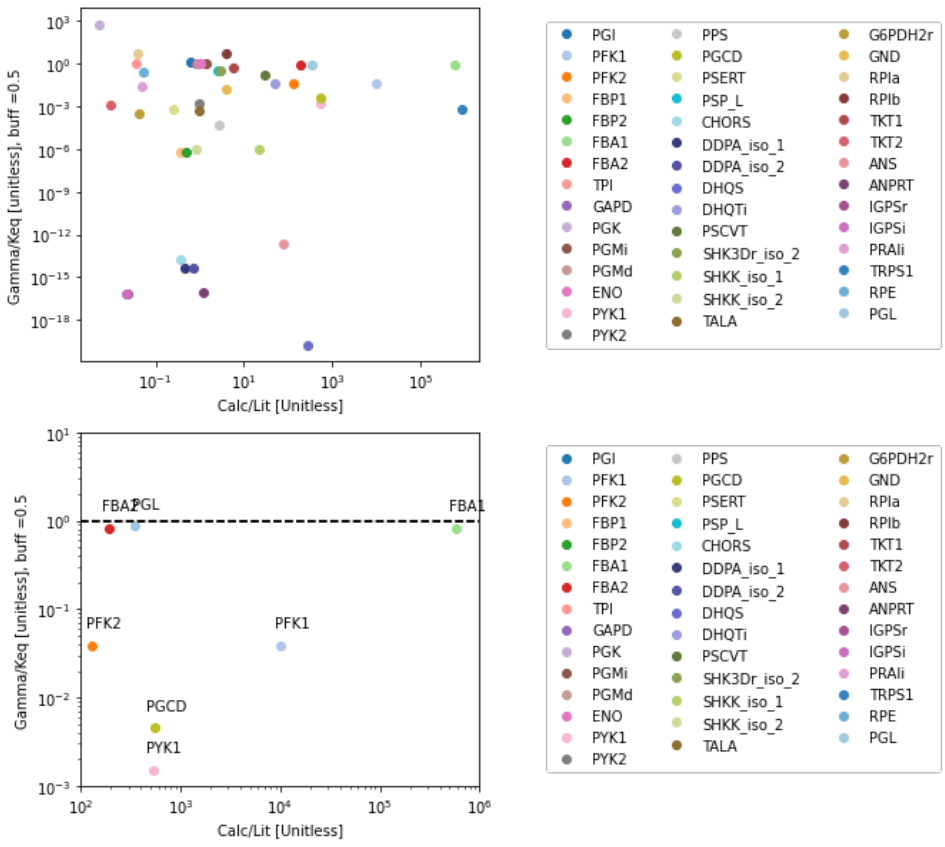

Figure 4.8: Comparison of Γ_{ss}/K_{eq} and Enzyme Concentration Ratio.
The top graph depicts Γ_{ss}/K_{eq} over the ratio of the calculated concentration of the enzyme to the literature concentration of the enzyme. Different colors correspond to different reactions. The bottom graph is an expanded view of the larger enzyme concentration ratios, with a line at $\Gamma_{ss}/K_e = 1$

Figure 4.8 shows the comparison of mass action ratio over K_{eq} versus enzyme concentration, in an attempt to determine whether closeness to equilibrium is correlated with enzymes having a higher concentration. Looking at the results, enzymes with a ratio of Γ_{ss}/K_{eq} close to 1 did not necessarily have a high enzyme concentration. Some enzymes with a high concentration have a Γ_{ss}/K_{eq} close to one (FBA2, PGL, FBA1), however others, such as PFK1 and 2 as well as PGCD, and PYK1 were not close to unity. This indicates that while closeness to equilibrium might play a role in poor fit, it isn't solely casual, and is likely a contributing factor to a poor fit.

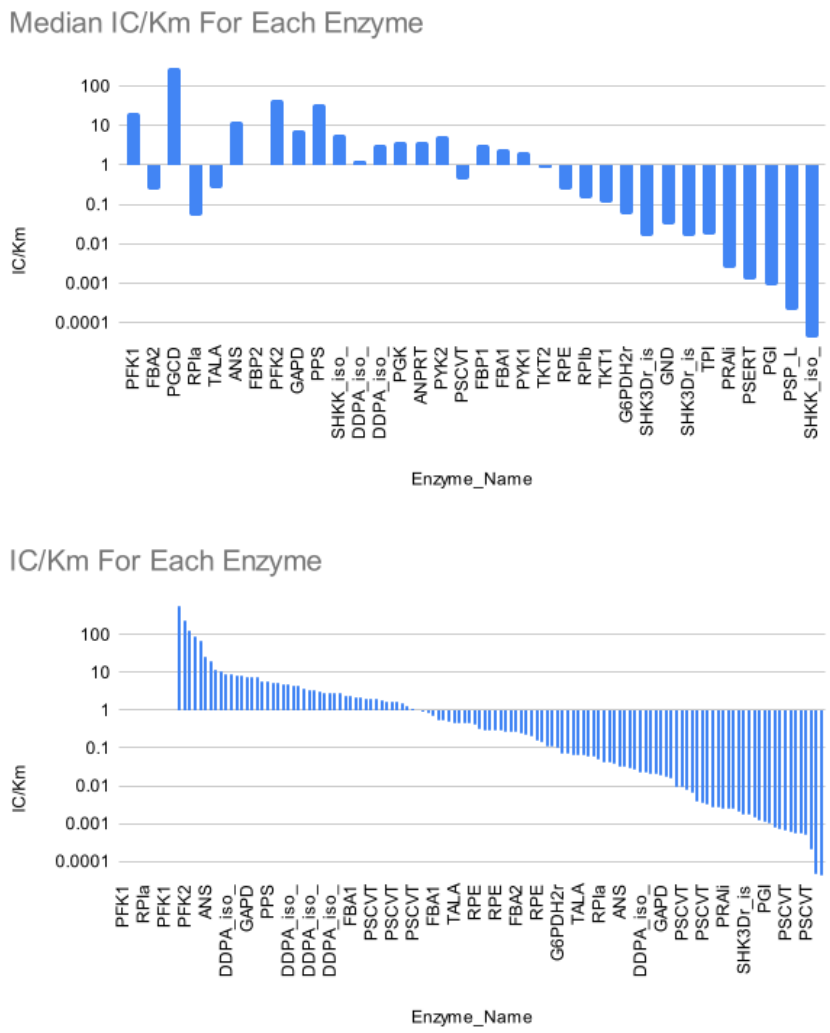

Figure 4.9: IC/K_m Graphs.
The top graph shows the Ic/Km for the median Km of each enzyme in the model. The bottom graph depicts the Ic/Km for each Km obtained for each enzyme in the model.

Figure 4.9 shows the IC/K_m value for each metabolite for each K_m used in generating the enzyme modules. PSCVT being a very studied enzyme features many of the K_m values. Noticeable are many K_m values for which the IC is lower than the K_m significantly, suggesting that these enzymes are at a significantly decreased activity. Looking at the median K_m values, one sees that the median K_m value for over half of the enzymes is greater than the initial

condition for that metabolite. This suggests that many of the enzymes in the model might not initially be at full activity.

Another culprit behind higher concentrations for enzymes than found in literature could be higher IC/K_i values. Table 2.1 shows the top five IC/K_i values. Noticeable is the presence of several of the enzymes with the highest calculated concentration, such as PGCD, FBA2, FBP2 and PFK1. In this model, PGCD, FBA2, and FBP2 likely are significantly inhibited, which might contribute to their elevated concentrations. Furthermore, since this model uses a flux split of 50% between isozymes, it should not be surprising that in the case of FBA1 and PFK2 also have calculated concentrations much higher than literature.

Table 2.1: Table of the Top Five IC/K_i Values.

