A DATA-DRIVEN MODELING PIPELINE FOR QUANTITATIVE SYSTEMS PHARMACOLOGY

By

CHRISTOPHER DENARO

A dissertation submitted to the

Graduate School–Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Benedetto Piccoli

And approved by

Benedetto Piccoli

Iman Dehzangi

Jeffrey S. Barrett

Maria E. Solesio

Camden, New Jersey

May 2025

ABSTRACT OF THE DISSERTATION

A Data-driven Modeling Pipeline for Quantitative Systems Pharmacology

by CHRISTOPHER DENARO

Dissertation Director:
Benedetto Piccoli

Computational methods are becoming commonly used in many areas of medical research. Recently, the field of Quantitative Systems Pharmacology (QSP) emerged to address the need for mathematical models to capture the complex interplay between therapeutic approaches and underlying disease pathophysiology. In a sense, QSP modeling incorporates approaches from Physiologically based pharmacokinetics (PBPK), systems biology, and disease modeling, with the explicit goal of model design and output being motivated by quantitative data. In this dissertation, we propose a computational framework utilizing a generalization of mathematical graphs called hypergraphs to model metabolic and disease networks; we apply this framework directly to the combination therapy treatment of Tuberculosis, combining heterogeneous data sources of gene expression levels via microarray data and metabolic networks. Next, we demonstrate that machine learning architectures that have classically been applied to natural language problems can be trained to learn network simulation equilibria, providing a proof-of-concept of traditional machine learning approaches to a biological setting. We transition into a literature review that examines the connection between infection with SARS-CoV-2 and neurodegeneration through the lens of mitochondrial function. Next, we motivate and explore applications of the computational framework used in treatment optimization of Tuberculosis to therapeutic biomarker identification in genetic variants of Parkinson's Disease (PD). We also review publicly available data sources for genetic variants of PD and propose a collaborative and integrative computational platform for future model development.

Acknowledgments

First and foremost, I wish to extend my gratitude to my committee for their guidance and expertise and, in particular, Dr. Benedetto Piccoli for shaping my development as a scientist. I greatly appreciate the opportunities and collaborations provided by the joint work with the Critical Path Institute: Critical Path for Parkinson's. I want to express my gratitude to the numerous members of Dr. Piccoli's lab for their invaluable help and support in my research projects. I especially want to thank Ryan Weightman for his continuous willingness to to help track down and squash bugs in my code, conversation to understand new research topics, and help navigating the bureaucracy of graduate school. Finally, I would like to thank my friends and family, especially my Mom, Dad, and brother for providing a robust support network throughout my higher education. To Alexa, I could not have undertaken this journey without your love, reassurance, and support, you have undoubtedly assisted me through the completion of my graduate program.

Contents

List of Figures viii

List of Tables x

List of Abbreviations xi

1 General Introduction 1

1.1 Background on Mathematical Modeling 1

1.2 Introduction to Applied Mathematical Modeling 4

1.3 Overview of Chapters 11

2 A pipeline for testing drug mechanism of action and combination therapies: From microarray data to simulations via Linear-In-Flux-Expressions: Testing four-drug combinations for tuberculosis treatment 13

2.1 Chapter Overview 14

2.2 Introduction 14

2.3 Pipeline Overview 17

2.3.1 Genetic Data 17

2.3.2 Data Curation 18

2.3.3 Metabolic Pathways 18

2.3.4 Drug E ects on Relevant Fluxes 19

2.3.5 Optimal Exchange Fluxes 19

2.3.6 Network Simulation 20

2.3.7 LIFE Methodology for Network Analysis 21

2.4 Pipeline in detail 22

2.4.1 Genetic Data 22

2.4.2 Data Curation 22

2.4.3 Metabolic Pathways 24
2.4.4 Drug E ects on Relevant Fluxes 25
2.4.5 Optimal Exchange Fluxes 30
2.4.6 PK Model 31
2.4.7 Network Simulation 33
2.4.8 LIFE Methodology for Network Analysis 33
2.5 Application to tuberculosis 36
2.5.1 TB: Genetic Data 37
2.5.2 TB: Data Curation 37
2.5.3 TB: Metabolic Pathways 38
2.5.4 TB: Optimizing Exchange Fluxes 41
2.5.5 TB: Drug E ects on Relevant Fluxes 42
2.5.6 TB: PK Model 42
2.5.7 TB: Network Simulation 43
2.5.8 TB: LIFE Methodology for Network Analysis 44
2.6 Conclusion 45

