Fall 11-30-2022

A Pipeline to Generate Deep Learning Surrogates of Genome-Scale Metabolic Models

Achilles Rasquinha
University of Nebraska-Lincoln, achillesrasquinha@gmail.com

A PIPELINE TO GENERATE DEEP LEARNING SURROGATES OF GENOME-SCALE METABOLIC MODELS

by

Achilles Rasquinha

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professors

Tomáš Helikar and Massimiliano Pierobon

Lincoln, Nebraska

November, 2022

A PIPELINE TO GENERATE DEEP LEARNING SURROGATES OF GENOME-SCALE METABOLIC MODELS

Achilles Rasquinha, M.S.

University of Nebraska, 2022

Advisers: Tomáš Helikar & Massimiliano Pierobon

Genome-Scale Metabolic Models (GEMMs) are powerful reconstructions of biological systems that help metabolic engineers understand and predict growth conditions subjected to various environmental factors around the cellular metabolism of an organism in observation, purely *in silico*. Applications of metabolic engineering range from perturbation analysis and drug-target discovery to predicting growth rates of biotechnologically important metabolites and reaction objectives within different single-cell and multi-cellular organism types. GEMMs use mathematical frameworks for quantitative estimations of flux distributions within metabolic networks. The reasons behind why an organism activates, stuns, or fluctuates between alternative pathways for growth and survival, however, remain relatively unknown. GEMMs rely on manual intervention during their curation and annotation process, which can potentially induce substantial experimental bias. Also, solution spaces that cater to the flux distributions can be sensitive to the addition, updates, and deletions of metabolites and reactions and gene-enzyme-reaction rules within the model. Therefore, the quest for optimality can often be lost due to the number of hyper dimensions represented by these networks

Recently, Deep Learning (DL) has played a significant role in building function approximators for highly complex input datasets correlating in extremely large hyper dimensions. In this thesis, to address the computational costs associated with the simulations of GEMMs, we use an interpretable learning-driven approach to build surrogate GEMM models that act as alternatives to existent Flux Balance Analysis (FBA)-based approaches for predicting intracellular fluxes of reactions. We exploit the network characteristics of a well-curated input organism and build a synthetic subset of the flux cone containing thermodynamically feasible reaction growth rates. We then feed this dataset into a deep generative model capable of reconstructing intracellular flux values of the input organism. We evaluate its efficiency based on time-to-construct, accuracy, and ease of use. To provide a fair comparative analysis, we explore our learning approach with other traditional regression-based models and test our pipeline on three different input organisms subjected to network reduction techniques and different hyperparameters.

DEDICATION

To the Almighty, my parents, mentor, and friends.

ACKNOWLEDGEMENTS

I would like to offer my sincere gratitude to my advisor and mentor, Dr. Tomáš Helikar, for his immense support over the past many years and for his guidance and encouragement to help me explore challenging ideas in the field of System Biology and Computer Science. I am extremely grateful to him for letting me be part of his research group at Helikar Labs, UNL, and for motivating me in my work and research in order to accomplish a common goal. His ambitious visions in building a Virtual Immune System (VIS) for Biological System Modelers worldwide have made this research offer scalable possibilities in this particular domain and, at the same time, have offered me a noble platform for experiential learning.

I would also like to offer my personal thanks to Dr. Massimiliano Pierobon and Dr. Juan Cui for their consistent support in helping me pursue my research. Also, a huge thank you to Dr. Ashok Samal for taking the time to participate on my committee.

A huge thank you to all the members of Helikar Labs (especially Dr. Bhanwar Lal Puniya, Sara Aghmari, and Robert Moore) for providing me with their constant guidance, support, and direction in helping me achieve my research goals.

Lastly, I would like to thank my parents, Rebecca Furbeck, Rahul Prajapati, Heena Puri (for their valuable feedback), and my friends for offering me their unconditional love and being a pillar of support. Finally, I would like to offer my sincere thanks to Shivkumar Jaju, Shivani Tamkiya, Akshat Goel, Utkarsh Hardia, Twissa Mitra, Sanket Shinde, Navya Singh, Niyukta Kanjariya and Sailesh Pujara for believing in me to consider Graduate Studies at the University of Nebraska.

Table of Contents

List of Figures

Chapter 1

Introduction

Genome-Scale Metabolic Modeling (GEMM) provides information on metabolic activity in an organism and is often considered a *de-facto* standard in modeling metabolic-based biochemical systems [1]. GEMM typically represents a snapshot of the overall complex metabolic activity within an organism of interest [2]. A well-reconstructed GEMM contains a set of reactions and the various metabolites constituting them, as well as the gene-protein reaction (GPR) interactions governing the different possible states across the metabolic network [3]. GEMMs can also be used to simulate and predict the distribution of metabolic fluxes across reactions within an organism purely *in silico* [4]. In practice, GEMMs are capable of predicting the metabolic flux distribution for a given set of stoichiometry- and mass-balanced-based reactions of a biochemical system using linear optimization techniques such as Flux Balance Analysis (FBA) [5], thereby predicting the overall growth rate of an organism or the rate at which a metabolite is produced.

GEMMs have been applied in the field of strain development to produce biotechnologically important materials and potential drug target discovery against infectious microorganisms (bacteria, viruses, etc.) [6]. Flux Balance Analysis (FBA), a popular metabolic analysis method is often used in resolving knowledge gaps

(gap-filling) within metabolic networks, thereby refining the overall systematic investigation of the metabolic activity of organisms purely *in silico* that would otherwise be time-consuming and hard to validate *in vivo* [7]. However, one of the biggest limitations of such a method is that it is only relevant if the genomic reconstruction of the network is closely similar to that of the actual wild-type observed. This means that the model is worthwhile for as long as it has been well-curated, thus limiting itself to generalization and robustness to variational change concerning annotations, conditions, and constraints [8].

The first ever known GEMM was one of the *Haemophilus influenzae RD* in 1999 - constituting 343 metabolites and 488 reactions to interpret phenotypic behavior from its genotypic information embedded within the model [9]. Currently, there are over 6239 manually and automatically curated GEMMs of various bacteria and eukaryotes across a vast number of databases [10]. In recent years, cell-type specific GEMMs have been reconstructed to understand diseases within humans, such as obesity, diabetes, etc., thanks to the Human Metabolic Reaction (HMR) series of GEMMs for tissue-specific and cancer-specific cell types [11]. Accurately predicting objectives leading to a disease, thereby narrowing the bridge between genotype (the encoded information within organisms) and its phenotype (distinguished characteristics and functions), remains a problem in exploration.

1.1 Research Challenges

Even with such limitations, there is still no doubt that GEMMs are extremely relevant in the field of metabolic engineering of biological systems [12]. However, with the rise in the number of GEMMs reconstructed (especially in the case of cell-specific

models of the human genome), there is a dearth need for more efficient and scalable analysis to understand and simulate the complexities of single or multi-cell organisms

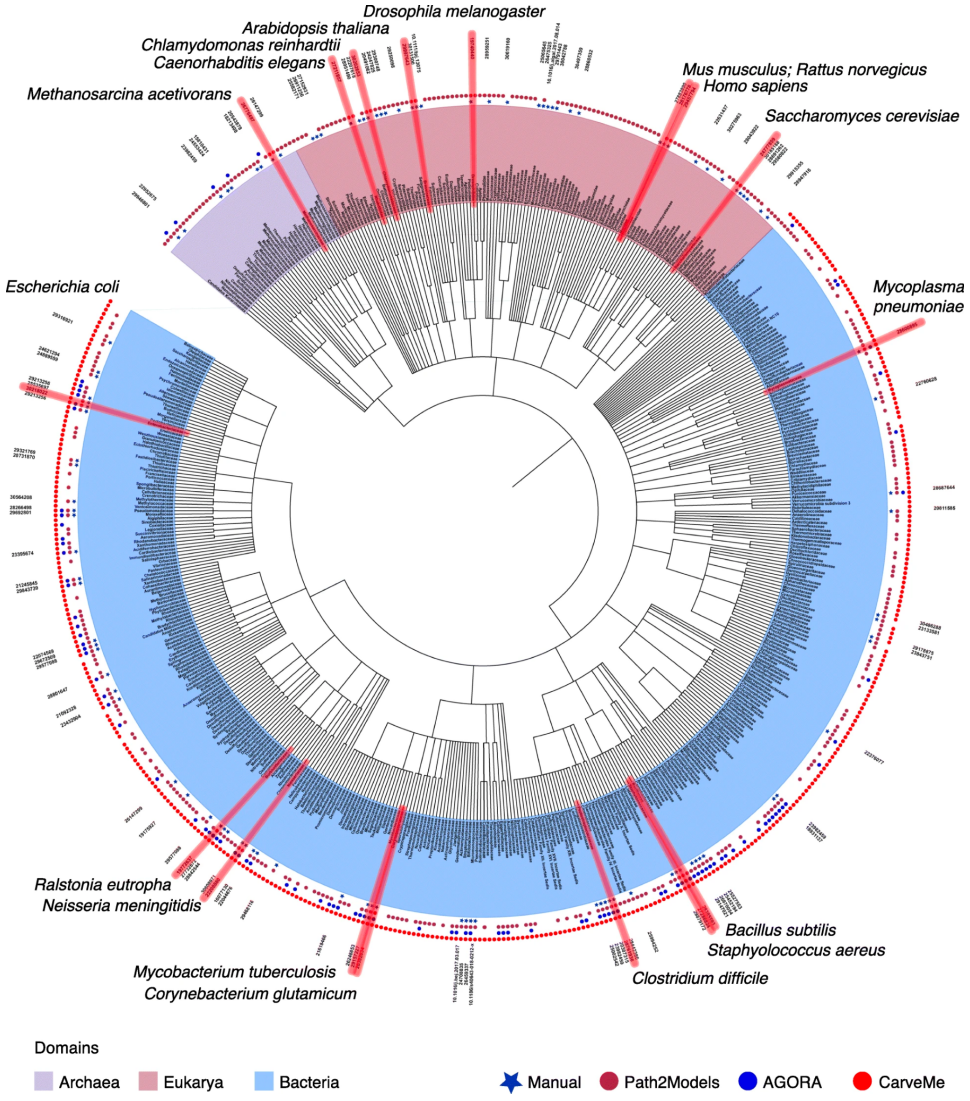

Figure 1.1. *A phylogenetic tree of all of the GEMs reconstructed to date at the family level.*

(Source - [10])

system *in silico*, particularly in understanding the behaviors and effects as well as the outcomes of introducing different pathogens within a multi-tissue and multi-organ system. When there is a widely accepted well-curated GEMM, reaction flux distributions predicted using FBA are generally almost consistent with actual lab experimental conditions [13]. In some cases, such predictions are further enhanced by inducing

additional information (testing the availability of certain compounds within metabolic pathways, introducing systematic chemical properties as additional information, biophysical capabilities, etc.) [14] [15]. As mentioned, although FBA at its core is a form of mixed integer linear-based optimization, is however limited to the annotated information provided within the model of interest, i.e., a feasible solution closest to the actual growth rates within the organism is only replicable if the model in interest has been qualitatively reconstructed. Therefore, a simple change in the constraint space, be it internal (gene knockouts) or external (expanding or shrinking bounded solution space) can heavily affect the growth estimates of such models [8]. Although FBA provides insights into the outcomes of external constraints imposed, it does not necessarily speak much on behalf of the constraint space itself. When the metabolic activity of different pathways is highly sensitive even with extraneous conditions, providing underlying insights into the genotype-phenotype relationship remains anything but complex. Also, efficient FBA fundamentally relies on the Moreover, a large model can present technical limitations in estimating such biological fluxes with an increase in experimental and environmental parameters. Similarly, simultaneously simulating a large number of metabolic models (say, a community of models for instance) could potentially lead to higher computational resource utilization.

1.2 Contributions

In this thesis, we examine the use of Deep Learning as a technique for building growth simulators as surrogates for Genome-Scale Metabolic Models (GEMMs) with a major goal of addressing the computational cost challenges associated with GEMM analysis. A state-of-the-art available Deep Learning model was used and enhanced to develop predictive simulators that accurately estimate and simulate the metabolic flux

distribution of a finely reconstructed and well-curated GEMM of an organism of interest to a very high degree of approximation. The generated and evaluated models from our Deep Learning-based pipeline are instantaneous in runtime compared to currently available traditional frameworks [5]. This elevates the scale to compute many GEMM simulations simultaneously and efficiently, particularly in use cases wherein the closest system reconstruction to the original biological system is required to achieve a better understanding of the organism in a given environment (for instance, a digital twin of the immune system) [16]. It is shown that the proposed pipeline can generate simulated models within constrained computational time and memory usage and it can mimic traditional approaches with little to no reconfiguration. Considering such alternative approaches during perturbation analysis for drug discovery in favor of currently available methods can enhance the overall pipeline for metabolic-based analysis of organisms, especially in cases where multiple and parallel flux simulations are expected to be executed in real-time. The work done within this thesis is available as an open-sourced package titled "DeepGEMM" for researchers and users to help build custom surrogates of an input GEMM of choice. It has been released under the GNU General Public License v3 and is currently being publicly developed and maintained on GitHub at https://github.com/HelikarLab/DeepGEMM.

1.3 Thesis Organization

In this section, we discuss the various Chapters covered in this thesis.

- Chapter 2 discusses the current trends and related work in Machine and Deep Learning with Genome-Scale Metabolic Modeling.
- Chapter 3 constructively defines the problem and the details of the implementation of our proposed framework to predict metabolic fluxes using a DL approach, experiments conducted to produce our dataset, and simulations to evaluate flux distributions.
- In Chapter 4, we discuss the details and results of applying our proposed methodology.
- Chapter 5 concludes the thesis and briefly discusses potential future work.

All chapters have been formatted based on the guidelines provided by the University of Nebraska-Lincoln. All references have been provided at the end of this thesis.

Chapter 2

Related Work and Background

2.1 Outline

This thesis is intended for readers from a wide range of backgrounds, especially those in the biological and computational sciences. As part of Chapter 2, we provide a comprehensive overview of prior works that have used Machine Learning and Deep Learning techniques integrated with metabolomics pipelines. Following a brief background, we outline our motivation by first bringing forward some caveats in existing works and then addressing them through our approach, thus defining our objectives. Lastly, we elaborate on some concepts to make sure that the reader understands specific terminologies, frameworks, and equations used throughout this thesis.

2.2 Background

Machine Learning (ML) is a field of computer science and mathematics that deals with algorithmic and statistical techniques for modeling and analyzing data [17]. The subject is closely related to the study of artificial intelligence (AI), though ML does not

always have a clear distinction between these two areas. In recent years, ML has turned out to be a vital tool in elevating flux analysis in GEMMs and helping understand environmental factors that affect cell phenotypes [18]. Giuseppe et al. estimated the production of lactate in Chinese Hamster Ovary (CHO) cells using linear regression on gene expression profiles [19]. Wu et al. could predict fluxomics of heterophobic bacteria based on their kind, types of substrate, aerobic/anaerobic conditions, and cultivation methods using an ensemble of ML methods, namely SVM, k-NN, and Decision Trees [20]. Folch-Fortuny et al. applied an unsupervised approach named Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS) to offer a more meaningful representation of metabolic pathways in *P. pastoris* [21]. Szapannos et al. [22] used genetic algorithms to help improve the prediction of gene-gene interaction networks to bridge the gap between empirical and computational studies in yeast metabolism.