Enzyme	Metabolite	Value	Parameter	K_i	IC/K_i
PGCD	ser__L_c	1.14E-04	['Kiu']	3.80E-06	30.09760474
FBA2	dhap_c	1.95E-04	['Kic']	9.00E-06	21.61639611
FBP2	pi_c	4.20E-03	['Kic']	0.00035	11.98720196
PGCD	ser__L_c	1.14E-04	['Kinc']	1.25E-05	9.14967184
PFK1	adp_c	6.31E-04	['Kic']	0.0002	3.154870777
FBA2	dhap_c	1.95E-04	['Kic']	0.00013	1.49651973

3.4. Case Study 1: Simulating the interventions in S28

To apply the model constructed, a strain produced by Chen and Zeng 2017 was examined. The authors of the study aimed to create a tryptophan production strain using rational methods. To mitigate the drain of plasmid expression on the cell, modifications were made to the stain, S28's, chromosome. In this strain, the wild type genes for *aroH*, *aroG* and *aroF* were deleted, in addition to the genes for the transporters *mtr, tnaA,* and *tnaB.* Furthermore, genes for feedback resistant variants of *aroG* and *serA* were incorporated into a

plasmid with a *tac* promoter and a strong ribosome binding site. This plasmid was incorporated into the chromosome where *aroH* was. Furthermore, a feedback resistant variant of *trpE* was substituted for the wild type gene with a strong ribosome binding site. In addition to this the strain replaced the wild type promoter for the tryptophan operon with the *trc* promoter in order to alleviate the inhibitory effect of the tryptophan. In order to study the effects of the interventions this study obtained high frequency sampled metabolite concentrations in order to study the changes in intermediates.

These modifications were modeled by overexpressing the *trpE:A* operon, *aroG*, and PGCD and, setting the forward rate constants for the inhibition equations of PGCD, ANS and ANPRT to zero. Furthermore, the study knocked out the permeases of tryptophan from the model, which due to the way that tryptophan export is set up, and due to a lack of permease kinetics is not reflected in the model. Furthermore, other modifications outside of the scope of the model are not modeled, such as the deletion of the lambda recombination system, and *rpsL*.

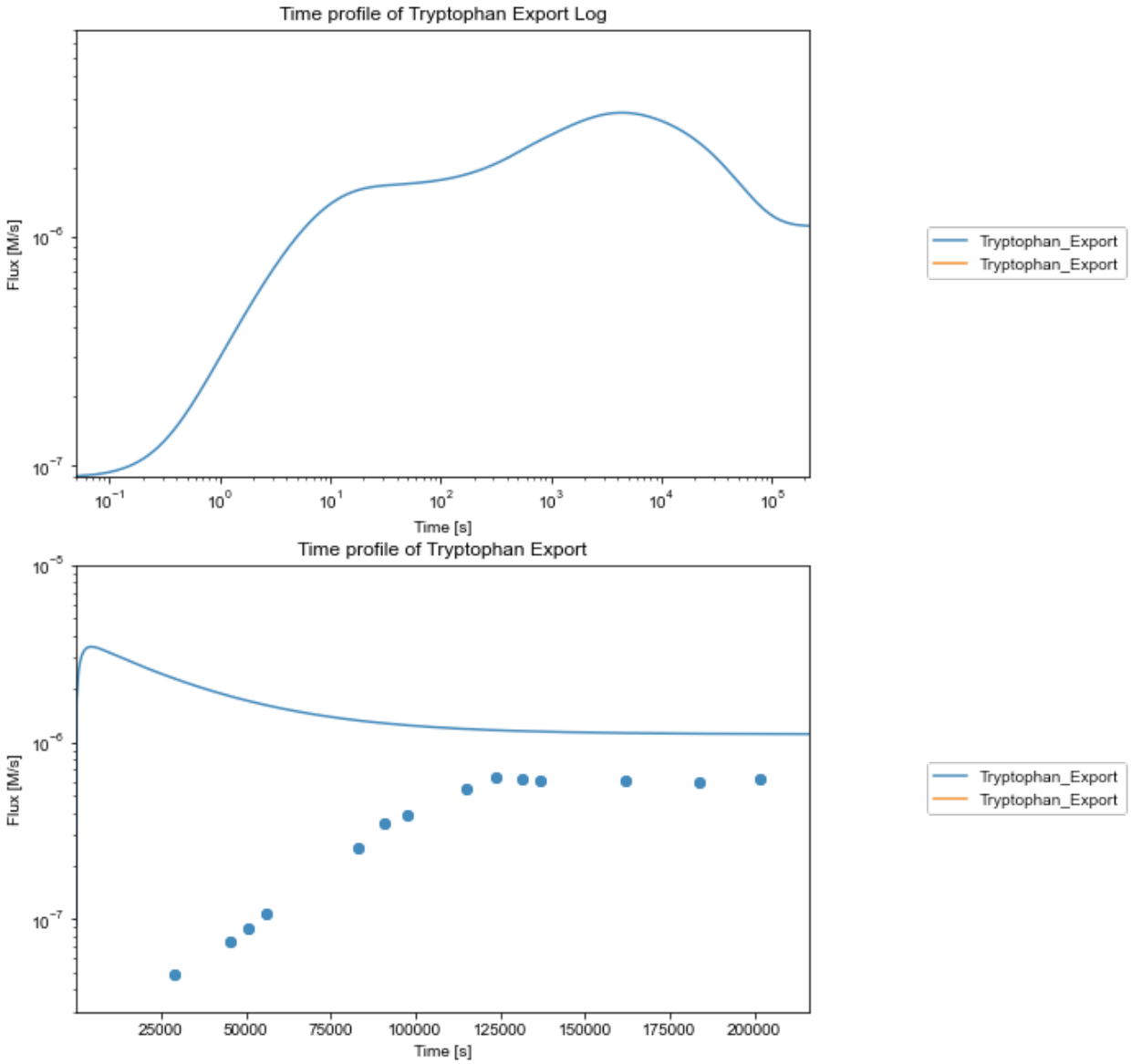

Figure 5.1: Tryptophan Export Flux of the Model in Response to the Interventions in S28.
The top graph depicts the tryptophan export flux over several different time scales, with the unmodified base model not shown due to its flux being zero. The Bottom graph depicts the tryptophan export flux over the duration of the experiment. Blue dots represent the productivity values taken from the experiment, which have units of M/s

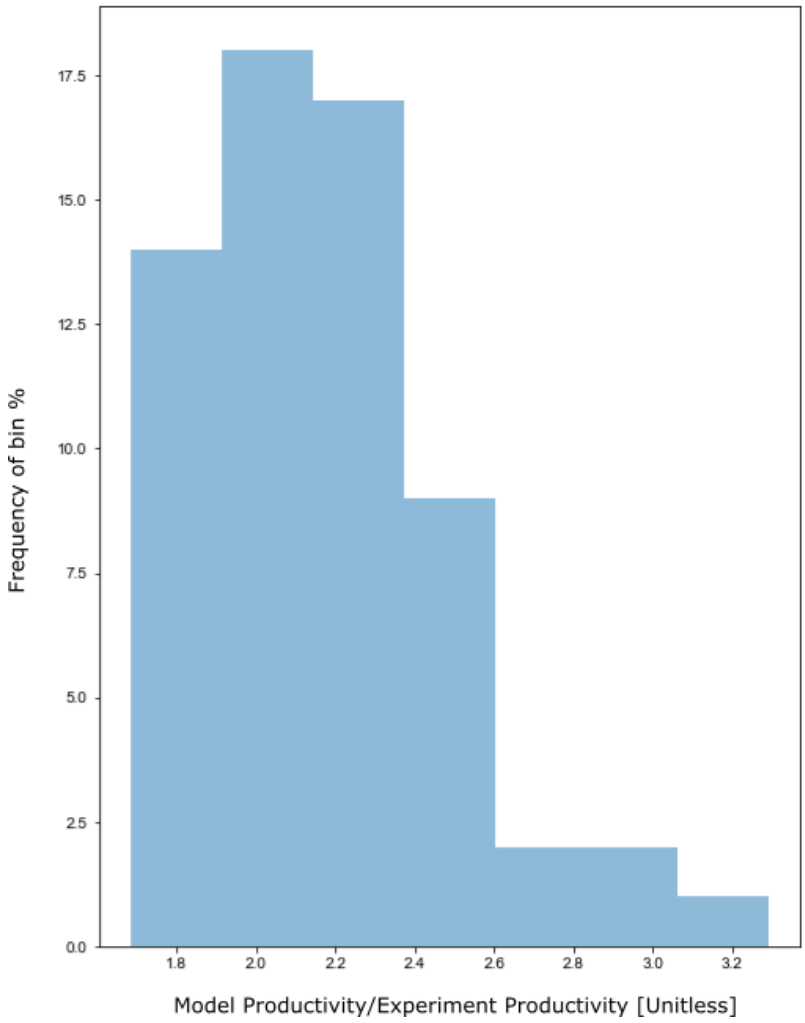

Figure 5.2: The Distribution of Ratios of Final Tryptophan Export Flux Over the Final Productivity Value. The frequency corresponds to the frequency of a given ratio value being categorized within one of the bins. The average ratio was 2.17 with a standard deviation of 0.316.