3 A deep language model to predict metabolic network equilibria 48
3.1 Chapter Overview 49
3.2 Introduction 49
3.3 Application of Transformer Architecture to Metabolic Networks . . 52
3.3.1 Metabolic networks and their equilibria 52
3.3.2 Approaches in modeling metabolic networks 55
3.3.3 Applicability of model to hypergraph networks 59
3.4 Deep learning architecture and network tokenization 61
3.4.1 Using deep learning models to predict equilibria - dataset generation 63
3.4.2 Predicting equilibria 66
3.4.3 Out-of-distribution generalization and results on real metabolic networks 67

3.5 Discussion 71
3.6 Datasets 75
3.7 Detailed results 78
3.7.1 Qualitative problem 78
3.7.2 Quantitative problems 78
3.8 The 29 metabolic networks used for testing 81
3.9 Conclusion 83

4 COVID-19 and neurodegeneration: The mitochondrial connection 85
4.1 Paper Overview 86
4.2 COVID-19 and neurodegeneration 86

5 Advancing precision medicine therapeutics for Parkinson's utilizing a shared quantitative systems pharmacology model and framework 96
5.1 Paper Overview 97
5.2 Introduction: Key scienti c and regulatory advances and enablers, development challenges and opportunities for precision medicine development in PD 98
5.3 Review of the genetic bases and associated pathways of PD and implications for biomarker development and disease strati cation . . 103
5.4 Role and impact of QSP in drug development and opportunity to advance PD therapeutic candidates through clinical proof of concept 105
5.5 Overview of existing initiatives and associated databases towards enabling e ective and e cient clinical path to proof of concept . . . 109
5.5.1 Sources of data for PD 110
5.6 Advances in regulatory landscape and utility of modeling and simulation for addressing and accelerating drug development regulatory milestones 113

5.6.1 Call to action for advancing precompetitive QSP models to support precision medicine strategies 115

6 Semi-mechanistic Modeling Tools with Applications to Biomarker Identi cation 118

6.1 A Quantitative System Pharmacology Framework for Parkinson's Disease 119

6.2 Integrating Multiple Data Types to Inform Disease Networks 124

7 Crowdsourcing proposal supporting patient engagement in Parkinson's Disease: A DRE-enabled, patient swarm approach to develop quantitative systems pharmacology (QSP) models 129

7.1 Paper Overview 130

7.2 Introduction 130

7.3 Representing Patients and Mobilizing the Patient Swarm 132

7.3.1 Leveraging existing technologies and data 135

7.3.2 Diversity in planned and completed clinical trials 136

7.3.3 Prototype Data Environment and the DRE future home . . 138

7.3.4 Engaging the swarm 140

7.3.5 Real-world patient level data 140

7.4 QSP Modeling Approach 142

7.4.1 Patient Engagement and Model Re nement with the Voice of the Patient 145

7.5 Current progress 147

7.5.1 QSP Modeling Progress - Rutgers Group 149

7.5.2 QSP Modeling Progress - University of Florida Group 151

7.6 Discussion - Delivering for Patients 152

List of Figures

1.1 Example Lotka-Volterra Model 6
1.2 Example SIR Model 7
1.3 Compartments in SIR Model 8
1.4 Example Hypergraph 10

2.1 Computational Pipeline Overview 17
2.2 Implementation of Drug E ects 25
2.3 Implementation of PK Data 32
2.4 Hyperedge vs Simple edge Representations 33
2.5 Simulations of Hyper- and Simple edge Systems 34
2.6 Equilibrium Analysis Plots 36
2.7 Average Value of Signi cant Genes Under Treatment 38
2.8 Mapping of Signi cant Genes to Pathways 40
2.9 L1 Ranking of Drug Combinations 41
2.10 Annotated KEGG Pathway 43
2.11 Metabolic Network Simulation Snapshot 46
2.12 Combination Therapy Metabolite Trajectories 47

3.1 Example Networks with and without Equilibria 53
3.2 Di erent Types of Biological Networks 54
3.3 Example of a Weighted Hypergraph 60
3.4 Example Corresponding Simple Graph 60
3.5 Encoding Procedure for Networks 62
3.6 Transformer Model Architecture 64
3.7 Synthetic Dataset Generation Scheme 65
3.8 Model Learning Curves 80

4.1 COVID-19 and Neurodegeneration Mitochondrial Commonalities . . 90

5.1 QSP Modeling and Drug Development Decision Making 106
5.2 Application of QSP Models Throughout Clinical Stages 107

6.1 QSP Model Framework 120
6.2 Example Compartmental Model 121
6.3 Low-level Disease Network 122
6.4 Annotated Low-level Disease Network 123
6.5 Disease Network Simulation Node Trajectories 123
6.6 Protein-Protein Interaction Network Schematic 125