Deep Learning (DL) refers to a subset of ML and AI [23]. It is particularly concerned with large Artificial Neural Networks (ANNs). The primary idea is that intelligence can be expressed as a set of layers, with each layer using the output from the previous layer to produce its next output. DL has been successful in a variety of non-linear problems and unstructured data, especially wherein ANNs help identify such important features and patterns in data [24]. The ability to find these patterns means DL can be applied to many different industries, including finance, health care, and even biological engineering [25].

ANNs are inspired by the human brain, and their structure is similar to that of human neurons [26]. Each neuron within the ANN constitutes a predefined activation function that translates the incoming input from each layer to an output signal strength that validates the input. The larger the scale of the network in terms of the number of neurons and layers, the "deeper" the network is. Such deep networks can be configured

for a wide range of tasks ranging from speech recognition, object classification on images, and text translation [27]. Due to the availability of high-throughput multi-omics data in modern biology, the data boom has also paved the way for integrating data reduction, selection, and translation tasks in systems biology with Machine and Deep Learning [28].

There have been a limited number of papers involved in the integration of Deep Learning and Systems Biology, especially in the area of Genome-Scale Metabolic Models (GEMMs) that use a kind of learning technique called Autoencoders (AE). AEs are a type of deep learning that is self-learning [29]. They can encode data into an efficient representation without any human intervention. In the last few years, autoencoders have produced significant results in image and speech recognition and in the process of machine translation [30]. As an example, Guo et al. [31] consider an approach titled 'DeepMetabolism' which consists of a 5-layered AE wherein the first 2 layers of the network are connected by Gene-Protein Rules (GPR) annotated within a GEMM model offered in SBML format [32]. Statistical simulations from Flux Balance Analysis (FBA) were used to connect the 3rd and 4th layers of the autoencoder network. Using experimentally measured transcriptomic profiles, the decoding layers of the AE were then expected to reconstruct phenotypic relationships from its genotype with very high accuracy.

Variational Autoencoders (VAE) are a type of Deep Learning technique that has been known to be successful in generating new and unseen data which closely resembles the original input [33]. They work by encoding such data into a latent vector which is then used to generate new unseen data within the same distribution using an encoder-decoder network. In their approach, Barsacchi et al. [34] used a paradigm dubbed GEESE (Gene Expression latEnt Space Encoder) that uses a β-VAE architecture to

relate gene expression profiles to the regulation of reaction fluxes by generating synthetic gene expression data.

2.3 Motivation

In this thesis, we thoroughly examine and build upon prior works of integrating Machine and Deep-Learning-based methods to experimentally generate GEMMs emulators and validate whether the proposed approach is both effective and efficient in terms of introducing it within metabolic engineering analysis pipelines. Even though Barsacchi et. al [34] works have introduced the possibility of such an approach, almost all prior works rely on a multi-omics strategy by introducing at least some form of experimentally conducted knowledge generated within laboratory experiments (gene expressions, transcriptomic profiles, etc.). Thus, the need for such experimental data limits the overall scope of expansion of such approaches to GEMMs for other organisms or cell types. In addition, steps to perform a systematic comparison of Deep Learning-based methods with other approaches have not been taken in the case of leveraging FBA. Moreover, advanced modeling approaches in Deep Learning have risen exponentially in recent years with the introduction of new and updated sophisticated networks that significantly overcome many drawbacks (training time, better data representation, etc.) of its predecessor networks [35]. In this thesis, we consider using a Wasserstein Conditional Generative Adversarial Network (with Gradient Penalty) (WCGAN-GP)-based approach [36] by using pure synthetic data generated using Monte Carlo simulations from reconstructed GEMMs available in standardized formats. This eradicates the need for additional multi-omics data by only using the refined truth embedded within gene-protein reaction (GPR) links available within the annotated model of interest. Conditional Generative Adversarial Networks (CGANs) have recently

been shown to be powerful semi-supervised alternatives to regression modeling techniques and generate novel data points [37]. To the best of our knowledge, this work would be the first to utilize WCGAN-GP as an alternative to standard regression models and validate its usage for even other regression-based problems.

The primary objective of such an approach is to

1. Validate whether such a 'black-box' model is capable of generating unseen genotype-phenotype relationships.
2. Reduce the computational time to perform FBA (by initially training the network with a considerable amount of data points in a constraint space subset).
3. Produce possible feasible solutions for constraints that were initially thermodynamically unfeasible.

To compare our approach, we consider a popular regression-based machine learning technique - Linear Regression [38] concerning a WCGAN-GP-based approach. Each method has its own learning and prediction times that are considered evaluation variables for our comparisons. At the same time, we also compare our WCGAN-GP-based approach across 3 GEMMs representing different organisms in nature - ranging from non-pathogenic prokaryotes to parasites to organisms of varying sizes (as observed in nature).

To keep compliance with FAIR research [39], we implement a high-throughput pipeline named 'DeepGEMM' using a customized self-implemented version of `cobrapy` [40] and the standard off-the-shelf available Deep Learning library - TensorFlow [41] as governing frameworks for reproducible analysis of this work and recreation of DL-based GEMM emulators. Our implementation of the DL pipeline is written in the Python language with 'convention over configuration' in mind. The choice of language and frameworks was based on using only the most recently advanced methods, ease of

use, and widely-available access, as well as the interpretability of software across all platforms. In terms of training and analysis, we use the Holland Computing Center's high-performance computing resources offered at the University of Nebraska-Lincoln to generate synthetic datasets and learning weights of the model emulators. This was done to ensure a standardized comparison across all methods concerning the proposed WCGAN-GP-based model.

2.4 Genome-Scale Metabolic Models

Synthetic Biology is a science that seeks to design and create new organisms and synthetic pathways by manipulating genomes resulting in new phenotypes [42]. In their eight-year-long project, Beyer et. al published their scientific insights into - the Golden Rice, created by introducing the β-carotene (a *vitamin A* precursor) biosynthetic pathway into *Oryza sativa* (traditional Asian Rice) [43]. Mehta et al. helped improve the juice quality, vine life, and lycopene in tomatoes [44] by using genetic modifications that resulted in an increase in polyamine, spermidine, and spermine accumulation during ripening. Systems Biology has the potential to reimagine how biologists design and apply genetic mutations *in silico*, leading to the discovery of new biological phenomena without the need to conduct experiments *in vivo* [45]. The domain consists of various mathematical and computational frameworks for modeling, simulation, and analysis of various biological systems, be it uni-, multicellular, or even a community of organisms interacting within an environment. Systems Biology has heavy applications in a variety of problems, such as cancer research and drug development [46]. In this thesis, we focus on a particular kind of computational modeling technique in Systems Biology that deals with the metabolism of biological systems called Genome-Scale Metabolic Models (GEMMs). Genome-Scale Metabolic Models (GEMMs) are a powerful

tool for understanding the metabolism within organisms and the fundamental processes and relationships that allow cells to grow, divide, and produce energy [47].

Note that such models are larger in size and can often contain more than thousands of species, reactions, and enzyme information for even small micro-organisms [10]. To reconstruct an organism of interest to its genome-scale, several algorithms have been developed to provide a well-curated reconstructed draft to a refined level of detail by extracting annotated genes from raw RNA sequence data [48] [49]. While we assume the availability of a well-reconstructed GEMM during our analysis, the work done within this thesis is not limited to the same. In the following section, we elaborate on the mathematical framework governing GEMMs - Flux Balance Analysis.

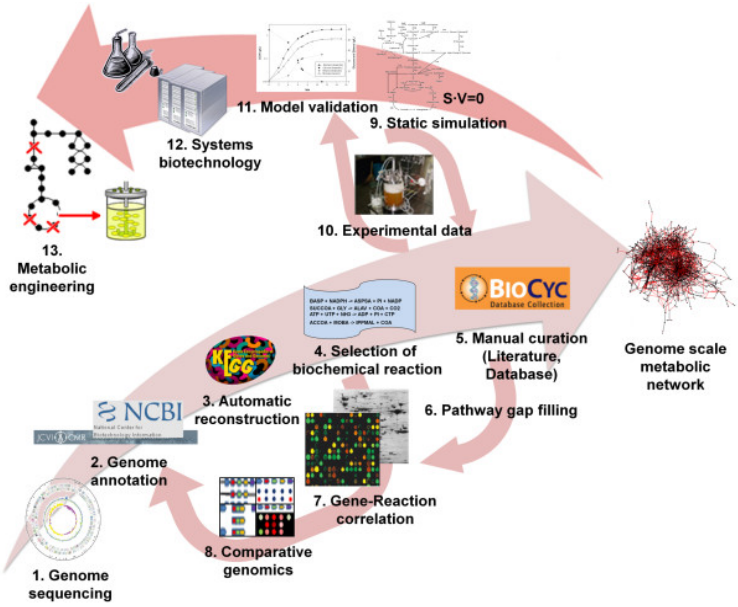

Figure 2.1 *Genome-Scale Metabolic Model reconstruction of Z mobilis. and analysis using a classical metabolic engineering pipeline.* (Source - [50])

2.5 Flux Balance Analysis

Flux Balance Analysis (FBA) is a well-known technique for modeling the metabolism of living systems. It is used to understand and predict the levels of metabolic intermediates, enzyme activity, and fluxes [5]. In this technique, constraints are imposed on a model to reduce the number of free parameters and also constrain the model to be consistent with experimental data. FBA then uses mathematical optimization frameworks that use linear equations to help maximize the growth of living cell populations [3].

As a result, such a technique helps modelers explore and understand the effects of small changes in nutrient concentration or other factors on the rapid growth of cells [51]. In this thesis, we are interested in devising a surrogate model for a given organism that is also capable of approximating growth rates within the organism subjected to certain environmental conditions by substituting a traditional FBA framework with a Deep Learning surrogate. The objective is not to completely replace traditional FBA but rather elevate the overall process of modeling biochemical metabolic pathways and, at the same time, effectively reduce computational costs.

At the core of FBA is a mathematical representation of the Genome-Scale Metabolic Model (GEMM) in a matrix format, often called the Stoichiometric Matrix (S) [52]. S represents the stoichiometry of each biochemical reaction interacting within the biochemical network of an organism. Generally, the stoichiometric matrix of a typical GEMM can be represented as follows:

$$\begin{bmatrix} S_{0,0} & S_{0,1} & S_{0,2} & \cdots & \cdots & \cdots & S_{0,r-3} & S_{0,r-2} & S_{0,r-1} \\ S_{1,0} & S_{1,1} & S_{1,2} & \cdots & \cdots & \cdots & S_{1,r-3} & S_{1,r-2} & S_{1,r-1} \\ S_{2,0} & S_{2,1} & S_{2,2} & \cdots & \cdots & \cdots & S_{2,r-3} & S_{2,r-2} & S_{2,r-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ S_{m-3,0} & S_{m-3,1} & S_{m-3,2} & \cdots & \cdots & \cdots & S_{m-3,r-3} & S_{m-3,r-2} & S_{m-3,r-1} \\ S_{m-2,0} & S_{m-2,1} & S_{m-2,2} & \cdots & \cdots & \cdots & S_{m-2,r-3} & S_{m-2,r-2} & S_{m-2,r-1} \\ S_{m-1,0} & S_{m-1,1} & S_{m-1,2} & \cdots & \cdots & \cdots & S_{m-1,r-3} & S_{m-1,r-2} & S_{m-1,r-1} \end{bmatrix}$$

Here, $S_{(i,j)}$ it represents the i^{th} metabolite and the j^{th} reaction in the metabolic network. m and r represents the cardinality of the metabolite and reaction sets respectively. Typically, the stoichiometric matrix of even the smallest of all organisms would be represented as a sparse matrix. In fact, the GEMM reconstruction of the smallest genome available - *M. genitalium* (iPS189) - consists of at least 274 metabolites and 262 reactions [53]. In addition, a GEMM also consists of a set of Gene-Protein Reactions (GPR) rules that can be represented by independent regulatory networks determining the potential activation states of reactions associated with it. Gene-Protein Reaction (GPR) rules represent the possibility of a reaction to achieve metabolism based on proteins produced by an organism, i.e., the effect of a gene on growth rates of metabolic reactions [5].

A GEMM consists of an objective function (or many) to be either maximized or minimized. In a biological sense, this is typically associated with the maximization of a biotechnologically important metabolite or reaction or minimizing its nutrient uptake. For most organisms, the primary objective for survival is to maximize their *biomass production* [54]. This is achieved by breaking down external compounds (oxygen, sugar, etc.) to multiply**.** Hence, one can derive a set of system equations from a GEMM containing a considerable number of distinct variables and a set of ordinary differential equations (ODEs). These equations represent the rate at which a chemical reaction occurs within the biochemical pathways of the organism, subject to a vast number of constraints. Ideally, the state of the system can be defined as follows:

maximize $\mathrm{v}_{objective}$

subject to

$$\begin{bmatrix} S_{0,0} & \cdots & S_{m-1,0} \\ \cdots & \cdots & \cdots \\ S_{0,r-1} & \cdots & S_{m-1,r-1} \end{bmatrix} \begin{bmatrix} v_0 \\ \cdots \\ v_{r-1} \end{bmatrix} = \begin{bmatrix} \frac{dM_0}{dt} \\ \cdots \\ \frac{dM_{m-1}}{dt} \end{bmatrix}$$

and

$$v_{l(j)} \leq v_j \leq v_{u(j)}$$

Here, v_j represents the growth rate of a chemical reaction (or simply, the metabolic flux), whereas here, M_i represents the concentration of a metabolite i. $v_{l_{(j)}}$ and $v_{u_{(j)}}$ represent the lower and upper bounds of the solution space. Notice that the vector S has been transposed to fit the equation correctly.

The flux through a chemical reaction is generally expressed as $\frac{mmol}{gDW}$ where $mmol$ is millimole, gDW^{-1} which is the dry weight of the organism. In the case wherein the objective is to maximize biomass production, we are specifically interested in the doubling rate of the organism. Hence, the unit for the same can be expressed as $\frac{mmol}{gDW.hr}$, where hr^{-1} represents the organism's doubling rate per hour. Flux Balance Analysis assumes a Pseudo Steady State Hypothesis (PSSH) of the biological system [56], i.e., it assumes that the concentration of each metabolite does not change over time, thus eliminating information related to enzyme kinetics. Therefore, the system can be minimized as a set of pure linear equations as follows:

$$\textbf{maximize } \mathrm{v}_{objective}$$

subject to

$$\begin{bmatrix} S_{0,0} & \cdots & S_{m-1,0} \\ \cdots & \cdots & \cdots \\ S_{0,r-1} & \cdots & S_{m-1,r-1} \end{bmatrix} \begin{bmatrix} v_0 \\ \cdots \\ v_{r-1} \end{bmatrix} = 0$$

and

$$v_{l(j)} \leq v_j \leq v_{u(j)}$$

We narrow down the system as a pure mixed integer linear programming (MILP) problem subjected to other constraints (in a biological sense, this generally relates to either limiting the growth rate of reactions or important nutrients). Figure 2.2 denotes the narrowing of the convex polytope solution space based on constraints imposed. In the case of biological systems subject to a large number of chemical reactions, such a polytope is hard to visualize in an n dimensional space.

The applications of Flux Balance Analysis in Systems Biology have been endless. Japhalekar et. al. were able to use FBA as a means to prove the overproduction of organic acids in Cyanobacteria (*Synechocystis sp. PCC 6803*) in dark anoxic conditions [56].