The effects of simulating the perturbations of the model are shown in Figure 5.1 On Top is a graph showing the dynamic response of the model to the changes mentioned in the methods section. The Dynamic response of tryptophan export, defined as the difference between the Sink of tryptophan needed for biomass and other reactions within the network is shown. The overshoot in this model occurs within a faster time scale than the change in the experiment, on the order of hours, or about 5000 seconds. Compared to the study's productivity data, the overshoot is higher and occurs about twenty hours earlier. Note that the literature

results in this case show a final productivity of about 6.7 times ten to the negative seventh. Compared to a final productivity at sixty hours of our model of 1.06E-6 M/s, the predicted productivity endpoint is only off the experimental value by 1.89 fold. To verify that this was not a fluke, an ensemble of sixty three models was run through the simulated perturbations corresponding to the experiment. The ratio between the final value of tryptophan export and the final productivity value from the study for each of the models was stored with results plotted in the histogram Figure 5.2. The standard deviation was 0.316 and the mean was 2.17. Since the average value was a 2.1 fold difference with a standard deviation of 0.6, one can presume that the initial ratio of 1.89 was not due to chance.

The above again shows a degree of agreement in performance between simulated perturbations in bioreactors, albeit with dynamics that are much faster than the experimental data. This is likely due to the fact that the increase in enzyme concentration, and the feedback resistance are modeled as occurring at t=0, instead of gradually building up within the system and being subject to constraints on protein production. Furthermore, since this model does not account for transcriptional dynamics and regulation, it is likely that the network serves to stabilize the cell compared to just the mass action kinetics of the reactions themselves.

3.5. Case Study 2: Simulating Anthranilate Feed-Forward Inhibition of IGPS.

The second case study is based on Chen et al. 2018. In this study the authors wanted to examine if, based on findings from other studies, anthranilate feed-forward inhibited enzymes in the tryptophan synthesis pathway. After conducting structural studies and relative activity experiments, the study found that anthranilate non-competitively inhibited IGPS, and obtained an IC50 of anthranilate for IGPS. In order to study the effects of anthranilate inhibition on tryptophan production, the study generated a feedback resistant form of IGPS. Furthermore,

based on studies of IGPS from other organisms, IGPS from *A. niger* and *S. cerevisiae* were incorporated into S28 from Chen and Zeng 2017. Both of these enzyme variants are activated by anthranilate instead of inhibited. The study found that strain with the resistant form of IGPS exhibited reduced growth compared to the wild type form. Furthermore, upon adding IGPS from *A. niger*, increases in yield and titer were obtained compared to the overexpressed wild-type control.

The authors found that anthranilate inhibited IGPS activity with an IC50 of 0.70 mM. Using a ½ IC50 ~ K_i, Furthermore, the authors used a pTrc99A, a plasmid with a trc promoter with a copy of feedforward resistant IGPS. While the authors studied the improvement on yield of *trpC* from *A. niger*, and *S. cerevisiae*, due to a lack of kinetic data, the activating effects of anthranilate in those two enzymes were not able to be modeled. As a result, the comparison for this case study was the S28 interventions combined with feedforward inhibition and resistance to inhibition of anthranilate. This was modeled by reducing the forward rate constant of both inhibition reactions to zero, preventing any inhibition of anthranilate. Furthermore, the wild type *trpC* overexpression in a plasmid behind a TAC promoter was compared to the preexisting *trpC*.

Figure 6.1 shows the response of the model to the simulated condition of anthranilate resistance. The dynamic profile of this response shows an immediate overshoot followed by a gradual decline. This occurs on the order of about a few hours, and reaches approximately its final flux fby the 100000 second mark, or about the 25 hour mark. In contrast to this, the tryptophan productivity (dots below) were approximately one order of magnitude less, with a later peak. In this figure, the plot of the response of S28 interventions is plotted with the response of the model to anthranilate resistance. The model shows negligible differences between the two simulated strains, and a miniscule difference of two final productivities being 6.1 times ten to the sixth, which is within the standard deviation of the previous case study. As a consequence, the impact of anthranilate inhibition seems to be limited.

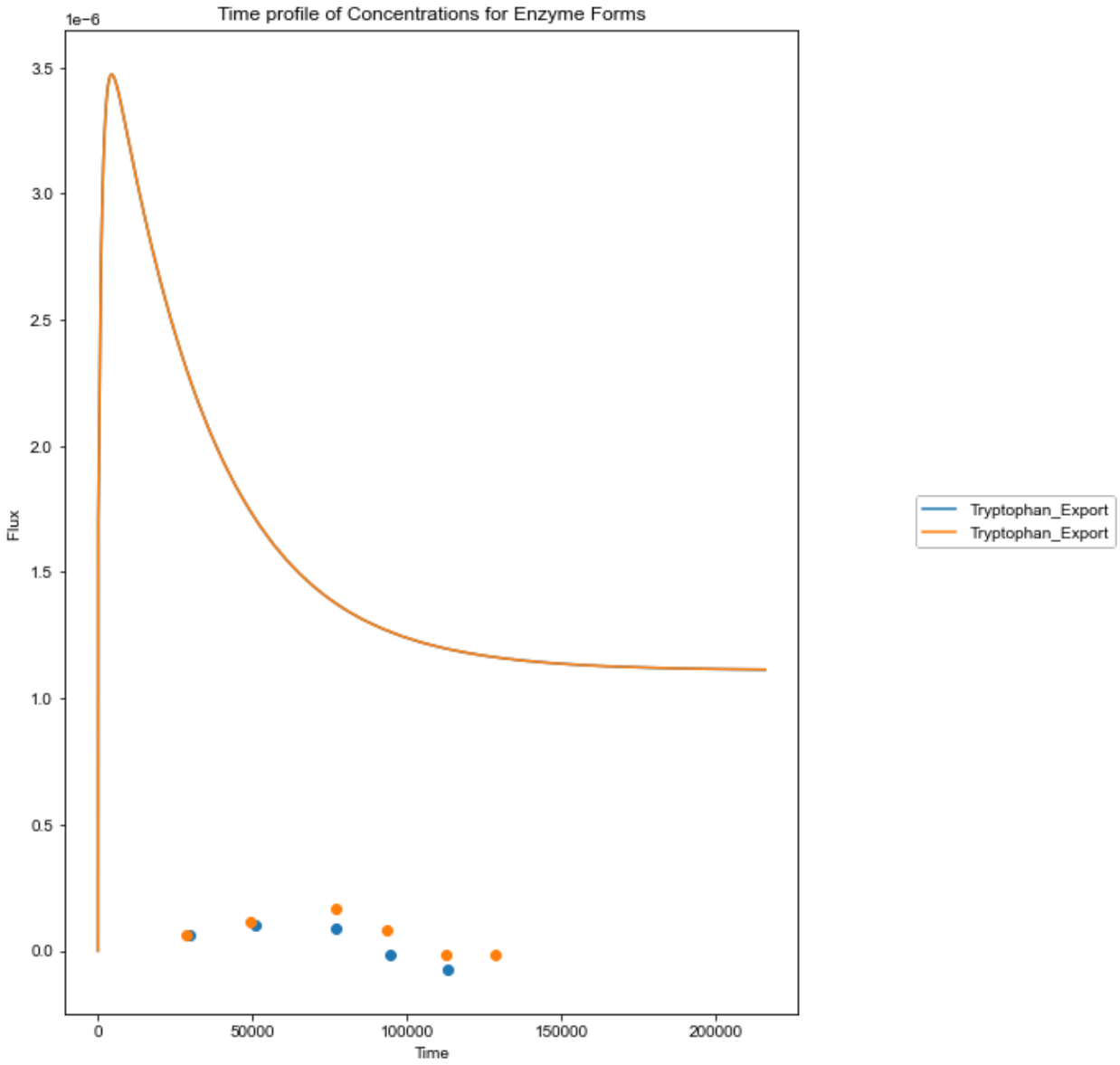

Figure 6.1: Simulation of the Interventions of Two Strains of S028 with (Blue) or without (Orange) Anthranilate Resistance.
The blue line represents the response of the model to anthranilate resistance, while the orange line represents the response of the model to the interventions done in strain S28. The blue points correspond to the productivity values of the resistant strain derived from Figure 5c from the study. The orange points correspond to the productivity values derived from the non-resistant strain.