7.1 DRE Diagram 139
7.2 DRE Integration with Academic Institutions Schematic 142
7.3 Generalized PD Model Development 144

List of Tables

2.1 Drug Dosage Information 37
2.2 Summary of Signi cant Genes 39

3.1 Qualitative Model Performance 66
3.2 Quantitative Model Performance 67
3.3 Out of Distribution Model Performance 68
3.4 Out of Domain Qualitative Model Performance 70
3.5 Out of Domain Quantitative Model Performance 71
3.6 Out of Domain Generalization for Weighted Experiments 72
3.7 Training Set Parameters 77
3.8 Qualitative Performance, Best vs. Low Dimension 78
3.9 Quantitative Performance at Varying Tolerances 79
3.10 Quantitative Accuracy Across Encoder Depths 80
3.11 Quantitative Accuracy Across Tolerances 81

5.1 Parkinson's Disease Data banks 109

6.1 PPI Network Protein and Disease Network Overlap 127
6.2 PPI Network Compartments and Disease Network Overlap 128

7.1 Public Parkinson's Disease Data Banks 137
7.2 Parkinson's Disease Clinical Trials 150

List of Abbreviations

A amyloid

AD Alzheimer's Disease

ADME absorption distribution metabolism and excretion

AI arti cial intelligence

AMP PD Accelerated Medicines Partnership for PD

APP amyloid precursor protein

BBB blood-brain barrier

CNS central nervous system

CPP Critical Path for Parkinson's

CSF cerebrospinal uid

DaTSCAN dopamine transporter scan

DLB Dementia with Lewy Bodies

DMR data management resource

DMT disease-modifying therapy

DRE digital research environment

DTI di usino tensor imaging

EMA European Medicines Agency

ETC Electron Transfer Chain

FBA Flux Balance Analysis

FDA Federal Drug Administration

FDR false discovery rate

FTD frontotemporal dementia

GCase glucocerebrosidase

GEO Gene Expression Omnibus

GP2 Global Parkinson's Genetic Program

GPU Graphics Processing Unit

GRN gene regulatory network

GSEA Gene Set Enrichment Analysis

GWAS genome-wide association studies

HIF-1 hypoxia-inducible factor 1

HPC High Performance Computing

IFN- interferon

IND Ivestigational New Drug

IPA Ingenuity Pathway Analysis

IR information retrieval

KEGG Kyoto Encyclopedia of Genes and Genomes

LIFE Linear-In-Flux-Expressions

LM language model

M membrane

M&S Modeling & Simulation

MAVS mitochondrial antiviral-signaling protein

MIDD model informed drug development

ML machine learning

MoA mechanism of action

MoD mechanism of disease

MRI magnetic response imaging

MT-CYB Mitochondrial Cytochrome B

mtDNA mitochondrial DNA

N nucleocapsid

NCE new chemical entities

NDS Neuronal Synuclein Disease

NFL neuro lament light

NLP natural language processing

nOH neurogenic orthostatic hypotension

Nsps non-structural proteins

ORF open reading frame

ORF1ab open reading frame 1ab

OXPHOS oxidative phosphorylation

PBPK Physiologically based pharmacokinetics

PCR polymerase chain reaction

PD-PROP Parkinson's Disease Patient Report of Problems

PDBP Parkinson's Disease Biomarker Program

PK pharmacokinetic

POC proof-of-concept

PPI protein-protein interaction

PPMI Parkinson's Progression Marker Initiative

PPP Public-Private Partnership

QSP Quantitative Systems Pharmacology

RET reverse electron transport

ROS reactive oxygen species

RyR2 ryanodine receptor 2

S spike

SAA seed ampli cation assay

SAM Signi cance Analysis of Microarrays

SARS-CoV-2 severe acrute respiratory syndrome coronavirus 2

SMT symptom-modifying therapy

TRE trusted research environment

UPDRS Uni ed Parkinson's Disease Rating Scale

WPPIN weighted protein-protein interaction network

Chapter 1

General Introduction

1.1 Background on Mathematical Modeling

The challenges faced during drug development are numerous and varied, spanning a wide range of elds including (but not limited to) biochemistry, molecular biology, systems biology, arti cial intelligence, regulatory practices, and computational costs. Within the past decade, there has been an expansion and adoption of a eld known as QSP, a eld characterized by the study of the dynamic interaction of a therapeutic agent with traditional systems biology frameworks. QSP expands on traditional Systems Biology approaches by focusing on mathematical models to characterize systems with an implicit focus on the availability of data sources across di erent scales. For instance, a QSP approach may include a mathematical model for the underlying metabolic network and incorporate pharmacokinetic (PK), pharmacodynamics, and patient outcome data. While the challenges faced in drug development are varied, we will introduce and focus on only a select few to motivate the applications within this dissertation. One such challenge in drug development arises from the reuse and reapplication of clinically tested drugs to new diseases; when there is an abundant number of available therapies with low toxicity and the ability to be combined, a particular challenge arises in nding the optimal treatment combination. More speci cally, given a set of possible drugs to form a combination therapy, the question of how to choose the optimal therapy becomes both physically and computationally expensive as the number of drugs increases. This issue arises due to the fundamental concept in combinatorics of choosing without replacement; for example, if there are 25 possible drugs to choose from, picking a four-drug combination results in over 10,000 possible combination