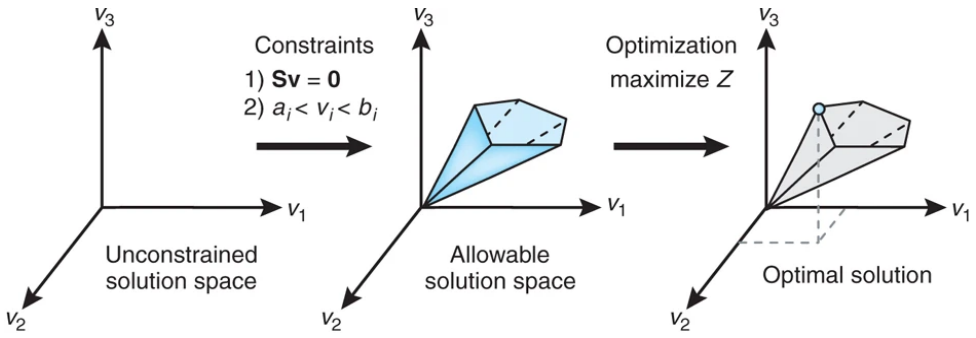

Figure 2.2 *Narrowing Solution Space due to constraints imposed by metabolic reaction bounds in Flux Balance Analysis* (Source - [5])

Da Veiga Moreira et al. were capable of optimizing citrate overproduction in yeast (*Yarrowia liplytica - iYali4*) using dynamic Flux Balance Analysis (an extension to FBA)

[57]. There is therefore no doubt that Flux Balance Analysis (FBA) has been a vital tool in Systems Biology having a wide range of applications [58]. In the next section, we will elaborate on the Deep Learning methodology used to help us build GEMMs emulators for organisms.

2.6 Wasserstein Conditional GANs *(with Gradient Penalty)*

Generative Adversarial Networks (GANs) are a class of Deep Learning algorithms that can create new unseen or modified data points by learning low dimensional representations of an input dataset. These neural networks use two deep learning models: a *generative* model that creates data and a *discriminative* model that evaluates data created by the generative network [59]. GANs have recently made tremendous progress in the field of bioinformatics, ranging from medical image analysis and processing to the realistic generation of single-cell RNAseq data [60]. Ghahramani et. al. were able to use a WGAN-GP to achieve a universal representation of the Epidermal Differentiation Complex (EDC) and predict cell state perturbations on gene expression profiles [61]. Recently, Cao et. al. used string-based Simplified Molecular-Input Line Entry System (SMILES) representations of compounds as input data into a GAN that generates new unknown valid molecules [62].

Traditionally, the two feed-forward neural networks act against each other like a zero-sum non-competitive minimax game to minimize the independent errors fed back into the network for learning. Ideally, the objective of a generator G is to trick the discriminator D by offering fake samples, whereas the objective of the discriminator is to not be cheated during its critiquing process. If x represents a real sample and p_r is the data distribution over x, then the critical decision made by D over x must be accurate by maximizing $\mathbb{E}_{x\sim p_r(x)}[\log D(x)]$. In the case of a fake sample $G(z)$ over an input noise z,

the discriminator must estimate a probability $D(G(z))$, closer to zero by maximizing $\mathbb{E}_{z\sim p_r(z)}[\log(1 - D(G(z)))]$. Hence, the overall loss function to be optimized by the GAN would be:

$$\min_G \max_D L(D, G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_r(z)}[\log(1 - D(G(z)))]$$

$$\min_G \max_D L(D, G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x))]$$

Note that the term $\mathbb{E}_{x\sim p_r(x)}[\log D(x)]$ does not cater to the generator during training. Due to this, GANs generally suffer from convergence instability as they attempt to find the Nash Equilibrium during convergence. This generally occurs when the distributions $p_r(x)$ and $p_g(x)$ are disjointed from each other.

2.6.1 Wasserstein Metric and K-Lipschitz Continuity

Movement from one distribution to the other can be intuitively thought of as moving a unit of dirt from the earth from one pile to the other. Therefore, the minimum energy cost it takes to transfer such a pile can be thought of as the distance between these distributions where the cost would be the amount of dirt moved times the mean distance between the piles. In the case of our GAN, the Wasserstein Metric between the data distributions $p_r(x)$ and $p_g(x)$ is given as follows:

$$W(p_r(x), p_g(x)) = \inf_{\gamma\sim\prod(p_r(x),p_g(x))} \mathbb{E}_{(x,y)\sim\gamma}\|x - y\|$$

Here, $\prod(p_r(x), p_g(x))$ represents a set of all possible joint probability distributions between $p_r(x)$ and $p_g(x)$. Arjovsky et. al [63] provides a modified formula to the above equation to compute the distance between discrete distributions as follows:

$$W(p_r(x), p_g(x)) = \frac{1}{K} \sup_{\|f\|_L\leq K} \mathbb{E}_{x\sim p_r(x)}[f(x)] - \mathbb{E}_{x\sim p_q(x)}[f(x)]$$

Here, $\sup$ is the supremum (opposite to $\inf$, the infimum). f is expected to satisfy the $\|f\|_L \leq K$ term to be K-Lipschitz continuous.

2.6.2 *Wasserstein Loss Function*

The GAN's discriminator is now trained to learn a K-Lipschitz continuous function to calculate the Wasserstein distance between the real and fake sample distributions. With a decrease in discriminator loss, the generator is then expected to produce fake samples closer to the actual distribution of the input dataset. If f is a K-Lipschitz continuous function subjected to parameters $\theta \in \Theta$, then the discriminator is expected to learn θ in order to find the best fit for f_θ with a loss function as follows:

$$L(p_r(x), p_g(x)) = W(p_r(x), p_g(x)) = \max_{\theta \in \Theta} \mathbb{E}_{x \sim p_r(x)}[f_\theta(x)] - \mathbb{E}_{x \sim p_q(x)}[f_\theta(g_w(z))]$$

The weights are then clipped within a pre-defined bounded region in order to enforce K-Lipschitz continuity of the function f_θ. However, this might be a poor approach towards restricting the function from being K-Lipschitz continuous. To overcome this, Gulrajani et. al offered an alternative solution by penalizing the gradient weights during the training process. On doing so, it promises a faster means for convergence of the two distributions being within proximity [36].

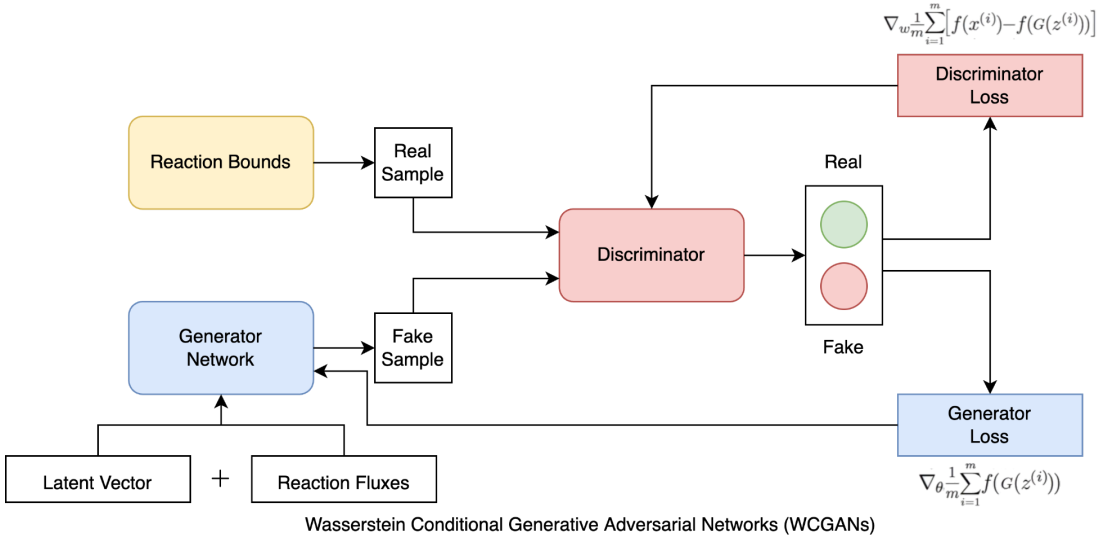

Figure 2.3 *Wasserstein Conditional GAN Architecture for DeepGEMM*

2.6.3 Conditional GANs

Conditional Generative Adversarial Networks (CGANs) are a type of GAN that can generate data conditioned on an input. The idea of CGANs is to have a generator model that outputs fake data given some signal and a discriminator model that tells the generator whether its output is real or fake based on that signal. Conditional GANs are considered to be a kind of semi-supervised learning approach since both the networks are conditioned to generate input closely resembling the one provided. We utilize this crucial feature of Conditional GANs to help the learning model approximate the flux value of a reaction based on input constraints. Figure 2.3 shows an architecture diagram of the proposed WCGAN-GP network with respect to our flux predictor.

In the next section, we provide a detailed outline of the proposed pipeline to output model surrogates of the input metabolism. We also exploit the network

characteristics of the input GEMM to minimize the number of feature dimensions required during GAN training and evaluate the validity of our approach in detail.

Chapter 3

Approach

3.1 Outline

In our approach, we first define our problem at hand and then outline an overview of the various modules that constitute our pipeline. Next, we describe how Monte Carlo sampling was used to create a synthetic dataset to be fed into our deep generative model. Following that, we use the original input metabolic network and consider the idea of dimensionality reduction through the minimization of its reactome. Finally, we outline the internal workings and configuration (hyperparameters and training) of our WCGAN-GP network.

3.2 Problem Definition

The main problem that we aim to resolve in this research is defined as follows: *Given a Genome-Scale Metabolic Model (GEMM) of an organism* M*, where* M *represents* (m, r, g)*,* m *is a set of metabolites,* r *is a set of reactions and* g *is a set of boolean-based gene-protein reaction (GPR) rules, train a Deep Generative Model to predict the corresponding flux distribution of* r *(say* J'*) by minimizing the error loss from its actual flux distribution* J

of r and at the same time, performing it efficiently in terms of computational time and resource utilization.

3.3 Pipeline Overview

Our DeepGEMM pipeline can be bifurcated into different modules as follows:

- ***Synthetic Data Generation:*** A Monte Carlo sampling of the convex polytope hyperspace was considered to generate reaction flux distributions of the input metabolic network across various constraints.
- ***Dimensionality Reduction (minimization of network):*** We reduce the genome-scale metabolic network to a minimal network by removing non-functional metabolic reactions that do not hinder target growth conditions yet maintaining the overall functionality of the base network.
- ***Parameterization and Training:*** We format our WCGAN-GP network corresponding to the dimensionality of the reduced metabolic network, feeding in our synthetic data for eventual training of our conditional GAN and its hyper-parameterization, thereby attempting to minimize the loss between the original flux distribution and the reaction fluxes generated from the GAN network.

3.4 Synthetic Data Generation

We propose a methodology to synthetically generate reaction flux distributions subjected to different kinds of model mutations and constraint limitations. Figure 3.1 illustrates the detailed workflow of generating such a synthetic dataset. In this approach, a Monte Carlo scheme of flux sampling random points from the solution space was considered as sample data points for the dataset.

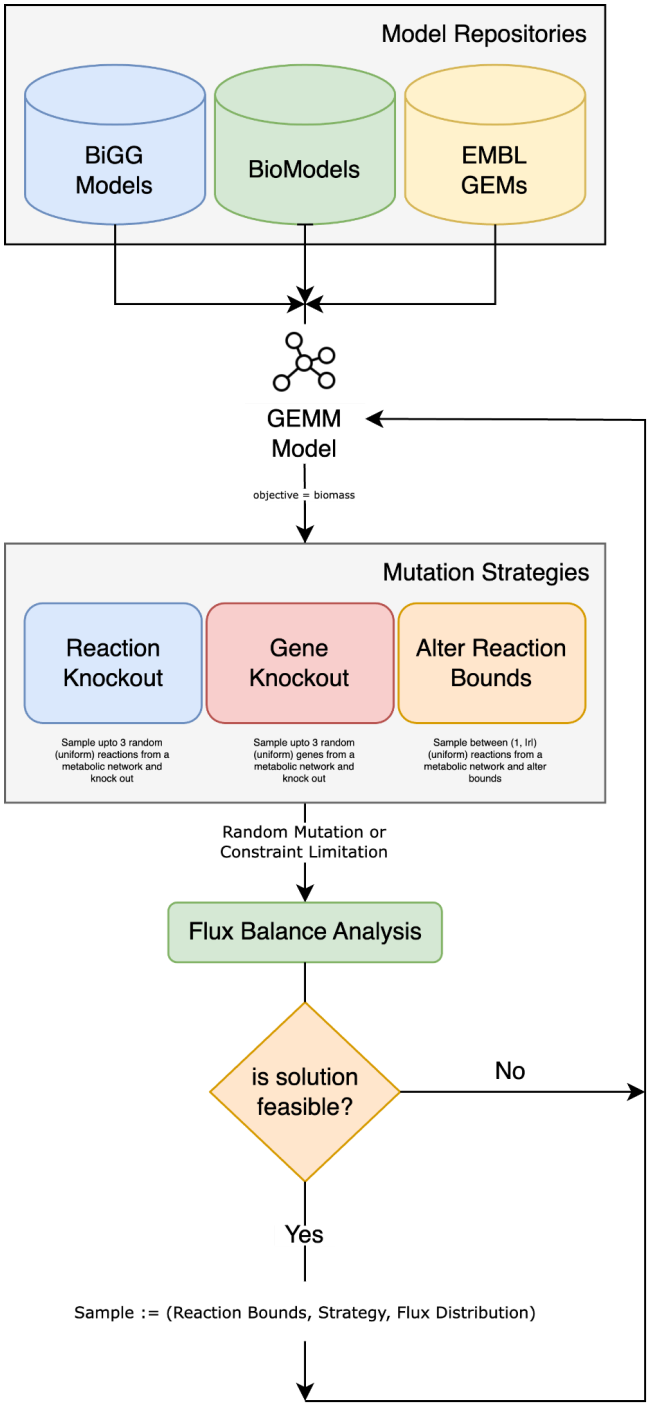

Figure 3.1 *Monte Carlo Flux Sampling for Synthetic Data Generation of a given Metabolic Network*

Monte Carlo simulations of the effective solution space of a GEMM has proven to be a useful approach in extracting important properties of the metabolic network - like the hypervolume of the convex polytope flux distribution space [64], topological and subsystem information of the network [65], thereby remaining robust to variations in constraints.

Given N (the number of sample points to be generated), we first consider an input metabolic model from a model repository (BiGG, BioModels, EMBL GEMs, etc.). [66] [67] [48] and randomly (uniformly) pick a mutation strategy to be applied to the model. Note that throughout the course of performing mutations or constraint-restrictions, unless undefined, we consider the pseudo reaction - *wild-type biomass* to be the defined objective for the input organism. The mutation/constraint-limitation strategies devised for this approach are as follows:

1. ***Reaction knockout (Mutation):*** An *in silico* reaction knockout is achieved by limiting the growth rate bounds to 0 $\frac{mmol}{gDW}$ of the reaction in the metabolic network. Under *in vivo* conditions, it is feasible to perform up to triple reaction knockouts for a given organism [68]. Thus, we consider random subsampling (uniform) of up to 3 reactions in this stage that performs a single, double or triple knockout analysis.
2. ***Gene knockout (Mutation):*** Similar to the approach mentioned in 1., we subsample up to three genes (random uniform) that perform a single or multi-knockout mutagenic simulation. A metabolic network generally contains gene-protein-reaction links as regulatory rules which indicate the relations between activation states of gene products and the corresponding enzyme catalyzation of reactions. Gene knockout mutations also offer Gene Coupling (dependence of the activation state of a single gene affects multiple reactions

across the metabolic pathway), information regarding essential genes and genetic robustness of the organism [69] [70].