Since this case study does not model the effects of IGPS from *A. niger* or *S. cerevisiae* , perhaps the very small change productivity is not surprising. With a K_i of 0.00035 M, the inhibition resulting from anthranilate would seem to be mild except at high concentrations. While the batch results showing decreased tryptophan production and productivity from the *trpC* anthranilate resistant strain, the paper cited that the mutation conferring feedback resistance also decreases the affinity for substrates, causing a decrease in rate. This change in rate was not modeled in this case study explicitly, and if included might produce different results. Furthermore, the difference between the productivity of the model and strains are noticeable in this case study. The average ratio of the final productivity of the anthranilate resistant strain and the modeled anthranilate resistant strain is -75 with a standard deviation of 6.21. This indicates that the model does not replicate the strain in this bioreactor set up well. Of note is that the data points used for this graph were taken from batch reactor results, which might not be reflected in the model, which uses exchange reactions with a known boundary concentration. In a batch reactor, external metabolite concentrations would decrease over time and not be replenished, causing a later decrease in productivity due to shifting growth phase. The model does not currently account for the decreasing external concentration of metabolites, though it could be modified to do so if the external concentrations were not clamped.

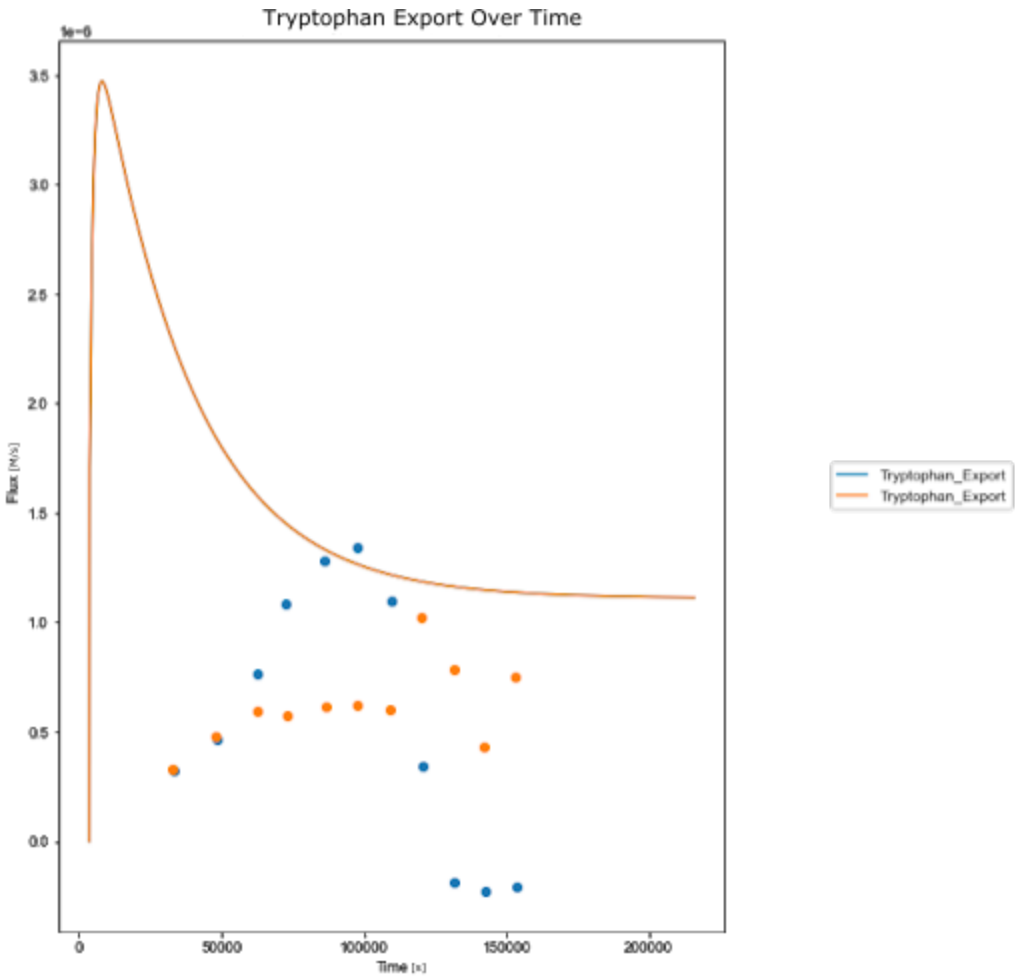

Figure 6.2: Simulation of the Overexpression of Wild-type *trpC*.
The response of the model to *trpC* overexpression behind a TAC promoter (blue line) is compared against the response of the strain with just the plasmid with the TAC promoter in S28. The dots correspond to productivity values obtained from Figure 8b of (Chen et al. 2018). Blue dots represent the productivity of the strain with the TAC plasmid without *trpC*, while orange represents the strain with *trpC* in front of the TAC promoter.

Figure 6.2. shows the response of the model to the intervention whose result are shown in Figure 8b from (Chen et al. 2018), that of the presence of the tac promoter plasmid versus that plasmid carrying a k-12 *trpC* gene. From the figure, it appears that there isn't a difference between the models, suggesting that the three times change in IGPS concentration alone isn't sufficient to change the kinetics greatly. This is consistent with the data points gathered, as the concentration end points of both strains are similar, and the two sets of points, while following

different trajectories, end up at similar points. The cause of the difference between these two strains, thus might be caused by other factors not accounted for within this model. However, looking at the final values, the model of the TAC promoter strain is only 1.6 times more than the final productivity value of the TAC promoter only strain from the study. This at least suggests that there is some commonality in final steady states between the study and the model.

3.6. Case Study 3: Simulating the Effect of a Feed Rate Profile on the Model

This case study focuses on modeling the effect of a feed rate profile as seen in (Dodge and Gerstner 2002). In this paper, the authors attempt to develop a feed rate profile to minimize excess glucose and therefore reduce the formation of acetate, which can inhibit tryptophan synthesis. To do so, the study initially attempted to add glucose in with a linear feed rate profile. However, this led to accumulation of glutamate, which reduced the production of tryptophan. Genetic modifications made to the strain are to improve tryptophan production were placing feedback resistant *trpE*, *aroG*, and *serA* genes behind a LacUV5 promoter, as well as similarly overexpressing the entire tryptophan operon. The wild type *serA* gene was knocked out, so that the plasmid would be maintained without need for antibiotic.

To model this the rates for the glucose feed rate profiles were determined from the graph and were modeled as a change to the boundary metabolite b. Furthermore, overexpression was modeled by increasing the concentration of each of the forms of DDPA_iso_1, ANS and PGCD respectively by 3, an amount corresponding to the change in efficiency due to deregulation. Furthermore, to model the final feed rate profile used in the study, a piecewise function was used comprising of the integrated forms of the equations making up each leg of the glucose feed rate profile shown in Figure 7.1.

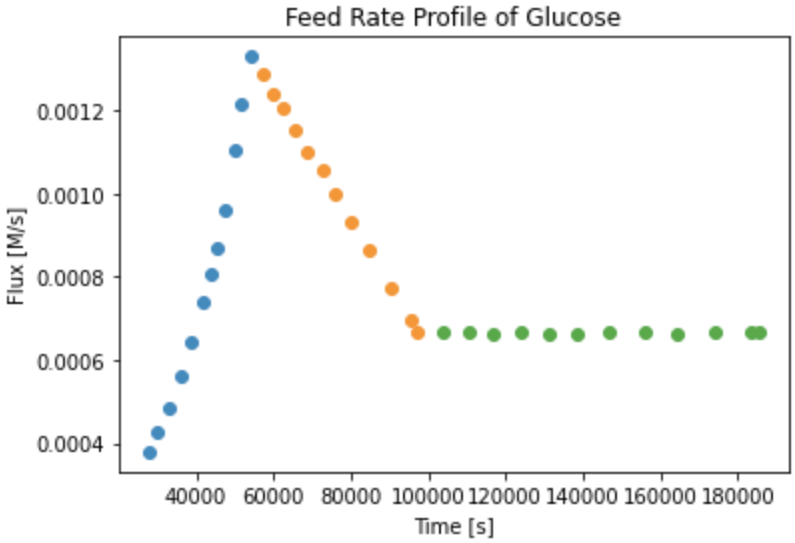

Figure 7.1: The Glucose Feed Rate Profile used in Case Study 3.
Points correspond to values taken from the feed rate profile graph from the paper. The first segment corresponds to an exponential increase up to a max flux (blue), followed by a linear decline (orange), followed by a constraint input flux.