therapies. As we will discuss in Chapter 2, nding the best combination therapy for the treatment of Tuberculosis is an ongoing eld of study. An altogether different challenge arises when there is a limited understanding of the mechanism of disease (MoD) or even the mechanism of action (MoA) of potential therapeutic interventions. In this situation, there is a problem of evaluating the e cacy of a trial therapy. A common approach is to identify biomarkers for the disease and the therapy; biomarkers are measurable characteristics (e.g., presence of a metabolite in a bio uid) that indicate (or support) speci c health outcomes or indicate outcomes of biological processes. We will propose a QSP model for genetic variants of Parkinson's Disease (PD) as a potential avenue to identify biomarkers associated with disease progression. Parkinson's Disease o ers a challenging application as PD treatment is one such area that su ers from a lack of disease-modifying therapies, relying on existing therapies that simply address symptoms.

The responses to these distinct di culties are understandably varied; we will focus on computational and mathematical methods of addressing these challenges in order to frame the discussions within this dissertation. Systems biology is a eld that aims to understand how biological systems interact { examining a system across many di erent scales, from micro- to macro-perspectives. For instance, in studying combination therapies as an approach to the treatment of Tuberculosis, systems can range from metabolic networks with Mycobacterium Tuberculosis all the way to the host system in areas of infection.

Mathematical modeling can de-risk traditional approaches in pharmaceutical development. Mathematical models can be broadly split into two di erent classes: mechanistic and non-mechanistic. Mechanistic models aim to capture the underlying biology/mechanism of disease, whereas non-mechanistic models aim to characterize the behavior of a biological system based on statistical analysis of outcomes.

QSP o ers a framework for mechanistic models; for instance, a model of drug concentration in blood may be implemented as a system of di erential equations

describing the concentration of blood in various compartments within the body. Additional mechanistic models may use as input the predicted drug concentration in a speci c compartment to predict the e ect that a drug may have on disease outcome. This approach can be supported by the non-mechanistic modeling of longitudinal markers in clinical trial outcomes.

Mathematical models must strike a balance between complexity and applicability; the number of parameters within a model correlates with the amount of data needed in order to t those model parameters. As one might expect, models with a large number of parameters may be able to more accurately describe a biological system, but the accuracy may be meaningless if there is not su cient data to nd the correct values of the parameters. When the available data in a subject eld is sparse, compromises on the detail of the model must be made.

There are two famous quotes in the mathematical modeling eld that characterize the relationship between model complexity and data availability; the rst is attributed to John von Neumann, \With four parameters, I can t an elephant, and with ve, I can make him wiggle his trunk." If given a biological system, we can keep adding parameters to our model to account for each detail of the system; of course, with each new parameter, we require additional data in order to t the model parameters. Additionally, there is the statistician George Box wrote, \All models are wrong, some are useful." We never aim to capture completely the underlying behavior of the system, just the required amount in order to apply the mathematical model to the problem.

Because of the reliance on available data to t model parameters, mathematical models are often created with prior knowledge about what data is available in the problem domain. In this dissertation, we introduce a model for combination therapies for the treatment of Tuberculosis { this work would look completely different if there were not already publicly available data about single-drug response in Mycobacterium Tuberculosis.

1.2 Introduction to Applied Mathematical Modeling

In order to reduce the barrier to entry for this dissertation, we provide a section that introduces mathematical concepts key to modeling in the areas discussed within. We will motivate and introduce basic differential equations, compartmental models, and mathematical graphs as tools to model various scales of biological systems.

In broad terms, differential equations are equations that relate functions and their derivatives. Recalling that the derivative of a function describes how that function changes (with respect to its parameters), differential equations describe how quantities change.

A first or second course in Calculus will generally introduce the "Fundamental Theorem of Calculus" which (broadly speaking) describes integration as an inverse process of taking the derivative. More precisely, Theorem 1.2.1 describes the relationship between antiderivatives and derivatives.

Theorem 1.2.1. Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined, for all $x \in [a, b]$, by $F(x) = \int_a^x f(t)\,dt$. Then F is uniformly continuous on $[a, b]$ and differentiable on the open interval (a, b), and $F'(x) = f(x)$ for all $x \in (a, b)$.