3. ***Altering Reaction Bounds (Constraint-Alterations):*** In the biological sense, alteration of reaction constraints typically correspond to altering growth conditions of reactions or nutrient uptake. We consider subsampling (randomly uniform) a subset of reactions and alter its lower and upper bounds (random uniform) within the range:
    - -1000 $\frac{mmol}{gDW}$ and 1000 $\frac{mmol}{gDW}$ if the reaction is reversible.
    - 0 $\frac{mmol}{gDW}$ and 1000 $\frac{mmol}{gDW}$ otherwise.

The artificially induced ranges generally cover the diffusion limits for many reactions since the diffusion rates for even the largest metabolites are approximately 100 mM.

We consider the dataset to comprise both natural and perturbed state information of the organism of interest [71]. Such mutation strategies are also viable to infer the minimization of metabolic adjustments (MOMA) required for the organism to survive and grow, thereby providing a diverse representation of the flux distribution cone within our dataset. We perform an FBA over the perturbed/altered metabolic model and predict the uptake rates of all reactions within the network using Linear Programming (LP). A simulation is considered successful with a sample point generated if the solution is feasible else it is disregarded from the synthetic dataset. Finally, a sample point consists of the lower and upper bounds of each reaction, the mutation strategy used to generate the corresponding flux distribution, and the solution fluxes itself, based on the constraints imposed through such mutation. The matrix dimensions

of the resultant dataset is $(3|r|+1, N)$ where $|r|$ denotes the cardinality of the reaction set of the input metabolic network.

We implement our synthetic data generator script in Python and parallelly perform perturbations for each strategy. For a sample model, `e_coli_core` comprising 72 metabolites and 95 reactions, the average time it took to generate 1 million sample points with 8 processors was approximately 9 hours. Figure 3.2 visualizes a plot between the size (metabolites and reactions) of 1213 GEMM networks over the time it takes to generate 1000 feasible synthetic data points.

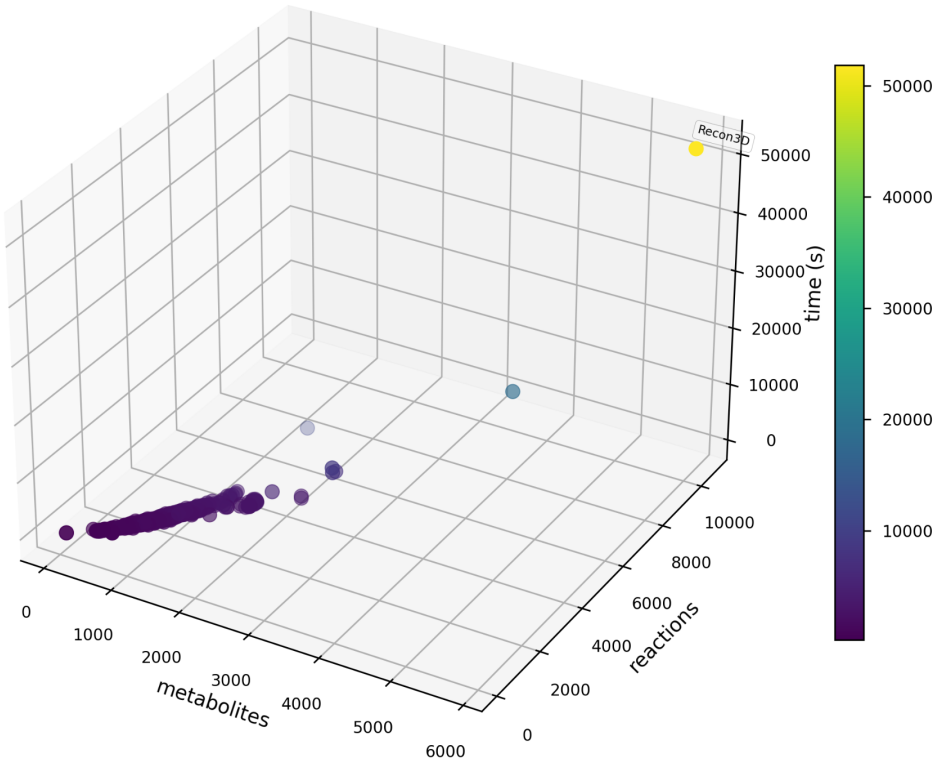

Figure 3.2 *Synthetic Data Generation of 1213 GEMMs over Time (seconds) (1000 samples)*

The networks were randomly selected from a large collection of GEMMs available from various model repositories. As a result, we conclude that the size of an annotated GEMM network, and the time necessary to generate feasible sample points follow a

linear relationship. This is natively trivial since larger networks contain sparser matrices to derive possible solutions.

Similarly, much can be said about the number of infeasible solutions generated using our Synthetic Data Generator with respect to the size of the metabolic network. Figure 3.3 visualizes the relationship between the size (reactions and genes) of 1213 GEMM networks over the number of infeasible solutions generated when $N = 1000$.

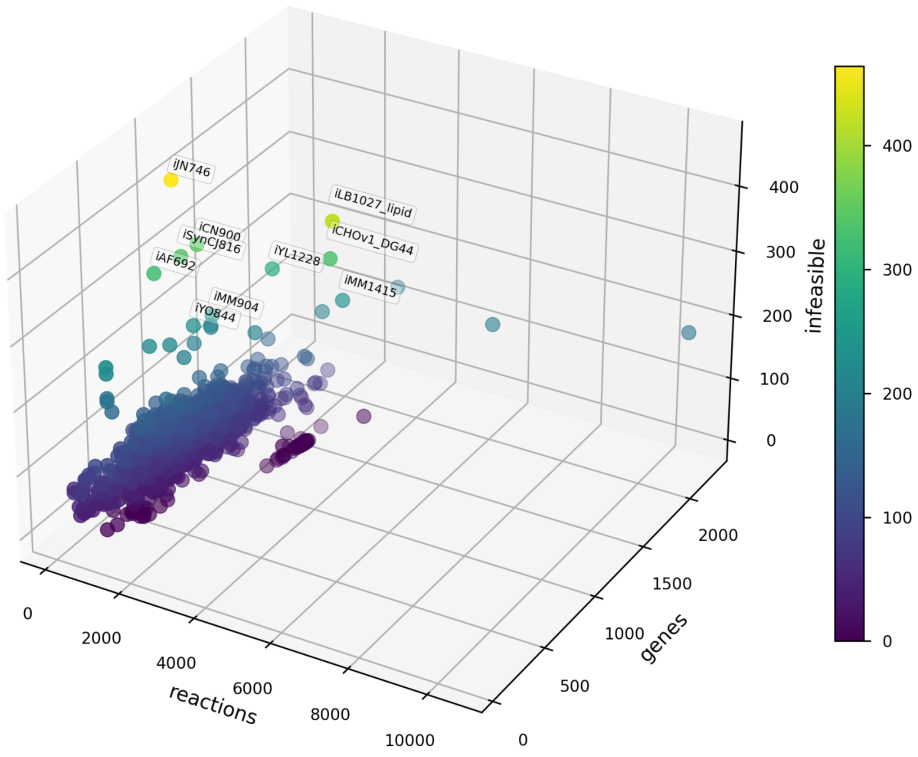

Figure 3.3 *Number of Infeasible Solutions of 1213 GEMMs w.r.t. size of the network (1000 samples)*

Based on our above plot, we observe the following:

- A tiny concentrated purple cluster towards the right of the large blob indicates the different strains of annotated *E. coli* by generating infeasible solutions through these strains. Thus, we conclude that random

mutations/constraint-limitations generated by our Monte Carlo simulations that cause lethal damage to organisms are generally consistent.

- Also, there is an inverse correlation between the number of infeasible solutions and the size of the network (particularly with respect to the number of reactions annotated within the model). This indicates that synthetic lethal damages to the network or performing any form of constraint restrictions limits the organism to opt for alternate pathways in order to survive. Furthermore, we observe that larger metabolic networks contain a large number of non-essential genes and reactions (since the number of infeasible solutions generated is lower even with a consistent number of random reaction-gene-knockouts or limiting flux boundaries).
- *Chinese Hamster* (*iCHOv1_DG44*), *Common House Mouse* (*iMM1415*) and the *Phaeodactylum tricornutum* (*iLb1027_lipid*) hold low survival rates even with sufficient mutations and a large annotated model.

In the next section, we consider the idea of dimensionality reduction using reactome minimization of the metabolic network.

3.5 Reactome Minimization

Dimensionality Reduction is the idea of reducing the number of dimensions within a dataset to reduce complexity, for instance by converting n-dimensional vectors into m-dimensional vectors. In Genome-Scale Metabolic Models, dimensionality refers to the size of the metabolic network in terms of the number of metabolites, reactions, and genes annotated within it.

The presence of a considerable amount of functional redundancy has been well-studied in the case of GEMMs. For instance, about 37-47% of reactions within the

metabolic networks of *E. coli* and *yeast* can be removed without hindering the organism's growth rate under any environment [72]. Even after multiple genetic variations, evolution has helped many organisms to consider alternative pathways in order to sustain lethal damage and maintain survivability [73] [74]. Almaas et. al. identified the existence of a metabolic core consisting of active reactions that have highly correlated flux variations irrespective of the growth conditions within the model [75].

Burgard et. al. were the first to identify a method to find the *minimal reactome* in an *E. coli* metabolic network [76] using a MILP approach**.** In addition, there are other approaches that take advantage of the characteristics of the underlying structure of the GEMM network. For instance, Jonnalagadda et. al used a graph-theory + MILP-based approach to reduce an organism's GEMM to a minimal required metabolism [77]. Recently, Lugar et. al was capable of utilizing matrix manipulation of the sparse stoichiometric matrix as an alternative to find a minimal reactome [78].

For our approach, we utilize the MinREACT algorithm offered by Sambamoorthy et al. [79] and study it in detail with respect to our pipeline**.** Unlike other approaches, the MinREACT algorithm leverages the very structure of the network by analyzing the reaction classes identified by performing a parsimonious FBA [80]. The MinREACT algorithm also results in a smaller reactome of the input metabolic network as compared to other prior approaches.

As an example, we perform the MinREACT algorithm on 50 GEMMs chosen randomly from each model repository. Figure 3.4 visualizes a plot between the size (metabolites and reactions) of 50 GEMM networks over the time it takes to produce a minimal reactome.

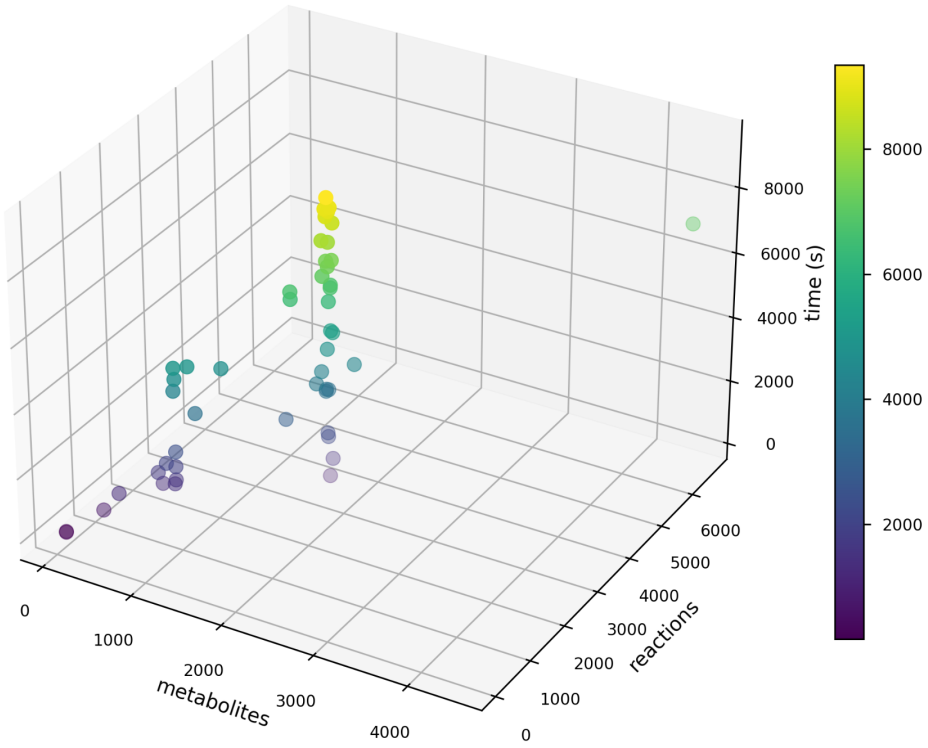

Figure 3.4 *Minimal Reactome Generation (using MinREACT) of 50 GEMMs over Time (seconds)*

Evidently, the network size increases the overall execution time of the MinREACT algorithm exponentially, directly correlating to the time it takes to achieve 'single lethals' (single reaction deletions) that ought to be discarded from the network. There also exists a metabolic core for almost all tested GEMMs that contains up to 1000 reactions within the compressed network irrespective of the size of the original input graph. Figure 3.5 confirms the above statement by depicting the size of the input metabolic network (metabolites and reactions) with respect to the number of reactions in its minimal reactome.