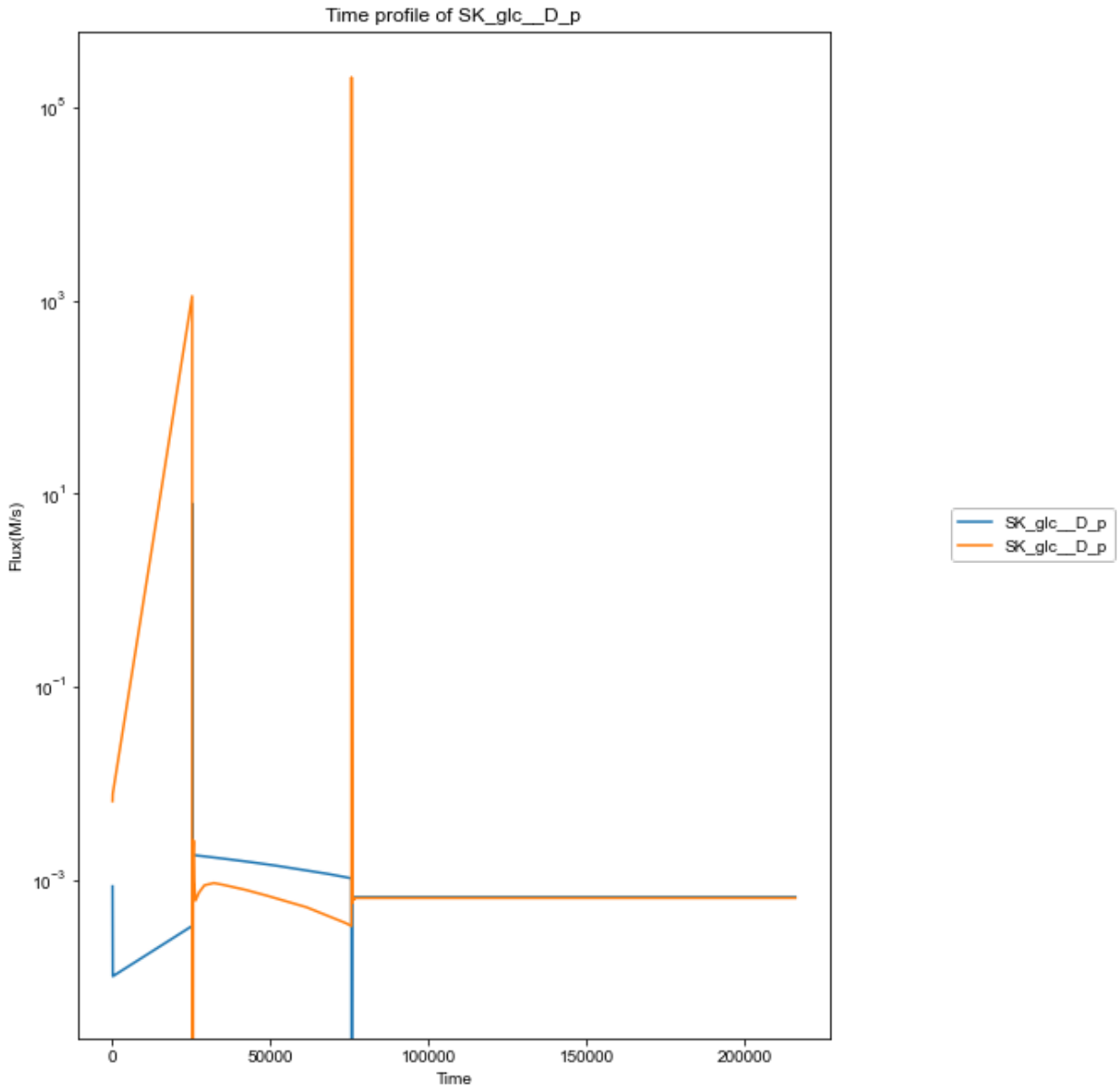

Figure 7.2: Comparison of the SK_glc__D_p Fluxes Perturbed by Two Different Feed Rate Profile Equations.
The feed rate profile perturbation obtained from fitting data points from fitting to a graph are in blue, while the feed rate profile perturbation obtained by integrating the equations provided in the paper are shown in orange.

The effect of the feed rate profiles on glucose import into the model are shown in Figure 7.2. In this figure it is apparent that the feed rate profile is affecting flux of glucose into the

periplasmic space with the first leg exhibiting an exponential increase (Y-axis is logarithmic) followed by a linear decrease, followed by a constant rate of feed. This is consistent with the feed rate profile graph shown in Figure 7.2.

The results of this are shown below in Figure 7.3. Looking at the graph, it is apparent that the feed rate profile used to perturb glc__D_b does not cause a change in tryptophan export flux larger than the effect of the overexpression on the model. However, of note is the fact that the peak of both of these legs approaches the peak productivity of the strain, and its final value is only 1.063 of the final value of the productivity. Looking at the top graph of glutamate concentration, one can see that compared to the non-perturbed model, glutamate concentration decreases and stays decreased. This is consistent with the sources findings that the modified glucose feed rate profile resulted in decreased glutamate accumulation compared to other feed rates. However the shape of the glutamate over time graph suggests that the change is more likely due to the effect of the overexpression and increased tryptophan flux, rather than the feed rate profile.

Since the bottom graph of the feed rate profile in Figure 7.3 perturbed strain shows an overshoot type behavior on a faster time scale than the feed rate profile, it seems that the feed rate profile is not sufficient to make a difference, at least not more so than the effect of overexpression and feedback resistance, which could be caused by several factors. Since the intervention was modeled as a change in the boundary metabolite over time, it might not cause a change in the import flux. Furthermore, in addition to the simultaneous overexpression of this model possibly causing this difference in dynamics, it could be possible that the lack of an explicitly tracked external metabolite generated by a flux might cause issues. This might be different if this feed intervention were modeled as a custom rate law. Promising however is that the final tryptophan export rate is similar to the final model productivity, suggesting at least some validity to the underlying model, and a similar limit to metabolism.