When we use differential equations as modeling tools, we say the differential equation is f from Theorem 1.2.1, and we look for its antiderivative, F.

The term "simulation" within the context of differential equations typically refers to the process of: (1) defining initial conditions for the system – meaning what value the antiderivative starts at, (2) establishing a "time-horizon" that defines how long the simulation will last – meaning what interval do we want to integrate over, and (3) numerically integrating the differential equations. There are many techniques for numerical integration, and the exact implementation of the technique is outside of the scope for the topics discussed in this dissertation.

A classic application of di erential equations in biology is the Lotka-Volterra model (typically shown in introductions to modeling in biology), which models the relationship between predator and prey animals within an ecosystem.

$$\frac{dx}{dt} = x \quad xy \tag{1.1}$$

$$\frac{dy}{dt} = \quad y + xy \tag{1.2}$$

In this setting, x and y are functions of time that correspond to the population levels of the prey and predator, respectively. The parameters of the model, , ; ; , correspond to the growth rate and death rate of the prey and predator species, respectively. We note that these equations describe the derivatives of x and y with respect to t, the time. Additionally, as described earlier in the introduction, the parameters of the model must be t using some observed data in order for this model's predictions to be applicable.

The Lotka-Volterra equations for predator-prey interactions within an ecosystem can be used to predict how the population of the predator species may change if there is a sudden gain/loss of the prey species (and vice versa). An example simulation of the Lotka-Volterra model is shown in Figure 1.1.

A perhaps even more well-known basic application of di erential equations in biology following the outbreak of COVID-19 in 2020 is the so-called SIR model, which describes the dynamics of \Susceptible", \Infectious", and \Removed" populations during an outbreak. In a SIR model, we describe how the population of S, I, and R change with respect to time { we de ne $\frac{dS}{dt}; \frac{dI}{dt}; \frac{dR}{dt}$.

$$\frac{dS}{dt} = \quad \frac{}{N} IS \tag{1.3}$$

$$\frac{dI}{dt} = \frac{}{N} IS \quad I \tag{1.4}$$

$$\frac{dR}{dt} = I \tag{1.5}$$

Figure 1.1: An example simulation of the Lotka-Volterra equations with initial conditions $x(0); y(0) = 4; 2$ and parameter values of $; ; ; = 1$. The time horizon of this simulation is $t \ 2 \ [0; 30]$.

In this setting, the meaning of the variables within the di erential equations is the following: $S(t); I(t);$ and $R(t)$ are functions that describe the number of people within the susceptible, infectious, and removed categories, respectively. The parameters of the model, $;$, correspond to the infection rate and recovery rate, respectively. An example simulation of Equation 1.3 is given in Figure 1.2, note the equilibrium (all trajectories are \ at") that occurs beginning around the time $t = 6$.

As one might notice, the traditional SIR model does not account for a plethora of real-world conditions such as birth and death rates (the total population size does not change during the simulation); re-infection with disease; or di erent variants of the disease { this hearkens back the famous quote from George Box and can be explained by saying that the majority of the infections during an epidemic can be described by these three simple equations. Of course, this does not mean that more complicated expansions of the SIR model do not exist, as there has been much work in this direction [341].

The simplest de nition of a mathematical graph (also called a network) is provided in De nition 1.2.1, but generally speaking, a graph contains nodes (also

Figure 1.2: An example simulation of the SIR equations with initial conditions $S(0); I(0); R(0) = (999; 1; 0)$ and parameter values of $; = 1:1; 1$. The time horizon for this simulation is $t 2 [0; 30]$.

called vertices) and edges. The relationship between nodes is described by the edges.

De nition 1.2.1 (Graph). A graph is an ordered pair, $G = (V; E)$, where V is a set of vertices and E is a set of edges, $E ff x; ygjx; y 2 V; x 6 yg$.

Note that in the most basic de nition of a graph, De nition 1.2.1, the edges have no direction; they are an unordered set of nodes (of size two). The concept of a graph can be expanded into a directed graph by introducing direction to the edges. An example of a directed graph may be a graph of a website, where distinct pages on the domain are nodes, and edges correspond to the ability to navigate from one page to another. In this dissertation, we will continue to expand on the idea of a graph, introducing various de nitions that aim to more accurately describe the relationship in biological networks.

Graphs can be applied to various biological systems of many di erent scales. A practical example of a graph is a food web, where nodes within the graph are species of animals in an ecosystem, and the edges within the graph correspond to the predator-prey relationship between species within the ecosystem. In this dissertation, we primarily focus on graphs generated by the relationships within

chemical reaction networks, metabolic networks, disease networks, or gene regulatory networks. In these networks, nodes often refer to chemical compounds, metabolites, or genes, where edges correspond to chemical reactions or gene regulations.