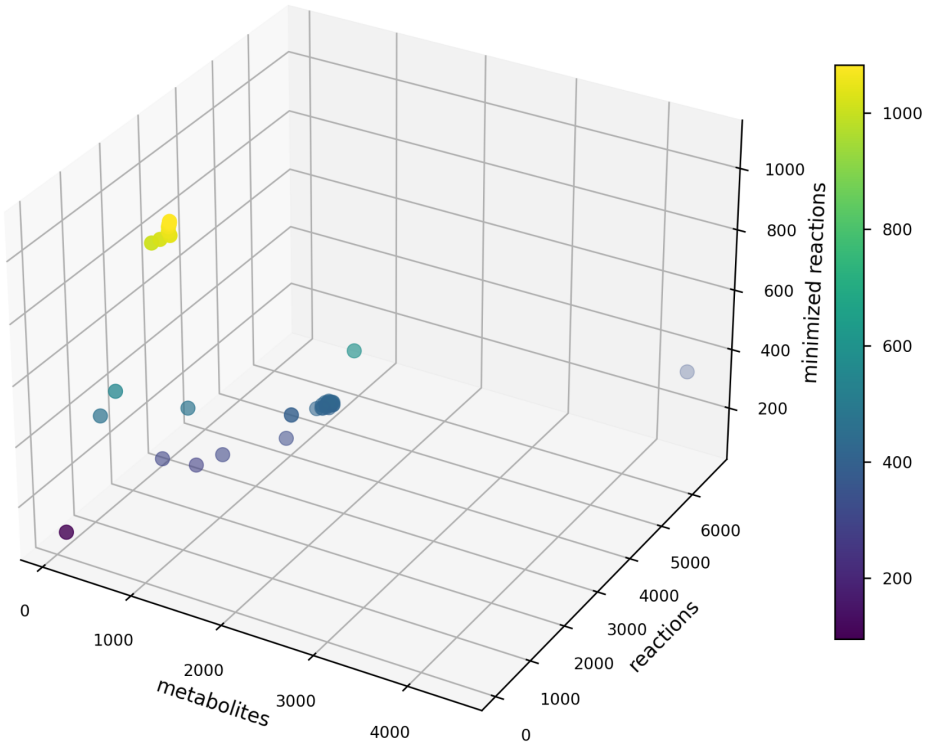

Figure 3.5 *Number of reactions within the minimized network w.r.t. size of original network (50 GEMs)*

Chapter 4

Results and Analysis

4.1 Outline

The experiments conducted for our results and analysis evaluate across multiple dimensions to answer many research-related questions. For our first dimension, we consider a comparative study of different models chosen based on time, accuracy, and feasibility. The next dimension considers a breadth-based analysis of the composition and nature of our input dataset based on our model choice. Finally, the third dimension caters to the effects of various hyperparameters introduced into our workflow pipeline. As mentioned, we consider a comparative evaluation across 3 well-annotated GEMMs namely - *e_coli_core (E. coli)*, *iIS312_Amastigote (Trypanosoma cruzi)*, and *iAB_RBC_283 (Homo sapiens).*

4.2 Configuration

To ensure unbiased comparison, we consider universally predefined hyperparameters across all our learning models. The Adam optimizer [81] was our primary choice for performing gradient-based optimization with a constant learning

rate of 0.001. For all our deep networks, we place a dropout layer after each dense hidden layer with a penalty rate of 0.2 (i.e., we penalize the weights of 20% of neurons in the preceding layer during training). This is to ensure that our models attempt to generalize well and avoid suffering from overfitting our synthetic dataset [82]. To ensure that the model achieves a neat yet diverse overview of the solution space, we consider a K-Fold cross-validation of 20% of the training set per epoch. We also consider performing an early stopping during training when the R^2 (coefficient of determination) of our validation set does not change after 10 epochs (training iteration) by a factor of 0.01; 50-100 epochs otherwise. We generate a dataset of over 1 million observations for each of our organisms and split each dataset into 80%-20% training-testing sets respectively. Prior to training, we normalize both the reaction bounds and the flux rates within the range [0, 1] as a preprocessing step. To avoid overfitting, we consider feeding our learning models a batch of our dataset (512 samples) during a forward feed.

In the case of our global states, we use *MAE (Mean Absolute Error)* as our global loss function across all gradient-based learning models (since MAE is a lot less sensitive towards outliers than estimating the Mean-Squared Errors). The MAE of an actual versus the predicted flux can be given as follows:

$$MAE = \sum_{i=0}^{N} \frac{|J_i' - J_i|}{N}$$

Here, J_i' represents the predicted flux value of reaction i whereas J_i represents its actual rate in $\frac{mmol}{gDW}$. In the case of our WCGAN-GP-based model, we use the cumulative losses collected from the individual generator and discriminator losses.

4.3 A DL emulator for *Escherichia coli strain K-12 substrain MG1655 in silico*

Escherichia coli is a genus of Gram-negative, rod-shaped bacteria that are mainly found in the lower intestine of warm-blooded organisms. *E. coli* can be pathogenic and cause disease by producing Shiga toxin, causing colitis and hemolytic uremic syndrome, and by contributing to the development of hemorrhagic fever with renal syndrome. The *E. coli MG1655* has been a well-studied genome with a well-curated Genome-Scale Metabolic Model which will be used as our test organism for analyzing our pipeline [83]. To explain in depth our approach, we consider the wild-type *E. coli str. K-12 substrain MG1655* genome-scale metabolic model [84] as a case study to run through our pipeline.

4.3.1 Baseline Model

In predicting biomass growth rate within an *in silico E. coli* GEMM, we consider Linear Regression as our initial model to benchmark our performance. Our Linear Regression model can be considered a simplified version of the McCulloch-Pitts Neuron [26] model with no activation function. Figure 4.1. Illustrates the layer configuration of our Linear Regression model. Note, the `DenseBlock` layer denotes an intermediate layer containing 190 neurons and 1 unit worth of a bias neuron.

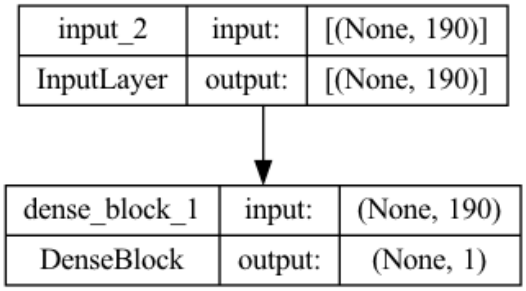

Figure 4.1 *Model Configuration of our Linear Regression Model for* `e_coli_core`

Even with its limitations, Simple Linear Regression performs well with a testing R^2 score of 0.7715 in the case when the input dataset comprises information associated with only the perturbed state of the GEMM model. We also noticed an early stopping at around 37 epochs. We achieved a validation loss of MAE of 0.0543 and an MSE of 0.0302 respectively. The average training time up to 38 epochs was 12.27 minutes, a comparatively impressive Time-to-Train (ToT) as compared to other learning-based models.

Figure 4.2 illustrates the actual versus predicted growth rates for the biomass objective in an *E. coli* GEMM specifically associated with the lethally damaged states of the model. A good reason as to why our baseline model works well in the case of predicting perturbed states is due to the fact that a GEMM model of small size generally lacks more insights into the organism's metabolic states, therefore stunting growth quicker than a well-annotated model.

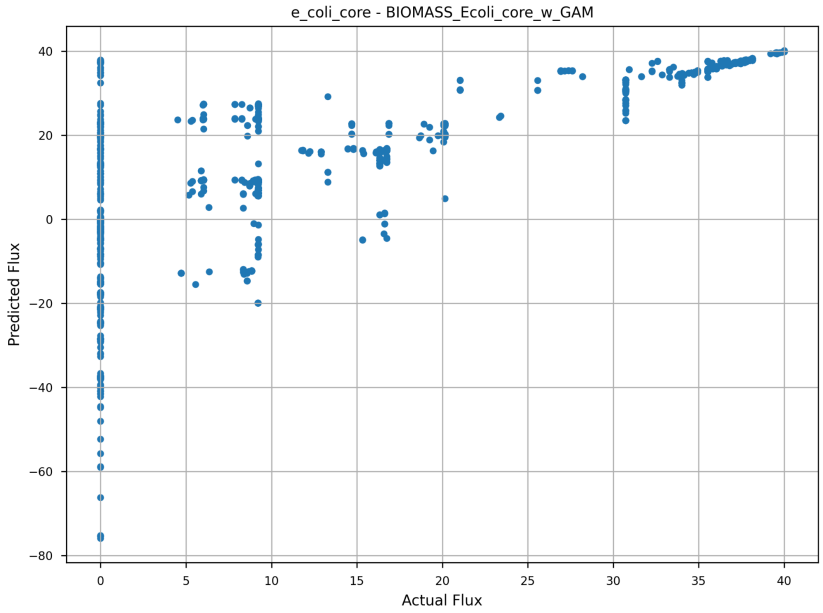

Figure 4.2 Actual versus Predicted flux rates of the biomass objective in E. coli. GEMM (perturbed, baseline)

In continuation, we consider predicting the objective based on the healthy state of the input organism. Our baseline model took up to 46 epochs with a quicker training time clocking at 8.36 minutes. However, the testing R^2 score stabilized at only 0.3795. A validation loss of 0.0485 (MAE) and 0.0159 (MSE) were achieved respectively. Figures 4.3 and 4.4 visualizes the training and validation losses and an improving R^2 score over each epoch for our healthy (with restricted nutrient uptake), perturbed, and combined model.

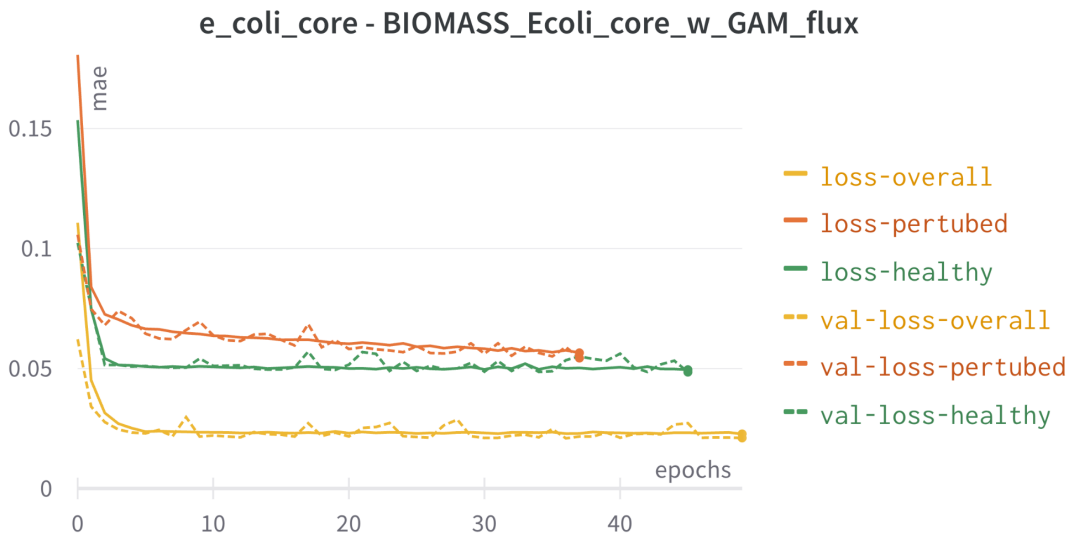

Figure 4.3 Training and Validation Losses (baseline)

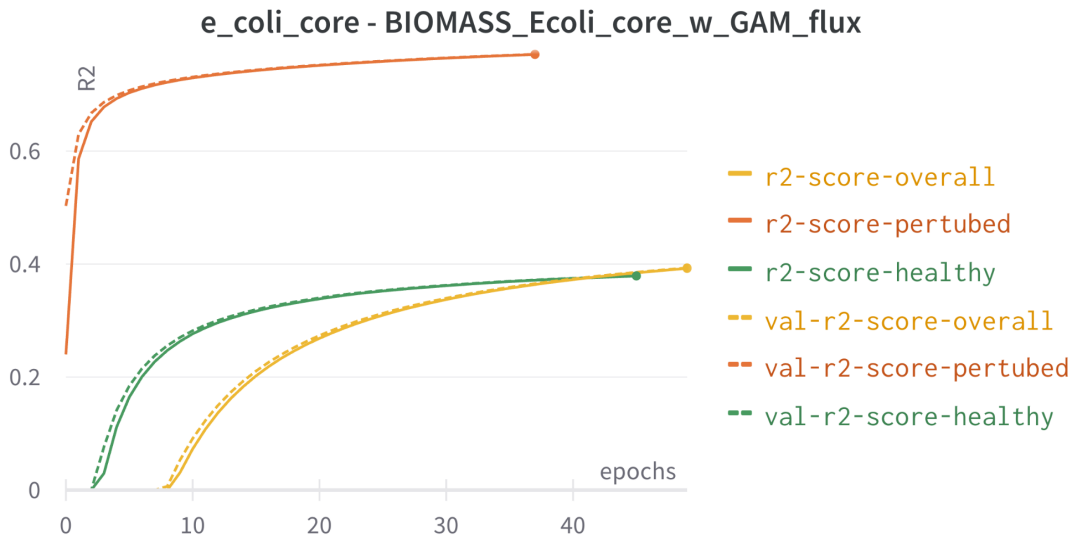

Figure 4.4 *Coefficient of Determination scores (baseline)*

In terms of the actual versus predicted growth rates for our objective pseudo-reaction, the model shows a relatively linear trend in Figure 4.5. For the sake of clarity, we visualize only 10,000 randomly chosen samples from our testing set containing 200,000 samples.

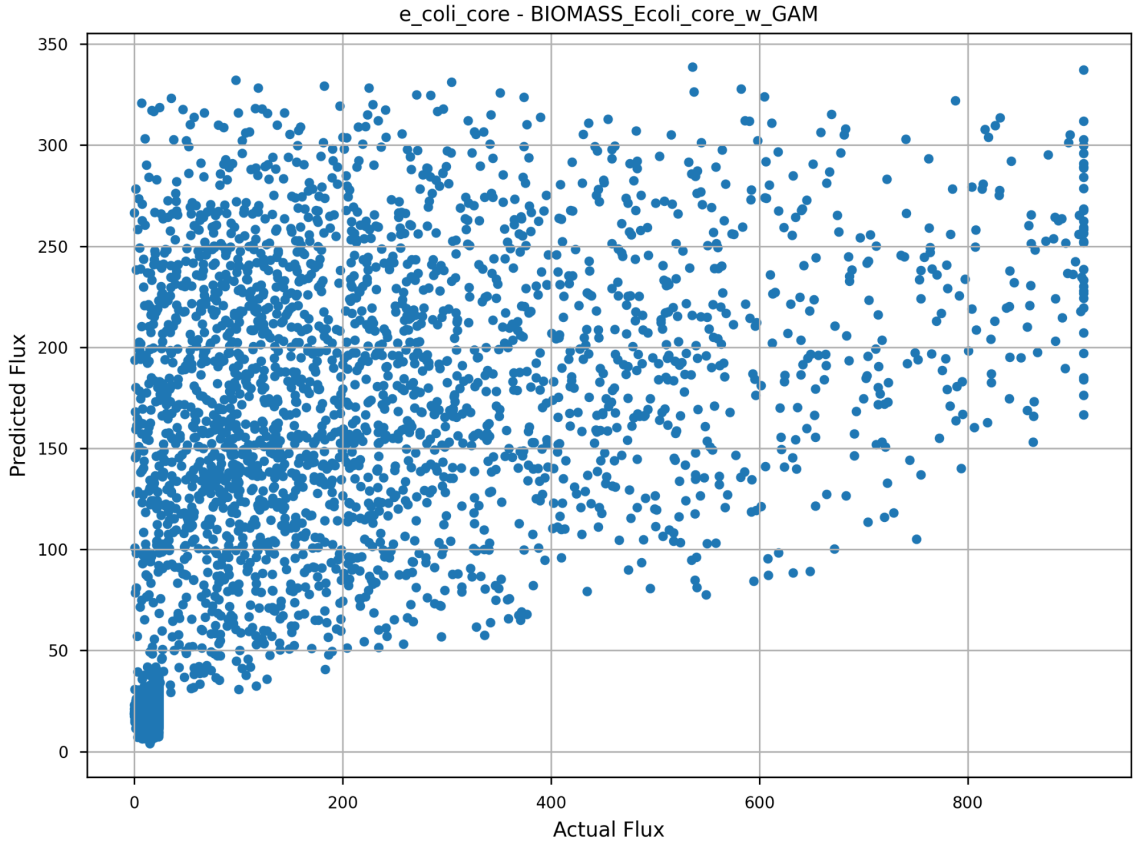

Figure 4.5 *Actual versus Predicted flux rates of the biomass objective in E. coli. GEMM (healthy, baseline)*

We combine both the healthy and perturbed states of the organism as our input dataset to check whether the lethally damaged states of the model penalize the capabilities of function approximation by our learning model. Our learning models can distinguish a healthy versus a perturbed state of the GEMM with a testing R^2 score of 0.3938. However, no early stopping was observed with the model completing within 50 epochs and with a validation loss of 0.0227 (MAE) and 0.00535 (MSE). Figure 4.6 visualizes the actual versus predicted flux in the case of the biomass pseudo reaction as our objective function.