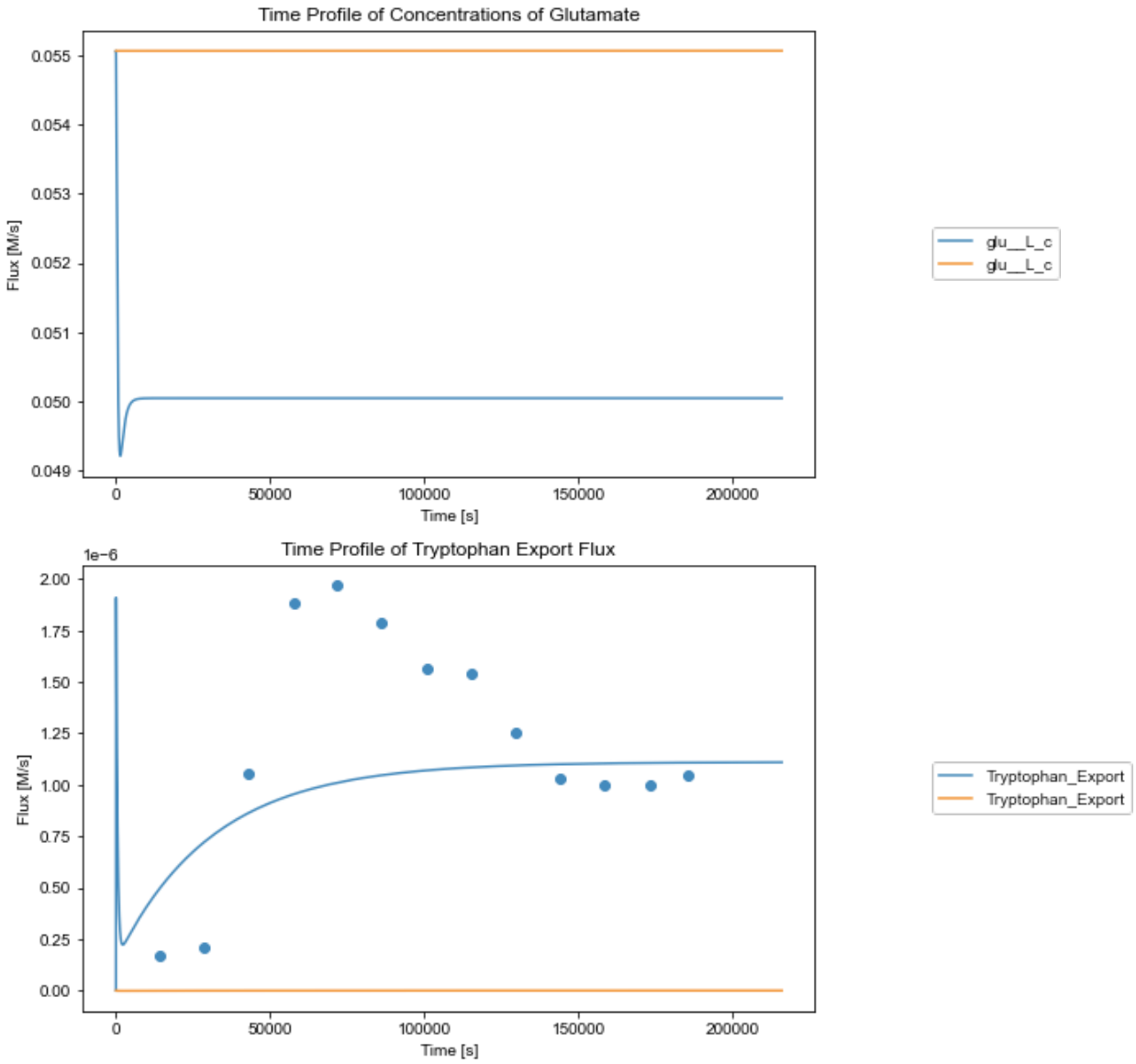

Figure 7.3: Tryptophan Export Response to the Feed Rate Profile.
Top: The model's response to the feed rate profile, overexpression and resistance perturbations are shown by the blue line, representing the concentration of glutamate over time. The orange line corresponds to the unperturbed model. Bottom: The model's response to the perturbations are shown in a graph of tryptophan productivity over time. The blue line represents the perturbed model, while the orange line corresponds to the unperturbed model. Blue points correspond to experimental productivity values obtained from differentiating tryptophan concentration measurements.

Discussion

In this work, a mass action model of tryptophan synthesis including microscopic enzyme kinetic constants was built and used to simulate perturbations as done in strain design experiments. A key goal of this work was to construct a mass action model that encompassed allostery, inhibition, and activation of enzymes in the model. Furthermore the goal was also to replicate the workflow done by (Haiman et al. 2021) on the network of tryptophan synthesis.

Underpinning this model was the baseline curation, which took the known set of reactions we had and translated them into kinetic data. One challenge of this work was to curate a set of data that could provide as much detail as possible to the model. Different enzymes had different amounts of data available, and consequently, required different amounts of gapfilling. This is reflected in some enzymes requiring gapfilling for k_{cat} and K_m values, and the dearth of enzymes with K_m values for both directions of the reaction. One consequence of this lack of complete data can be imprecise fits in the MASSef step. This is expressed by a wide distribution of k_{cat} values seen in the later steps. As a result of this limitation, the model can be limited by the limited available data.

Following the curation step, the MASSef workflow took these rate constants and converted them into elementary rate constants for an enzyme module. Due to differing amounts of data, different enzyme modules had varyingly wide ranges for k_{cat}s. Often, as seen in the comparison between PGL and SHK3Dr_iso_2, the more data points used, the worse the fit, but the more constrained the rate constants were. This need to improve how constrained the microscopic rate constant is, is also the impetus for obtaining K_m's for each metabolite for both directions of a reaction. Lacking K_m's in both directions can lead to a wider range of rate constants, and thus inconsistent results. One factor in fitting enzyme modules was not just how much data was available, but how much data should be used. Some enzyme modules, such as

PSCVT, have a very large amount of kinetic information, due to being an enzyme of interest for research. As a result much information is available, and not all of it is the same quality, and including more data, means a worse fit to the mechanism due to internal inconsistency. Furthermore, a greater amount of kinetic information also requires greater time to fit, increasing the time to fit a higher number of fits for a better enzyme module.

To set up the model that takes these microscopic rate constants, the MASSpy workflow was set up using a thermodynamic fit for the metabolite concentrations, fluxes and K_{eq}s. This was done in an attempt to ensure consistency between multiple sources of metabolite concentrations and fluxes. As a result, there was greater consistency in the results, manifested as smaller adjustments in the metabolite concentration following the QP optimization. Furthermore, fluxes also adjusted less due to similar reasons. This might be a good approach to take when incorporating new metabolomic data for expanding model coverage. This is critical, as metabolites lacking initial conditions often have very low solved values, being assigned a value around the lower end cutoff value of 1E-7. Prior to finding values for tryptophan, and anthranilate, these metabolites had a very low value from the flux solution. Thus, using a thermodynamic adjustment of the metabolite, flux, and K_{eq}'s one can ensure more consistent results and obtain greater model stability.

Another key aspect of the model is the assembly of the model itself from a PERC based model. Looking at the results of the model construction, one can see that the mass action model is able to obtain a steady state. However, a glance at the large number of boundary conditions for the model, indicates that there is a large amount of reactions not accounted for within the model. Due to these being modeled as boundary reactions, it could mean that the model is limited to the neighborhood of the fluxes used to set up the model. Furthermore, this compression of dynamic phenomena into exchange reactions could also lead to unrealistic responses to perturbations seen in the case studies.

Upon constructing the model, the concentration ratios of the enzyme modules were compared. While it might seem that the concentration of the modules are off, with only 31 of 42 enzyme modules being within two orders of magnitude of their literature value, this number is the result of implementing flux fitting, thermodynamic adjustment, and refinements to the fitting workflow. Factors that could lead to increased enzyme concentrations can include inhibition, closeness to equilibrium, low IC/km values, and bad enzyme module fits. Furthermore, other improvements can be made by adjusting other parameters of the model that can increase fluxes relative to the thermodynamics such as by altering the irreversible flux split. Looking at the IC/K_m values for the enzyme modules, one can also see that many of the enzyme modules had metabolites whose concentrations were below their K_m's significantly with enzymes such as SHK3Dr_iso_1, and PSCVT. This indicates that many of these enzymes are likely operating significantly below their maximum flux, which, while indicating room for responsiveness to perturbations, also means inefficiency. Furthermore, looking at the enzymes with higher IC/K_i values, one can see that some of the enzymes that have very high concentrations compared to their literature values (FBA1, PGCD) also have higher IC/K_i values, indicating that inhibition is a factor affecting fitting. Furthermore, when the enzymes have K_{eq}'s that are close to their mass action ratios, it appears that they also might have issues with a higher calculated concentration. Possible improvements made to this method could be made by explicitly accounting for inhibition by turning inhibition off during the fitting step.

Case study 1 was an attempt to both show a proof of concept of the method of overexpression and reconstruct a functioning strain. While modifications made to the TRN and to enzymes not included in the scope of this model were not included, changes were included where possible. One interesting aspect of the performance of the model compared to the experiment is the rapid overshoot and slow return to the new steady state. Compared to the productivities seen in the literature, the overshoot was higher, and much faster. This could be in part due to modeling overexpression and mutations as changes implemented at time zero. In

real cells, overexpression is not instantaneous, and time is required to increase the enzyme concentration of affected enzymes. Furthermore, overexpression of enzymes on plasmids causes metabolic load (Glick 1995), which would decrease the ability of the cell to produce tryptophan. Since the other sinks of tryptophan are accounted for in the sink reaction, this could be the reason why the paper's productivity values had a much shallower overshoot, and took much longer to reach the maximum value. While the agreement of tryptophan_export flux and the final productivity of the experiment seems to be promising, the boundary metabolite concentration was on the order of one millionth M. However, the increase in export is promising, and indicates the first principles of increasing enzyme concentrations leading to increased flux, and ergo increased concentration of tryptophan. However, intracellular concentrations of tryptophan did not increase to very high levels. One reason for this could be that the boundary reactions serve to artificially keep the fluxes of tryptophan being produced closer to the fluxes used to set up the model, possibly giving artificial resistance to perturbations. Finally, the agreement between the different ratios of the different models suggest that the performance of the model is consistent.