We can combine the concepts of di erential equations and graphs in the following manner: we begin by encoding a process as a graph, then we can use the structure of that graph to motivate di erential equations. To illustrate this process, we use the simplistic example of a basic SIR model. Figure 1.3 demonstrates the relationship between \Susceptible", \Infectious", and \Removed" populations during a pandemic. Stated simply, susceptible populations can become infectious, and after a certain period of the disease, the infectious population can become removed (e.g., develop immunity). We see these relationships encoded in the graph of Figure 1.3 { edges denote the possibility of one population owing to another population; the direction of the edges tells us the direction of the ow.

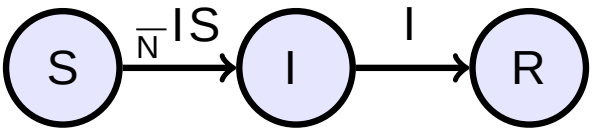

Figure 1.3: A graphical representation of the compartments from a SIR model. Each node (blue circle) corresponds to a compartment within the model. Compartments (left to right) are \Susceptible", \Infectious", and \Removed". The edges (black arrows) denote the relationship between compartments. Population ows in the direction of the arrow { susceptible people can become infectious, and infectious people can be removed. Edges are labeled with the corresponding terms from Equations 1.3.

The graph from Figure 1.3 can be used to inform the dynamics of a system of equations. Recalling Equations 1.3, we can attribute terms in the equations to each edge within the graph, where the sign of the term is dependent on the direction of the edge { negative terms correlate to edges leaving a node, where positive terms correlate to edges entering nodes.

A similar approach as the one outlined above can be used for other biological networks (such as metabolic networks). Nodes within these graphs correspond

to metabolites, and edges within the graphs correspond to biological processes that take one metabolite to another. We typically impose dynamics that relate to the amount of the metabolites (similar to how the dynamics in 1.3 relate to the number of people). An important concept throughout is the further generalization of a mathematical graph { we recount the de nition of a hypergraph, which allows edges to connect more than two nodes; this generalization is particularly important to capture the behavior of chemical reactions that include more than one substrate or product. De nition 1.2.2 provides a more formal de nition of an ordered hyper graph; while De nitions 1.2.1 and 1.2.2 may look similar, De nition 1.2.2 denotes a hyperedge as an ordered pair of sets, not single elements.

De nition 1.2.2 (Hyper Graph). A hyper graph is an ordered pair, $G = (V; H)$, where V is a set of vertices (or nodes) and H is a set of hyperedges, H $f(X; Y)jX; Y \quad Vg$.

Figure 1.4 provides a basic example of a hypergraph from De nition 1.2.2 consisting of two directed hyperedges and ve nodes. The nodes of the (hyper) graph are represented by blue circles, $v_0; : : : ; v_4$, and the hyperedges are represented by white circles, $h_1; h_2$. Using the notation from the De nition 1.2.2, $h_1 = (f v_0 g; f v_1; v_2 g)$ and $h_2 = (f v_1; v_2 g; f v_3; v_4 g)$; notice that the entries within the hyperedge are sets as opposed to single vertices. In Chapter 2, we will further expand on the concept of hypergraphs to so-called metabolic ubergraphs; however, the concept is similar. Furthermore, when using hypergraphs to represent metabolic networks, hyperedge h_1 may correspond to a dissociation chemical reaction, whereas hyperedge h_2 may correspond to a chemical reaction with two reactants and two products.

When studying biological systems through the lens of mathematical modeling, we often care about the \steady-state" behavior. The \steady-state" behavior of a system of di erential equations is formally referred to as the equilibria of the system { it is important to note that the equilibria of the system are a function of the parameters and initial conditions of the system. An informal description

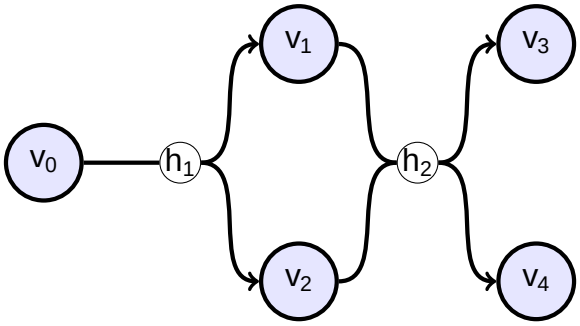

Figure 1.4: This gure depicts a small network consisting of ve nodes $v_0; : : : ; v_4$, and two directed hyperedges, h_1 and h_2. Nodes are depicted by blue circles, and hyperedges are labeled using small white circles. Arrows indicate the direction of the hyperedges.