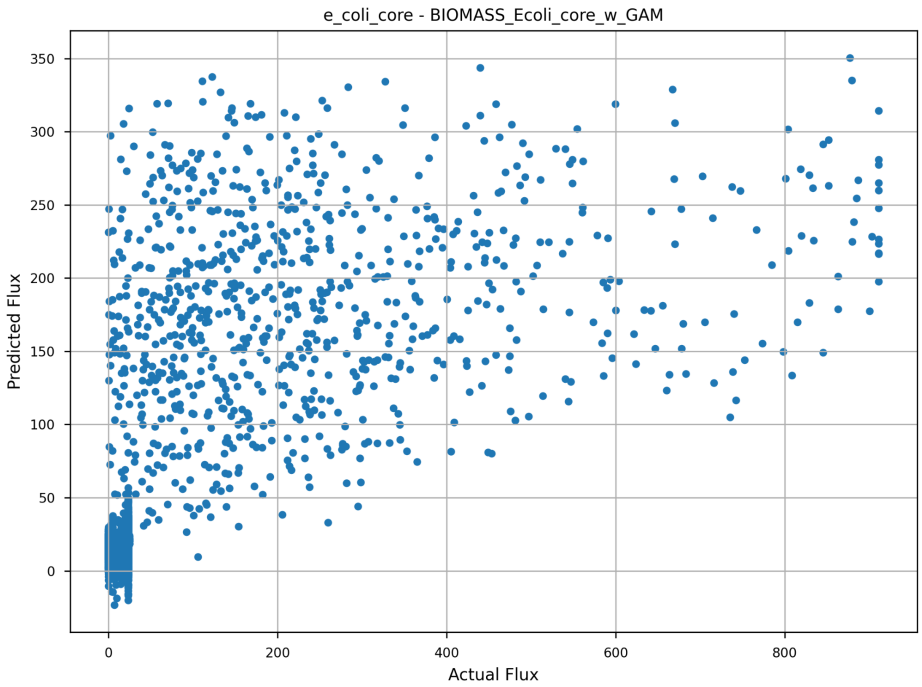

Figure 4.6 *Actual versus Predicted flux rates of the biomass objective in E. coli. GEMM (overall, baseline)*

4.3.2 WCGAN-GP Model

The WCGAN-GP-based model achieves a testing R^2 score of 0.9558 in just 14 epochs by helping predict the corresponding objective flux based on synthetic lethal reactions and gene knockouts. Clearly, our DL approach stands out learning essential from non-essential gene sets by observing the probability distribution of the flux vector cone. A validation loss of 0.0221 (GAN loss) was achieved by our model. Figures 4.7 and 4.8 visualizes the training and validation losses and improving R^2 scores over each epoch for our WCGAN-GP-based approach. Both, training and validation losses for each of our models drop significantly within 10 epochs worth of training with a gradual and smooth increase in its coefficient of determination. However, our DL approach tends to be slower in terms of training time as compared to our baseline model by an average factor of 3 clocking at 30 minutes (average) until halt.

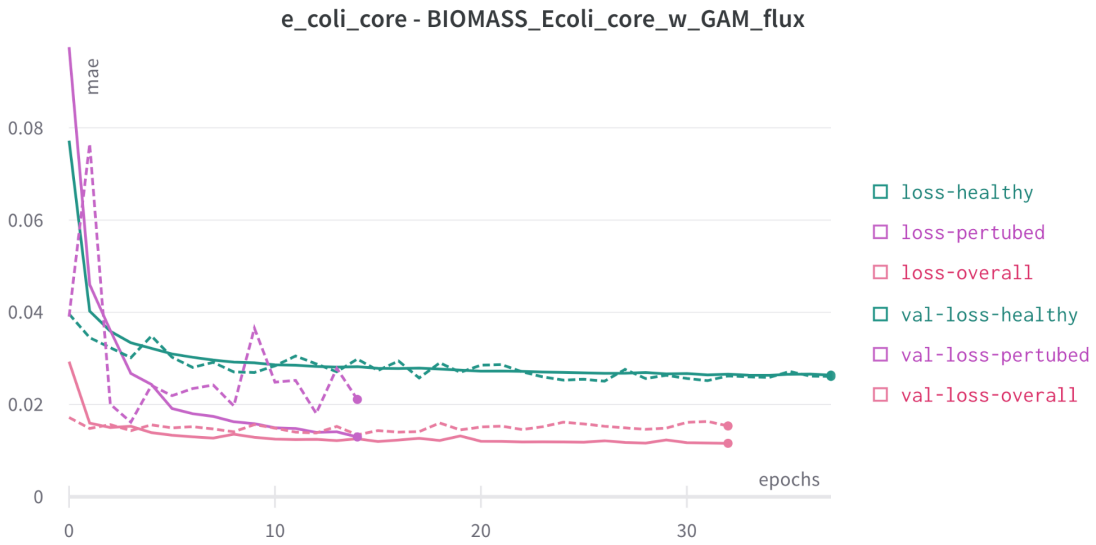

Figure 4.7 *Training and Validation Losses (WCGAN-GP)*

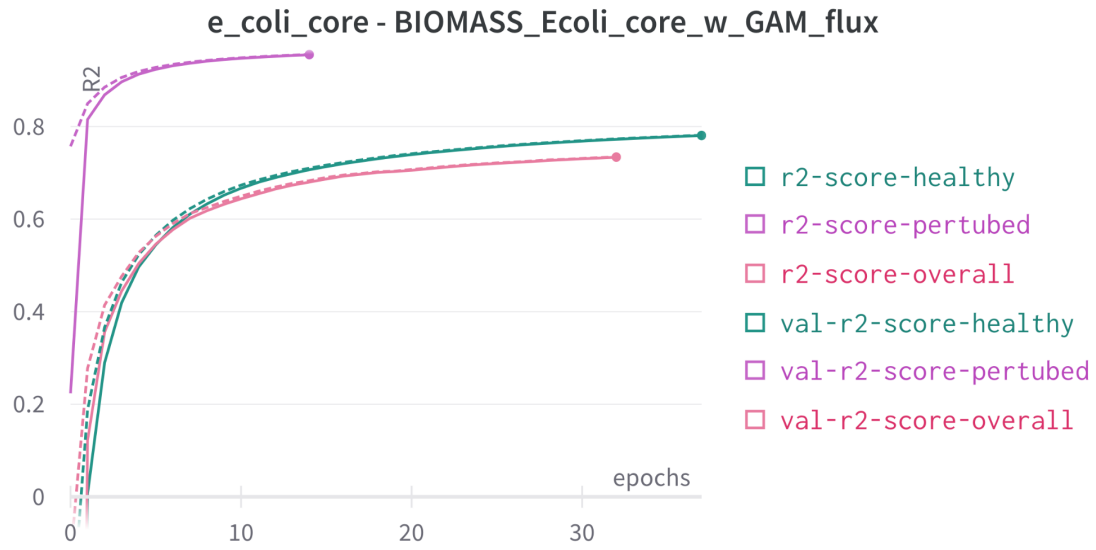

Figure 4.8 *Coefficient of Determination scores (WCGAN-GP)*

On combining both the healthy and perturbed states of the organism as our input dataset, our DL model achieves an R^2 score of 0.7335 on the testing set and a validation loss of 0.01536 halting at 32 epochs. Figure 4.9 represents the best fit estimation of predicted fluxes of $v_{objective}$ of the *E. coli* GEMM.

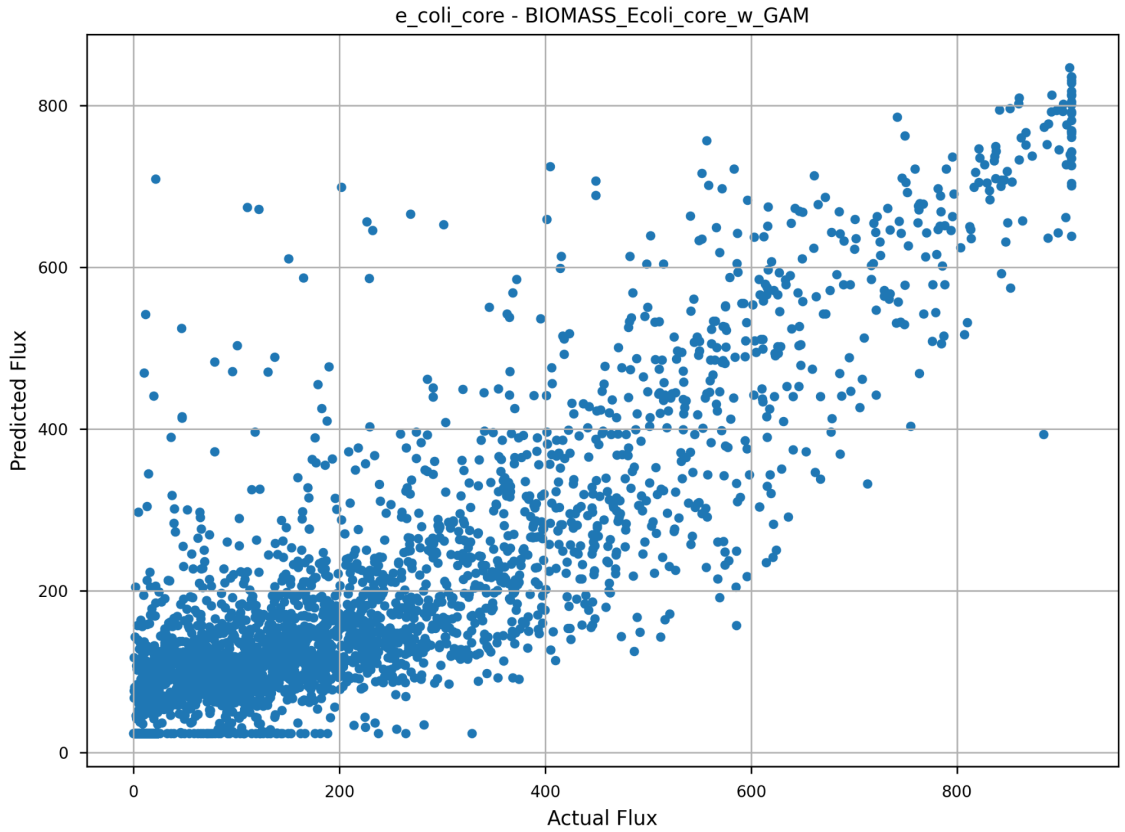

Figure 4.9 *Actual versus Predicted flux rates of the biomass objective in E. coli. GEMM (healthy, WCGAN-GP)*

Evidently, the plot follows a higher degree of linearity as compared to our baseline model. A noteworthy limitation within FBA's SIMPLEX approach is that it is generally hard to estimate whether a combination of given constraints leads to multi-optimal states of growth within the model and if so, the values of all possible sets of values of such states. This also holds true in the case of our DL-based pipeline when unaware of whether a prediction is part of a multi-optimal solution set. Figure 4.10 represents the best fit estimation of predicted fluxes of $v_{objective}$ of the *E. coli* GEMM for our complete input dataset. This relationship portrays the capability of our DL model to approximate biologically relevant growth estimates based on different and distinct

distributions of the convex polytope solution space. It offers a linear relationship trend between the actual versus predicted flux values irrespective of the model state.

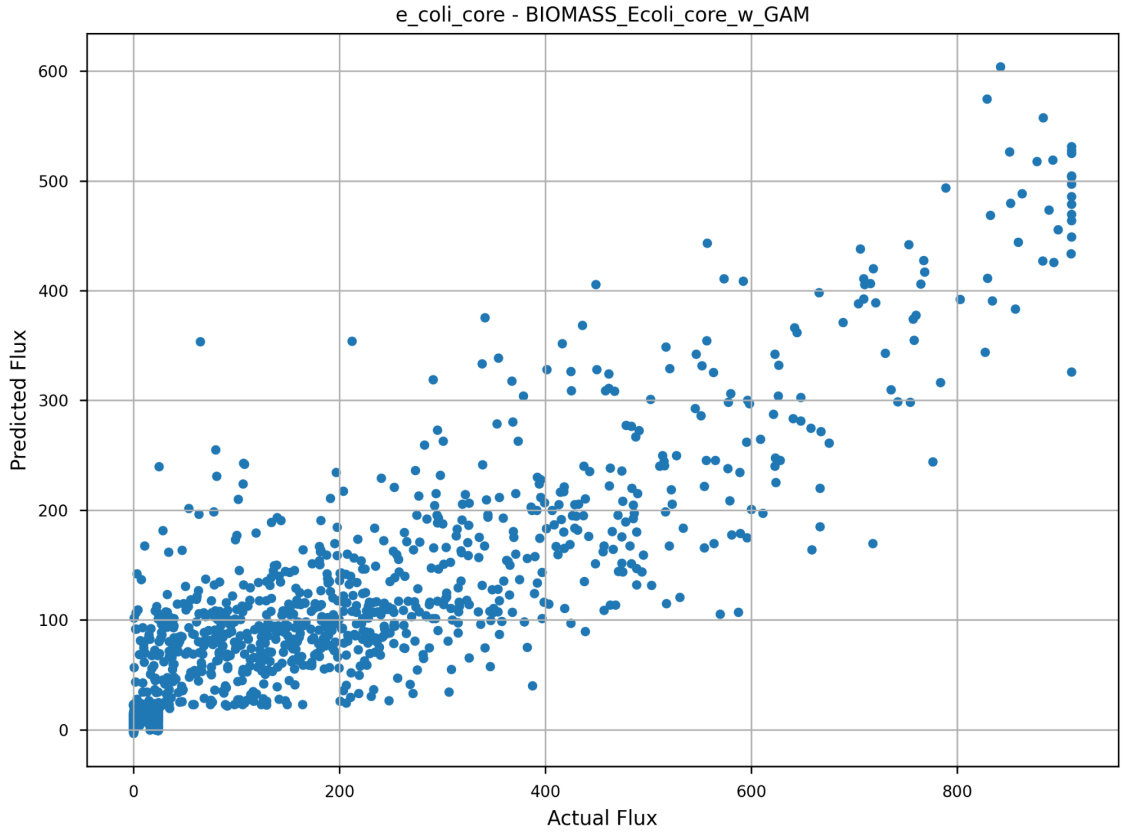

Figure 4.10 Actual versus Predicted flux rates of the biomass objective in E. coli. GEMM (overall, WCGAN-GP)

4.4 DL emulators for *Trypanosoma cruzi* and *Homo sapiens in silico*

For general-purpose use-cases, we consider two additional well-curated GEMMs *iIS312_Amastigote* and *iAB_RBC_283* and evaluate each of them with respect to our DL-based pipeline.

Trypanosoma cruzi is a parasite that causes Chagas disease. The parasite lives in the blood, heart, and digestive tract of an infected person. It is transmitted by the bite of an infected bug called a "kissing bug". The most common symptoms of Chagas disease are fever, fatigue, body aches, or swelling around the bite wound. The infection can cause serious complications such as heart disease and intestinal inflammation that can lead to death. Interestingly, this parasite is also a well-studied organism in the sphere of metabolic engineering [85] [86] [87].