Case study 2 was studying the impact of athranilate inhibition on the strain simulated in the first case study. While the anthranilate inhibition did not appear to have a noticeable effect on the model dynamics, it is reasonable considering that the K_i for anthranilate is roughly one to two orders of magnitude higher than the steady state model initial condition. Furthermore, while the graphs were similar, the overall performance of the two strains seen in the dots below are also somewhat similar. Also, while the model does not replicate the final tryptophan productivity, the productivity data was from a batch reactor, featuring both a higher initial glucose concentration, and decrease in external glucose over time. Since the model assumes a constant external boundary metabolite concentration, it is reasonable that the final flux of both strains is dissimilar to the model output. When the model was used to recreate the interventions

of overexpressing IGPS in a fed batch reactor, the model produced a final tryptophan export flux similar to the experimental data. This could be due to the fed-batch reactor including replenishment of glucose, as well as a lower external concentration. The lack of accounting for metabolic load in this model from overexpressing the plasmid without IGPS on it, could explain the lack of a difference between the two graphs for the fed-batch reactor, as the metabolic load of overexpression could lead to depleted cellular resources, which would limit the ability to produce more tryptophan.

Case study 3 was studying the impact of a feed rate profile on tryptophan production. Several different studies have used different glucose and other metabolite feed rate profiles in order to optimize tryptophan production. While it might seem like the model does not respond to the feed rate profile. Looking at the sink reaction of the glucose into the system, one can see that the external concentration changes. The reason why the feed rate profile does show changes in the model, is likely due to the effect of the overexpression being more dominant. Furthermore, due to the structure of the model, it is possible that the model might be more resistant to the external perturbations.

As previously mentioned, there are several limitations to our approach in building this model. Uneven data quality and quantity can potentially lead to imprecise fits for microscopic rate constants, which can in turn affect the calculated concentration of the enzyme for a given reaction, which in turn alters the behavior of an enzyme in response to overexpression and other perturbations. Modeling mutations and overexpression without accounting for additional efficiency changes or metabolic load could be an issue, as (Chen et al. 2018) mentioned that the IGPS mutation that conferred anthranilate resistance also decreased activity. Furthemore, by not modeling the load of overexpression, the full cost of the increased protein production is not seen in the cell. In addition, simulating the concentration change at time zero likely also is not reflective of the comparatively more gradual increase in enzyme concentration, and the subsequent cellular responses to it. As previously mentioned, the high number of boundary

reactions also can lead to many normally closed pathways being more open than they would be in nature. Furthermore, due to the lack of kinetics scrapped for indole, information for cellular indole toxicity was not monitored, which could be a concern if indole reaches above 0.005M, at which point it becomes toxic to the cell (Wang, Ding, and Rather 2001). Finally, due to the mass action nature of this model, and its focus on metabolism, there is no account of the effects of transcriptional regulatory network changes, which were often an intervention done in strain design experiments.

Other labs have attempted to use systems biology in order to model the effect of strain design interventions on the tryptophan synthesis pathway. Rizk and Liao 2009 used an ensemble modeling approach to screen an ensemble of models paired with enzyme tuning data from overexpression data in order to obtain an ensemble of lin-log kinetic models glycolysis and the pentose phosphate pathway (Rizk and Liao 2009). The work flow employed by the authors is almost the reverse of the workflow employed in this study. While this study attempted to use kinetic information, metabolite information, and fluxes to construct a model to simulate overexpression and other interventions, (Rizk and Liao 2009) used overexpression data to build a model of DDPA flux. While their approach was able screen for models that qualitatively behave similarly to experimental findings, they are testing more to see if a combination of overexpressions can be predicted through their model. They do not simulate the steps of individual enzyme mechanisms. Furthermore, while their model can be screened to show that some of the models can predict overexpression effects, the amount of change, and the rate of change cannot be determined from their model. (Schmid et al. 2004) used a set of lumped reactions for tryptophan synthesis coupled with a model of the transcriptional regulatory network of tryptophan. Schmid was able to determine that upon optimizing the entire network for tryptophan production, tryptophan production increased to 260% percent of its prior value for a model of a tryptophan producing strain. This is similar to the increase of tryptophan export from the steady state value, to the value of about 1.3 $x10^{-6}$ M/s found from case study 1.

Furthermore, the model used lacks detailed inhibition and mechanistic information for each enzyme, and, while it can simulate a response to a glucose perturbation in real time, the model produced by this paper is able to simulate perturbations made to the expression and metabolic regulation of reactions directly. Other approaches to modeling tryptophan metabolism have also used experimental data to generate a model (Jing et al. 2018) fitting to a traditional yield based kinetic model. While the model is accurate, it is much more phenomenological and limited in depth. Furthermore, other lin-log kinetic models have been used to study the effects of perturbations on tryptophan metabolism (Maria, Mihalachi, and Gijiu 2018), such as by studying the effects of oscillations on the tryptophan and glycolytic pathways, finding that oscillations in the glycolytic pathway affected the tryptophan pathway, but not vice versa. This is in contrast to the results of workflow three, as the feed rate profile (a bigger perturbation than a pulse) was simulated, and the effect on tryptophan export was not discernible. One aspect that sets this model apart from past models of tryptophan synthesis is its reliance of models of enzyme mechanisms built from kinetic data. This first principles approach to modeling the flux through enzymes in the network is not commonly done due to difficulty in acquiring and curating data for these models. As more metabolomics and fluxomics information becomes available for *E. coli*, hopefully similar models to that proposed by this paper can gain yet further refinement.

In conclusion, this work has shown a workflow to construct a mass action kinetic model of tryptophan metabolism from a set of curated kinetic data and metabolic data. Furthermore, this work also shows preliminary results in using this model to replicate interventions done in three different experiments. While the dynamics of the models were different from that of the case studies, the similarity of the model results to experimental productivities provides impetus for further work on this type of model.

This thesis, in part is currently being prepared for submission for publication of the material.

Glass, Kevin; Zielinski, Daniel; Palsson, Bernhard. The thesis author was the primary

investigator and author of this material.

References

Bennett, Bryson D., Elizabeth H. Kimball, Melissa Gao, Robin Osterhout, Stephen J. Van Dien, and Joshua D. Rabinowitz. 2009. “Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in Escherichia Coli.” *Nature Chemical Biology* 5 (8): 593–99.

Boer, H. A. de, L. J. Comstock, and M. Vasser. 1983. “The Tac Promoter: A Functional Hybrid Derived from the Trp and Lac Promoters.” *Proceedings of the National Academy of Sciences of the United States of America* 80 (1): 21–25.

Carmona, S. B., F. M. Moreno, F. Bolívar, G. Gosset, and A. Escalante. 2015. “Inactivation of the PTS as a Strategy to Engineer the Production of Aromatic Metabolites in Escherichia Coli.” *Microbial Physiology* 25 (2-3): 195–208.

Chassagnole, Christophe, Naruemol Noisommit-Rizzi, Joachim W. Schmid, Klaus Mauch, and Matthias Reuss. 2002. “Dynamic Modeling of the Central Carbon Metabolism of Escherichia Coli.” *Biotechnology and Bioengineering* 79 (1): 53–73.

Chen, Lin, Minliang Chen, Chengwei Ma, and An-Ping Zeng. 2018. “Discovery of Feed-Forward Regulation in L-Tryptophan Biosynthesis and Its Use in Metabolic Engineering of E. Coli for Efficient Tryptophan Bioproduction.” *Metabolic Engineering* 47 (May): 434–44.

Chen, Lin, and An-Ping Zeng. 2017. “Rational Design and Metabolic Analysis of Escherichia Coli for Effective Production of L-Tryptophan at High Concentration.” *Applied Microbiology and Biotechnology* 101 (2): 559–68.