of the equilibria of a system is, \What is the end state of the simulation when the variables have stopped changing?" For example, we see that the simulation in Figure 1.2 has reached an equilibrium state because the values of S, I, and R are no longer changing (the lines are at). We typically run simulations long enough to capture the steady-state behavior of the system. A motivating example of why we care about the equilibria of a system is the following: consider a mathematical model of a metabolic network within a bacterium when exposed to an antibiotic. If we wish to compare two di erent antibiotics, we can compare the di erence in equilibria achieved in the simulation under each antibiotic. The ow of the \experiment" is the following:

1. Simulate the metabolic network without any antibiotic
2. Record the equilibrium when the bacterium is exposed to no antibiotic (ie, the values of the metabolites)
3. Simulate the network with the rst antibiotic and record the new equilibrium
4. Simulate the network with the second antibiotic and record the new equilibrium
5. Compare the de ection of the metabolites between: (1) the untreated system, (2) the rst antibiotic, (3) the second antibiotic

In this example, comparing the de ection in the equilibria is just one method of comparing two di erent antibiotics; we could also justi ably look at the maximum rate of de ection of metabolites. As one might expect, mathematical modeling gives us many ways to compare di erent treatments on biological systems, motivating the development of metrics for drug e cacy, cost functions for treatment optimization, and summary statistics for large biological systems. Later, we motivate and introduce metrics for estimating therapeutic e cacy in di erent systems.

1.3 Overview of Chapters

Chapter 2 proposes a framework for simulating metabolic networks; discusses rigorous mathematical foundations for a generalized form of a hypergraph, and applies the framework to an example domain { namely, treatment optimization for multi-combination therapies of Tuberculosis. This chapter begins by introducing a computational pipeline that combines metabolic network information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) with microarray gene expression data from the Gene Expression Omnibus (GEO). The microarray data provides information on how a single treatment a ects gene expression within Mycobacterium Tuberculosis; we combine multiple microarray datasets to represent the e ect of combination therapies. Additionally, we identify areas of the metabolic network associated with the genes that are di erentially expressed to model the e ect of combination therapies on a network scale. Chapter 3 demonstrates a novel application of a transformer architecture to biological networks, providing insight into the fact that a language model may be able to learn complex network structures and applies this approach to publicly available metabolic networks from the Kyoto Encyclopedia of Genes and Genomes (KEGG). This chapter outlines the generation of small, medium, and large networks to act as a training dataset for the transformer architecture. In this project, we are interested in predicting whether networks exhibit equilibria when simulated according to standard di er-

ential equations (qualitative) and if a network does exhibit an equilibrium, what is its value (quantitative). We then evaluated the performance of the qualitative and quantitative learning on the di erent network sizes and evaluated the model on metabolic networks from KEGG. Chapter 4 provides a review of the early evidence for the link between neurodegeneration and severe acrute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Throughout this chapter, we provide a review of the literature (prior to October 2022) that proposed a link between COVID-19 and mitochondrial dysfunction; as such, the review addresses the neurodegenerative implications of COVID-19 through the lens of mitochondrial function. Chapter 5 provides a review of the common biological processes implicated by GBA1 and LRRK2 variants of Parkinson's Disease (PD), provides a review of previous applications of mechanistic models within the clinical trial pipeline, describes the available data sources for a biomarker-focused model of PD, and motivates the need for a mechanistic model of PD. This chapter also outlines the general areas of pre-clinical and clinical stages of drug development while simultaneously summarizing applications of QSP models to pre-clinical and clinical drug development. Chapter 6 demonstrates the individual components of a QSP model for genetic variants of PD with the model focusing on therapeutic biomarker identi cation as motivated in Chapter 5 and 7; this chapter also introduces tools to support a QSP model when there is a lack data availability. Chapter 7 discusses how involvement of PD patient populations within traditional model development pipelines (across various computational platforms) may lead to better patient outcomes and research, this chapter also proposes a collaborative framework for general model development across several institutions.

Chapter 2

A pipeline for testing drug mechanism of action and combination therapies: From microarray data to simulations via Linear-In-Flux-Expressions: Testing four-drug combinations for tuberculosis treatment

This chapter was originally published online on May 19, 2023, and is published in nal edited form as [77].