We use a similar configuration of the DL pipeline as described in Section 4.2. However, the metabolic network is significantly larger as compared to our previous study with 609 metabolites and 519 reactions. To reduce its dimensionality, we perform a reactome minimization of the network by knocking out the set of reactions that are not part of the minimized network. On performing the MinREACT algorithm over *iIS312_Amastigote*, we get a minimized reactome containing 89 biologically relevant reactions. Our input data feed is then a subset corresponding to this minimized reaction set of the network. To improve training time, we use transfer learning [88] by using trained weights on the diseased state with other types. This encourages a quicker learning curve since adjustments to the hyperspace of the convex polytope has been viewed before by the learning model. Unlike *E. coli.,* the parasitic GEMM shows an inverse trend with respect to learning our input dataset.

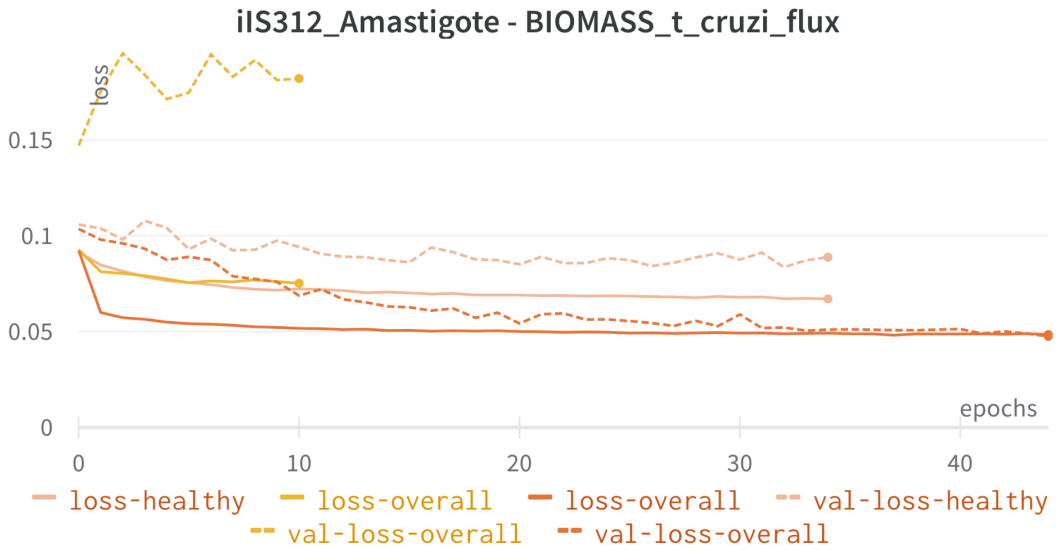

Figure 4.11 *Training and Validation Losses (iIS312_Amastigote, WCGAN-GP)*

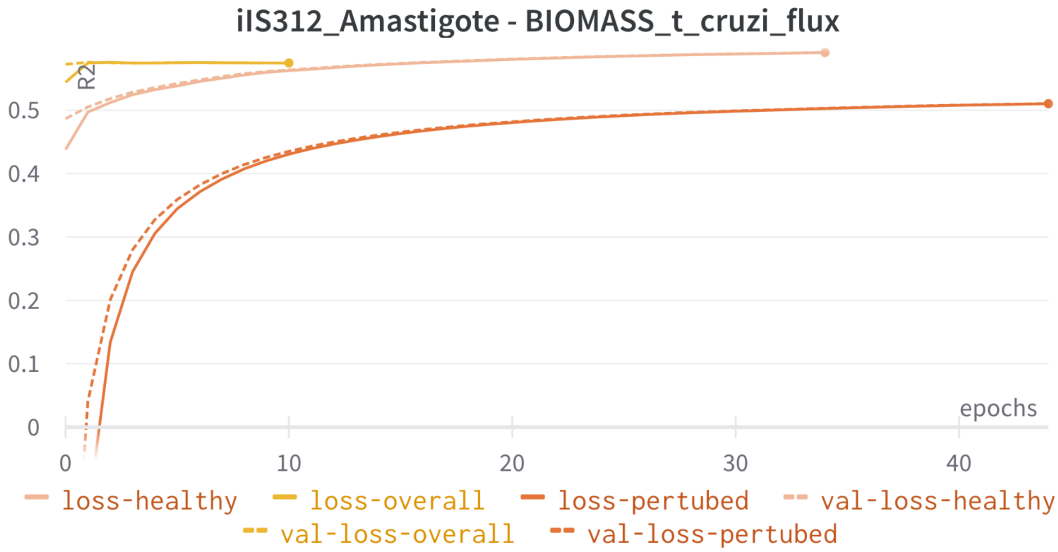

Figure 4.12 *Coefficient of Determination scores (iIS312_Amastigote, WCGAN-GP)*

Figures 4.11 and 4.12 visualizes the training and validation losses and improving R^2 scores over each epoch for our WCGAN-GP-based approach in the case of *iIS312_Amastigote* genome. We achieve a relative R^2 score of approximately 0.6 for all our data types at an average of 40 epochs. Interestingly, our diseased model states perform poorer as compared to a normal counterpart. Another interesting thing to

notice is that our learning models are capable of estimating the flux cone's outer boundaries. Figure 4.13 represents the best fit estimation of predicted fluxes of $v_{objective}$ of *Trypanosoma cruzi*.

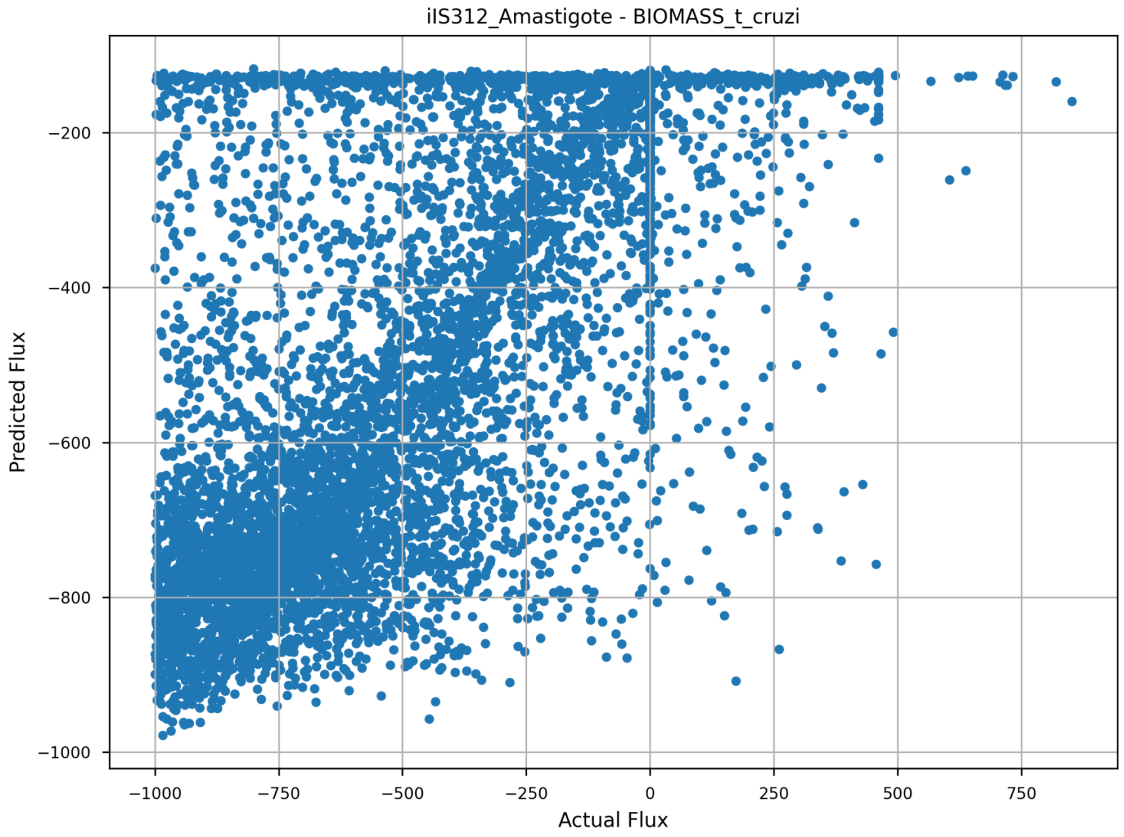

Figure 4.13 *Actual versus Predicted flux rates of the biomass objective iIS312_Amastigote in GEMM (overall, WCGAN-GP)*

In the case of our *Homo sapiens (iAB_RBC_283)* model, we minimize the reactome from 342 metabolites and 469 reactions to a metabolic core consisting of 71 reactions. Generally, the target objective in the case of humans (primarily in the case of animal cell neurons) is the transportation of Na+/K+-ATPase across the cell membrane. Figures 4.14. and 4.15. visualizes the training and validation losses and improving R^2 scores over each epoch for our WCGAN-GP-based approach in the case of *iAB_RBC_283* GEMM.

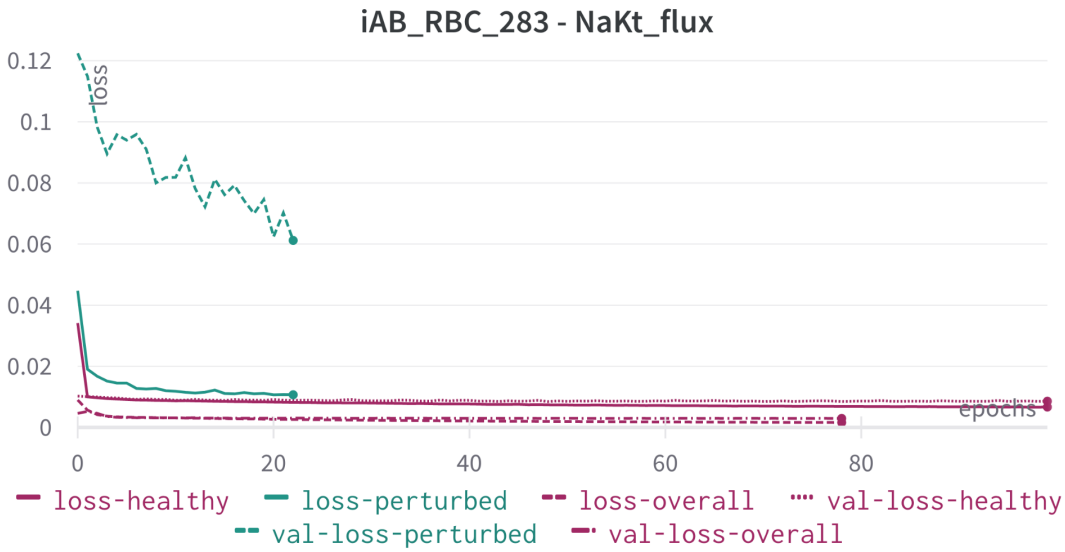

Figure 4.14 *Training and Validation Losses (iAB_RBC_283, WCGAN-GP)*

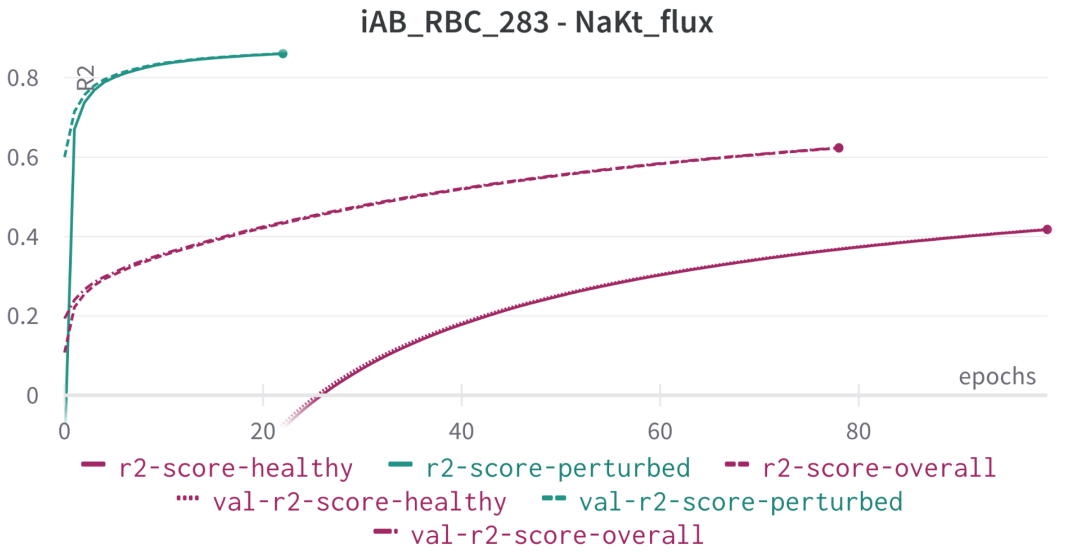

Figure 4.15 *Coefficient of Determination scores (iAB_RBC_283, WCGAN-GP)*

Impressively, *iAB_RBC_283* achieves with a testing R^2 score of 0.8615 for perturbed-based predictions. This is promising since the parameterized model is capable of learning the gene lethality of the input organism. On combining both the

healthy and perturbed states of the organism as our input dataset, our DL model achieves an R^2 score of 0.6240 on the testing set and a validation loss of 0.0295 halting at 78 epochs. Figure 4.16 represents the best fit estimation of predicted fluxes of $v_{objective}$ for the *Homo sapiens* GEMM.

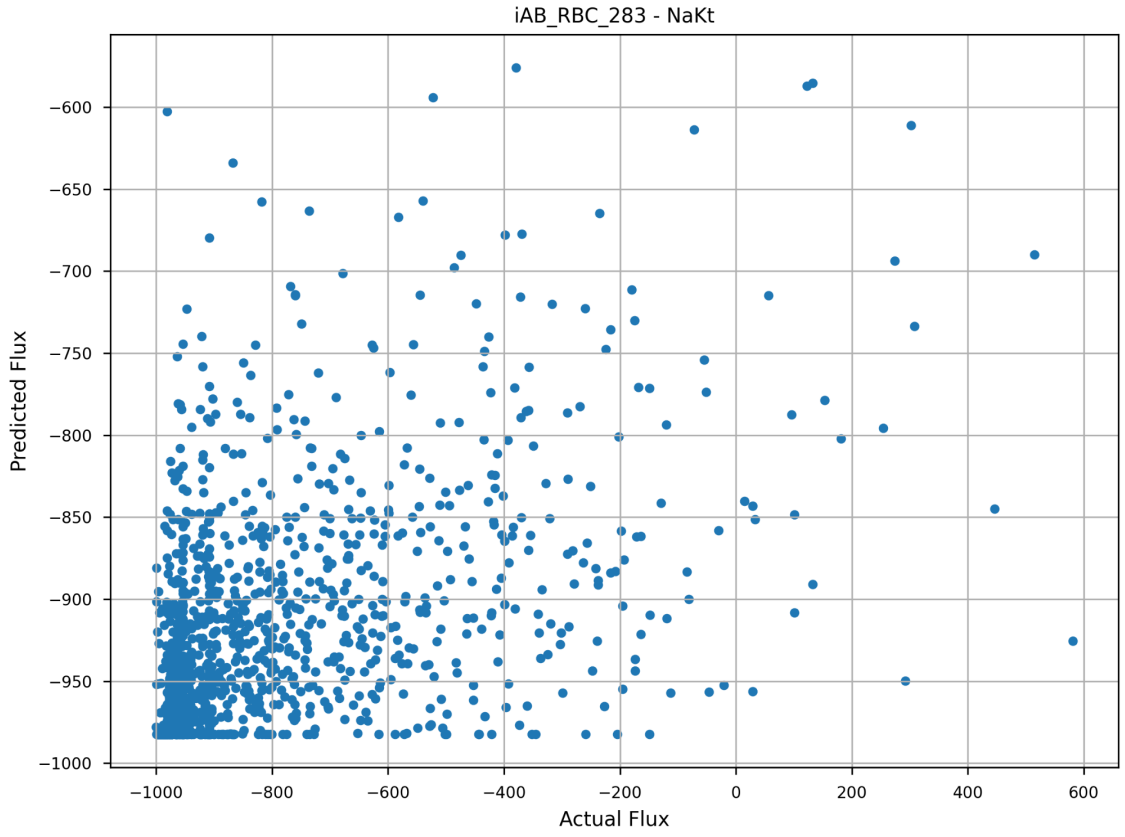

Figure 4.16 *Actual versus Predicted flux rates of iAB_RBC_283 - Na+/K+ ATPase*

4.5 Computational Efficiency

While the time taken to train a surrogate GEMM is certainly incomparable with respect to current traditional frameworks, we consider the time it takes to alter/mutate and compute its objective. Figure 4.17 explores this relationship with a benchmark graph. We use the open-source MILP solver - GLPK as our default solver to compute the objective flux of an e_coli_core model. The benchmark was performed on a Macbook Air 2021 (M1) on a single thread.