Davidi, Dan, Elad Noor, Wolfram Liebermeister, Arren Bar-Even, Avi Flamholz, Katja Tummler, Uri Barenholz, Miki Goldenfeld, Tomer Shlomi, and Ron Milo. 2016. “Global Characterization of in Vivo Enzyme Catalytic Rates and Their Correspondence to in Vitro Kcat Measurements.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (12): 3401–6.

D’Este, Martina, Merlin Alvarado-Morales, and Irini Angelidaki. 2018. “Amino Acids Production Focusing on Fermentation Technologies – A Review.” *Biotechnology Advances* 36 (1): 14–25.

Dodge, Timothy C., and Jeffrey M. Gerstner. 2002. “Optimization of the Glucose Feed Rate Profile for the Production of Tryptophan from Recombinant E Coli.” *Journal of Chemical Technology and Biotechnology* 77 (11): 1238–45.

Du, Bin, Zhen Zhang, Sharon Grubner, James T. Yurkovich, Bernhard O. Palsson, and Daniel C. Zielinski. 2018. “Temperature-Dependent Estimation of Gibbs Energies Using an Updated Group-Contribution Method.” *Biophysical Journal* 114 (11): 2691–2702.

Freedman, David, and Persi Diaconis. 1981. “On the Histogram as a Density estimator:L 2 Theory.” *Zeitschrift F�r Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 57 (4): 453–76.

Gerosa, Luca, Bart R. B. Haverkorn van Rijsewijk, Dimitris Christodoulou, Karl Kochanowski, Thomas S. B. Schmidt, Elad Noor, and Uwe Sauer. 2015. “Pseudo-Transition Analysis Identifies the Key Regulators of Dynamic Metabolic Adaptations from Steady-State Data.” *Cell Systems* 1 (4): 270–82.

Glick, Bernard R. 1995. “Metabolic Load and Heterologous Gene Expression.” *Biotechnology Advances*. https://doi.org/10.1016/0734-9750(95)00004-a.

Haiman, Zachary B., Daniel C. Zielinski, Yuko Koike, James T. Yurkovich, and Bernhard O. Palsson. 2021. “MASSpy: Building, Simulating, and Visualizing Dynamic Biological Models in Python Using Mass Action Kinetics.” *PLoS Computational Biology* 17 (1): e1008208.

Heckmann, David, Colton J. Lloyd, Nathan Mih, Yuanchi Ha, Daniel C. Zielinski, Zachary B. Haiman, Abdelmoneim Amer Desouki, Martin J. Lercher, and Bernhard O. Palsson. 2018. “Machine Learning Applied to Enzyme Turnover Numbers Reveals Protein Structural Correlates and Improves Metabolic Models.” *Nature Communications*. https://doi.org/10.1038/s41467-018-07652-6.

Henry, Christopher S., Linda J. Broadbelt, and Vassily Hatzimanikatis. 2007. “Thermodynamics-Based Metabolic Flux Analysis.” *Biophysical Journal* 92 (5): 1792–1805.

Ikeda, Masato. 2006. “Towards Bacterial Strains Overproducing L-Tryptophan and Other Aromatics by Metabolic Engineering.” *Applied Microbiology and Biotechnology* 69 (6): 615–26.

Jiang, Ming, and Haoran Zhang. 2016. “Engineering the Shikimate Pathway for Biosynthesis of Molecules with Pharmaceutical Activities in E. Coli.” *Current Opinion in Biotechnology* 42 (December): 1–6.

Jing, Keju, Yuanwei Tang, Chuanyi Yao, Ehecatl A. del Rio-Chanona, Xueping Ling, and Dongda Zhang. 2018. “Overproduction of L-Tryptophan via Simultaneous Feed of Glucose and Anthranilic Acid from Recombinant Escherichia Coli W3110: Kinetic Modeling and Process Scale-Up.” *Biotechnology and Bioengineering* 115 (2): 371–81.

Jol, Stefan J., Anne Kümmel, Vassily Hatzimanikatis, Daniel A. Beard, and Matthias Heinemann. 2010. “Thermodynamic Calculations for Biochemical Transport and Reaction Processes in Metabolic Networks.” *Biophysical Journal* 99 (10): 3139–44.

Lee, Jin-Ho, and Volker F. Wendisch. 2017. “Biotechnological Production of Aromatic Compounds of the Extended Shikimate Pathway from Renewable Biomass.” *Journal of Biotechnology* 257 (September): 211–21.

Liu, Lina, Muhammad Bilal, Hongzhen Luo, Yuping Zhao, and Hafiz M. N. Iqbal. 2019. “Metabolic Engineering and Fermentation Process Strategies for L-Tryptophan Production by Escherichia Coli.” *Processes* 7 (4): 213.

Liu, Lina, Xuguo Duan, and Jing Wu. 2016. “L-Tryptophan Production in Escherichia Coli Improved by Weakening the Pta-AckA Pathway.” *PloS One* 11 (6): e0158200.

Liu, Qian, Yongsong Cheng, Xixian Xie, Qingyang Xu, and Ning Chen. 2012. “Modification of Tryptophan Transport System and Its Impact on Production of L-Tryptophan in Escherichia Coli.” *Bioresource Technology* 114 (June): 549–54.

Lütke-Eversloh, Tina, and Gregory Stephanopoulos. 2008. “Combinatorial Pathway Analysis for Improved L-Tyrosine Production in Escherichia Coli: Identification of Enzymatic Bottlenecks by Systematic Gene Overexpression.” *Metabolic Engineering* 10 (2): 69–77.

Maria, Gheorghe, Marina Mihalachi, and Cristiana Luminita Gijiu. 2018. “In Silico Optimization of a Bioreactor with an E. Coli Culture for Tryptophan Production by Using a Structured Model Coupling the Oscillating Glycolysis and Tryptophan Synthesis.” *Chemical Engineering Research and Design* 135 (July): 207–21.

Masuo, Shunsuke, Shengmin Zhou, Tatsuo Kaneko, and Naoki Takaya. 2016. “Bacterial Fermentation Platform for Producing Artificial Aromatic Amines.” *Scientific Reports* 6 (1): 25764.

McCloskey, Douglas, Sibei Xu, Troy E. Sandberg, Elizabeth Brunk, Ying Hefner, Richard Szubin, Adam M. Feist, and Bernhard O. Palsson. 2018. “Adaptive Laboratory Evolution Resolves Energy Depletion to Maintain High Aromatic Metabolite Phenotypes in Escherichia Coli Strains Lacking the Phosphotransferase System.” *Metabolic Engineering* 48 (July): 233–42.

Rizk, Matthew L., and James C. Liao. 2009. “Ensemble Modeling for Aromatic Production in Escherichia Coli.” *PloS One* 4 (9): e6903.

Rodriguez, Alberto, Juan A. Martnez, Noem Flores, Adelfo Escalante, Guillermo Gosset, and Francisco Bolivar. 2014. “Engineering Escherichia Coli to Overproduce Aromatic Amino Acids and Derived Compounds.” *Microbial Cell Factories* 13 (1): 126.

Rohatgi, Ankit. 2017. “WebPlotDigitizer.” Austin, Texas, USA.

Santos, Christine Nicole S., Wenhai Xiao, and Gregory Stephanopoulos. 2012. “Rational, Combinatorial, and Genomic Approaches for Engineering L-Tyrosine Production in Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (34): 13538–43.

Schmid, Joachim W., Klaus Mauch, Matthias Reuss, Ernst D. Gilles, and Andreas Kremling. 2004. “Metabolic Design Based on a Coupled Gene Expression—metabolic Network Model of Tryptophan Production in Escherichia Coli.” *Metabolic Engineering* 6 (4): 364–77.

Schmidt, Alexander, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, Benjamin Volkmer, Luciano Callipo, Kèvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. 2016. “The Quantitative and Condition-Dependent Escherichia Coli Proteome.” *Nature*

Biotechnology 34 (1): 104–10.

Taymaz-Nikerel, Hilal, Marjan de Mey, Cor Ras, Angela ten Pierick, Reza M. Seifar, Jan C. van Dam, Joseph J. Heijnen, and Walter M. van Gulik. 2009. “Development and Application of a Differential Method for Reliable Metabolome Analysis in Escherichia Coli.” *Analytical Biochemistry* 386 (1): 9–19.

Wang, D., X. Ding, and P. N. Rather. 2001. “Indole Can Act as an Extracellular Signal in Escherichia Coli.” *Journal of Bacteriology* 183 (14): 4210–16.