2.1 Chapter Overview

Computational methods are becoming commonly used in many areas of medical research. Recently, the modeling of biological mechanisms associated with disease pathophysiology have bene ted from approaches such as Quantitative Systems Pharmacology (brie y QSP) and Physiologically Based Pharmacokinetics (brie y PBPK). These methodologies show the potential to enhance, if not substitute animal models. The main reasons for this success are the high accuracy and low cost. Solid mathematical foundations of such methods, such as compartmental systems and ux balance analysis, provide a good base on which to build computational tools. However, there are many choices to be made in model design, that will have a large impact on how these methods perform as we scale up the network or perturb the system to uncover the mechanisms of action of new compounds or therapy combinations. A computational pipeline is presented here that starts with available-omic data and utilizes advanced mathematical simulations to inform the modeling of a biochemical system. Speci c attention is devoted to creating a modular work ow, including the mathematical rigorous tools to represent complex chemical reactions, and modeling drug action in terms of its impact on multiple pathways. An application to optimizing combination therapy for Tuberculosis shows the potential of the approach.

2.2 Introduction

The eld of mechanistic modeling lies at the intersection of mathematical and biological sciences [15, 38, 177, 310]. The aim of mechanistic models is to capture and simulate underlying biological systems which are relevant to an applied problem. This eld has been expanding as the availability of computational power and mechanistic data becomes more prevalent [44, 20, 151, 66, 219]. There has also been an increase in big data in the systems biology eld, which makes establishing model priors more readily available. Note that mechanistic models contrast arti cial in-

telligence models as they are typically built with reference to known biological systems - contrasting the \black box" behavior of arti cial intelligence models. In this paper, we will introduce a pipeline that was created with the intention to investigate drug mechanism of action for single and combination therapies.

There exist many approaches to mechanistic modeling, and this area is rich and shows promises of expanding [252, 171]. There has been a call for this type of modeling as it o ers a cheaper alternative to running translational and pre-clinical studies. Moreover, the accessibility of large genomic data sets makes populating necessary model priors more feasible.

Prior research in this eld includes several methods of modeling - some more easily tied to underlying biology than others. Graph theory approaches, ordinary di erential equation approaches, boolean network modeling, and agent-based modeling all serve to o er a foundation for mechanistic models [38]. In this paper, we will discuss a pipeline that uses a variation of the typical Flux-Balance-Analysis and ordinary di erential equations approach.

As a motivation for the necessity of this pipeline, there is an ever-present interest in model-informed drug development (MIDD) [252, 171]. Traditional methods of drug development are costly and time-consuming and do not translate well for some diseases. Some treatments for diseases also lack robust biomarkers for drug-target engagement, making it di cult to gauge treatment performance during clinical trials [175]. Mechanistic modeling o ers a possibility to alleviate both of these issues. For instance, it may be infeasible for certain diseases to have all or most of the potential treatments tested in a pre-clinical or translational model [101, 320]. Speci cally, the treatment for Tuberculosis may involve a combination of 3-4 antibiotics over the course of 6 months. In order to demonstrate how quickly the treatment space for combination therapies increases, we show an example of the number of possible 3 and 4-treatment combinations if there are a total of 20 treatments. If there are 20 potential antibiotics, then there are 20 choose 3 (1140) plus 20 choose 4 (4845), resulting in 5985 potential combinations to test.

Mechanistic modeling o ers an avenue to narrow down this wide eld of potential combinations, aid in combination design, and prioritize combinations to test in pre-clinical models, hence de-risking and informing the translation into the clinic.

In the following sections, we will describe a conceptual, computational pipeline that describes how to capture pathologically-relevant metabolic pathways and utilizes available genetic data to predict treatment e cacy. As expected, the pipeline requires two key types of data: genetic data for gene expression levels and metabolic pathway data (e.g. RNAseq and metabolomics). This pipeline is built o of a variation of FBA techniques known as Linear-In-Flux-Expressions (LIFE) which encompass more complex dynamics with respect to traditional FBA [210]. A description of the LIFE method can be found in section 2.4.4. The LIFE methodology writes a system of ordinary di erential equations as$\dot{x} = S(x)\ f$, as opposed to$\dot{x} = S(f)\ x$, leveraging the natural linearity of biological systems w.r.t. uxes when encoding metabolic networks using hypergraphs. Non-linearities are captured via the implementation of Michaelis-Menten kinetics. In turn, the approach allows a better computational characterization of system equilibria and simulations [209, 218].

The computational pipeline allows for the simulation of single and combination therapies over metabolic pathways and is composed of multiple stages. We begin the paper with a Section 2.3 that o ers a rough overview of each of the steps, see Figure 2.1. This serves as a soft introduction to the language of the paper. The following Section 2.4 provides a detailed explanation of these same steps as in Figure 2.1. Then, in Section 2.5 we provide a concrete example by applying the pipeline to a speci c problem: the treatment selection for Tuberculosis expanding the methods of [212].