GLPK works comparatively better than our DeepGEMM framework with up to 100 alterations but, however, shows a lack of scalability as the number of alterations/mutations increase and diverge over time. Meanwhile, our DeepGEMM framework exhibits flux computations in almost constant time irrespective of the number of models required to compute. This is due to the fact that models, once trained, can efficiently compute solutions for multiple instances irrespective of the number of instances itself. This significantly boosts the processing time it takes to compute multiple instances of model mutations simultaneously.

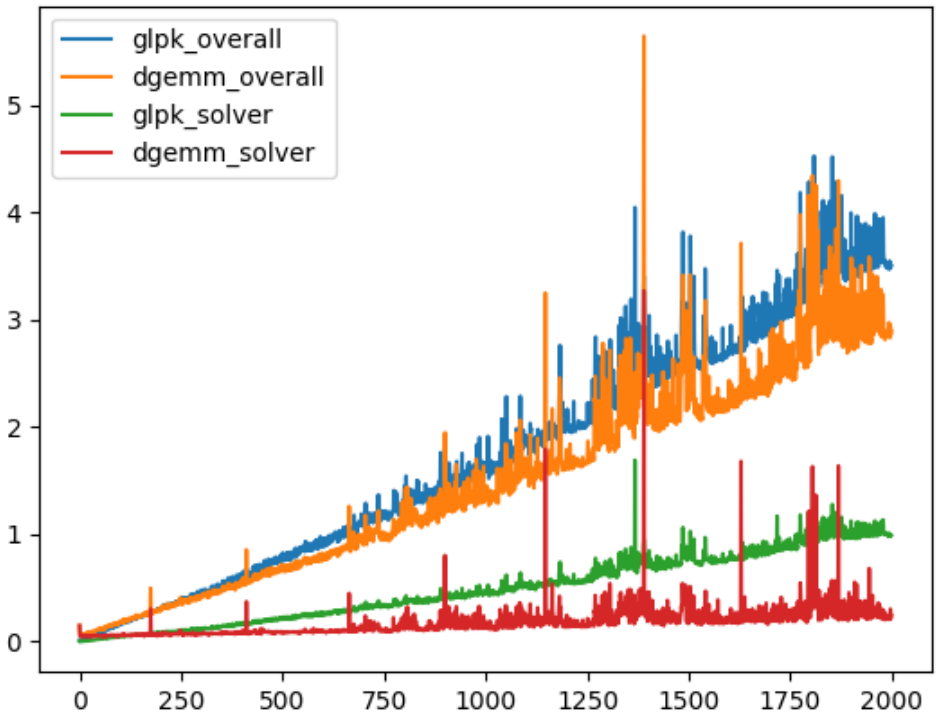

Figure 4.17 *number of alterations/mutations versus time to compute flux.*

Chapter 5

Summary and Future Work

Unwinding the facts within the genotype-phenotype relationship helps us get closer to understanding what makes biological systems behave the way they do. Genome-Scale Metabolic Models (GEMMs) offer us a snapshot of the metabolic activity within the biochemical pathways of a given organism under certain constraints. Accurately predicting the growth of wild-type reactions subjected to environmental constraints purely *in silico* is an improvement process in the field of Systems Biology. In this thesis, we used a Wasserstein Conditional GAN *(with Gradient Penalty)* (WCGAN-GP) to help learn and generate flux distributions of interacting reactions within metabolic pathways by using annotated Genome-Scale Metabolic Models and purely synthetic data generated using Monte-Carlo-based simulations and Flux Balance Analysis (FBA). First, we trained our WCGAN-GP using synthetically generated data from 3 GEMMs representing different organisms.

In this thesis, we have limited our research to certain assumptions. First, wild-type growth in the real biological world is directly correlated with respect to time. FBA, in its purest form, is also limited to the organism being observed in steady-state alone. In our future work, we consider broadening the applications of DeepGEMM to

more modeling paradigms - namely kinetic models, PBPK (physiologically-based pharmacokinetic) models, agent-based and multi-scale models.

References

1. Schilling, Christophe H., et al. "Genome-Scale Metabolic Model of Helicobacter Pylori 26695." Journal of Bacteriology, vol. 184, no. 16, Aug. 2002, pp. 4582–93. PubMed, https://doi.org/10.1128/JB.184.16.4582-4593.2002.
2. Henry, Christopher S., et al. "High-Throughput Generation, Optimization and Analysis of Genome-Scale Metabolic Models." Nature Biotechnology, vol. 28, no. 9, Sept. 2010, pp. 977–82. www.nature.com, https://doi.org/10.1038/nbt.1672.
3. Thiele, Ines, and Bernhard Ø. Palsson. "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction." Nature Protocols, vol. 5, no. 1, 2010, pp. 93–121. PubMed Central, https://doi.org/10.1038/nprot.2009.203.
4. Terzer, Marco, et al. "Genome-scale Metabolic Networks." WIREs Systems Biology and Medicine, vol. 1, no. 3, Nov. 2009, pp. 285–97. DOI.org (Crossref), https://doi.org/10.1002/wsbm.37.
5. Orth, Jeffrey D., et al. "What Is Flux Balance Analysis?" Nature Biotechnology, vol. 28, no. 3, Mar. 2010, pp. 245–48. www.nature.com, https://doi.org/10.1038/nbt.1614.
6. Raškevičius, Vytautas, et al. "Genome-Scale Metabolic Models as Tools for Drug Design and Personalized Medicine." PLoS ONE, vol. 13, no. 1, Jan. 2018, p. e0190636. PubMed Central, https://doi.org/10.1371/journal.pone.0190636.

7. Kim, Tae Yong, et al. “Recent Advances in Reconstruction and Applications of Genome-Scale Metabolic Models.” Current Opinion in Biotechnology, vol. 23, no. 4, Aug. 2012, pp. 617–23. ScienceDirect, https://doi.org/10.1016/j.copbio.2011.10.007.
8. Karthik Raman, Nagasuma Chandra, Flux balance analysis of biological systems: applications and challenges, Briefings in Bioinformatics, Volume 10, Issue 4, July 2009, Pages 435–449, https://doi.org/10.1093/bib/bbp011
9. Edwards, Jeremy S., and Bernhard O. Palsson. “Systems Properties of the Haemophilus InfluenzaeRd Metabolic Genotype *.” Journal of Biological Chemistry, vol. 274, no. 25, June 1999, pp. 17410–16. www.jbc.org, https://doi.org/10.1074/jbc.274.25.17410.
10. Gu, Changdai, et al. “Current Status and Applications of Genome-Scale Metabolic Models.” Genome Biology, vol. 20, no. 1, June 2019, p. 121. Springer Link, https://doi.org/10.1186/s13059-019-1730-3.
11. Väremo, Leif, et al. “Novel Insights into Obesity and Diabetes through Genome-Scale Metabolic Modeling.” Frontiers in Physiology, vol. 4, Apr. 2013, p. 92. PubMed Central, https://doi.org/10.3389/fphys.2013.00092.
12. Alberto Noronha, Jennifer Modamio, Yohan Jarosz, Elisabeth Guerard, Nicolas Sompairac, German Preciat, Anna Dröfn Daníelsdóttir, Max Krecke, Diane Merten, Hulda S Haraldsdóttir, Almut Heinken, Laurent Heirendt, Stefanía Magnúsdóttir, Dmitry A Ravcheev, Swagatika Sahoo, Piotr Gawron, Lucia Friscioni, Beatriz Garcia, Mabel Prendergast, Alberto Puente, Mariana Rodrigues, Akansha Roy, Mouss Rouquaya, Luca Wiltgen, Alise Žagare, Elisabeth John, Maren Krueger, Inna Kuperstein, Andrei Zinovyev, Reinhard Schneider, Ronan M T Fleming, Ines Thiele, The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D614–D624, https://doi.org/10.1093/nar/gky992

13. Edwards, Jeremy S., and Bernhard O. Palsson. "Metabolic Flux Balance Analysis and the in Silico Analysis of Escherichia Coli K-12 Gene Deletions." BMC Bioinformatics, vol. 1, no. 1, July 2000, p. 1. BioMed Central, https://doi.org/10.1186/1471-2105-1-1.
14. Lularevic, Maximilian, et al. "Improving the Accuracy of Flux Balance Analysis through the Implementation of Carbon Availability Constraints for Intracellular Reactions." Biotechnology and Bioengineering, vol. 116, no. 9, Sept. 2019, pp. 2339–52. DOI.org (Crossref), https://doi.org/10.1002/bit.27025.
15. Hoppe, Andreas, et al. "Including Metabolite Concentrations into Flux Balance Analysis: Thermodynamic Realizability as a Constraint on Flux Distributions in Metabolic Networks." BMC Systems Biology, vol. 1, no. 1, June 2007, p. 23. Springer Link, https://doi.org/10.1186/1752-0509-1-23.
16. Laubenbacher, R., et al. "Building Digital Twins of the Human Immune System: Toward a Roadmap." Npj Digital Medicine, vol. 5, no. 1, May 2022, pp. 1–5. www.nature.com, https://doi.org/10.1038/s41746-022-00610-z.
17. Jordan, M. I., and T. M. Mitchell. "Machine Learning: Trends, Perspectives, and Prospects." *Science*, vol. 349, no. 6245, July 2015, pp. 255–60. *DOI.org (Crossref)*, https://doi.org/10.1126/science.aaa8415.
18. Sridhara, Viswanadham, et al. "Predicting Growth Conditions from Internal Metabolic Fluxes in an In-Silico Model of E. Coli." PLOS ONE, vol. 9, no. 12, Dec. 2014, p. e114608. PLoS Journals, https://doi.org/10.1371/journal.pone.0114608.
19. Giuseppe Magazzù, Guido Zampieri, Claudio Angione, Multimodal regularized linear models with flux balance analysis for mechanistic integration of omics data, *Bioinformatics*, Volume 37, Issue 20, 15 October 2021, Pages 3546–3552, https://doi.org/10.1093/bioinformatics/btab324
20. Wu, Stephen Gang, et al. "Rapid Prediction of Bacterial Heterotrophic Fluxomics Using Machine Learning and Constraint Programming." PLOS Computational

Biology, vol. 12, no. 4, Apr. 2016, p. e1004838. PLoS Journals, https://doi.org/10.1371/journal.pcbi.1004838.

21. Folch-Fortuny, A., et al. "MCR-ALS on Metabolic Networks: Obtaining More Meaningful Pathways." Chemometrics and Intelligent Laboratory Systems, vol. 142, Mar. 2015, pp. 293–303. ScienceDirect, https://doi.org/10.1016/j.chemolab.2014.10.004.
22. Szappanos, Balázs, et al. "An Integrated Approach to Characterize Genetic Interaction Networks in Yeast Metabolism." Nature Genetics, vol. 43, no. 7, July 2011, pp. 656–62. www.nature.com, https://doi.org/10.1038/ng.846.
23. LeCun, Yann, et al. "Deep Learning." Nature, vol. 521, no. 7553, May 2015, pp. 436–44. www.nature.com, https://doi.org/10.1038/nature14539.
24. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27-48.
25. Vijayakumar, Supreeta, et al. "A Hybrid Flux Balance Analysis and Machine Learning Pipeline Elucidates Metabolic Adaptation in Cyanobacteria." IScience, vol. 23, no. 12, Dec. 2020, p. 101818. ScienceDirect, https://doi.org/10.1016/j.isci.2020.101818.
26. McCulloch, Warren S., and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." The Bulletin of Mathematical Biophysics, vol. 5, no. 4, Dec. 1943, pp. 115–33. Springer Link, https://doi.org/10.1007/BF02478259.
27. Deng, Li. "A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning." APSIPA Transactions on Signal and Information Processing, vol. 3, ed 2014, p. e2. Cambridge University Press, https://doi.org/10.1017/atsip.2013.9.
28. Tang, Binhua, et al. "Recent Advances of Deep Learning in Bioinformatics and Computational Biology." Frontiers in Genetics, vol. 10, 2019. Frontiers, https://www.frontiersin.org/articles/10.3389/fgene.2019.00214.

29. Bank, Dor, et al. Autoencoders. arXiv, 3 Apr. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2003.05991.
30. Inoue, Tadanobu, et al. "Transfer Learning from Synthetic to Real Images Using Variational Autoencoders for Precise Position Detection." 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 2725–29. IEEE Xplore, https://doi.org/10.1109/ICIP.2018.8451064.
31. Guo, Weihua, et al. DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing. arXiv, 8 May 2017. arXiv.org, https://doi.org/10.48550/arXiv.1705.03094.
32. Keating, Sarah M., et al. "SBML Level 3: An Extensible Format for the Exchange and Reuse of Biological Models." Molecular Systems Biology, vol. 16, no. 8, Aug. 2020. DOI.org (Crossref), https://doi.org/10.15252/msb.20199110.
33. Kingma, Diederik P., and Max Welling. Auto-Encoding Variational Bayes. arXiv, 1 May 2014. arXiv.org, https://doi.org/10.48550/arXiv.1312.6114.
34. Barsacchi, Marco, et al. GEESE: Metabolically Driven Latent Space Learning for Gene Expression Data. bioRxiv, 11 July 2018. bioRxiv, https://doi.org/10.1101/365643.
35. Fournier-Viger, Philippe. Too Many Machine Learning Papers? | The Data Mining Blog. Accessed 10 Oct. 2022.
36. Gulrajani, Ishaan, et al. Improved Training of Wasserstein GANs. arXiv, 25 Dec. 2017. arXiv.org, https://doi.org/10.48550/arXiv.1704.00028.
37. Aggarwal, Karan, Matthieu Kirchmeyer, Pranjul Yadav, S. Sathiya Keerthi, and Patrick Gallinari. "Regression with conditional gan." *arXiv preprint arXiv:1905.12868* (2019).
38. Maulud, Dastan, and Adnan M. Abdulazeez. "A review on linear regression comprehensive in machine learning." *Journal of Applied Science and Technology Trends* 1, no. 4 (2020): 140-147.

39. Jansen, Mascha. “FAIR Principles.” GO FAIR, https://www.go-fair.org/fair-principles/. Accessed 11 Oct. 2022.

40. Ebrahim, Ali, Joshua A. Lerman, Bernhard O. Palsson, and Daniel R. Hyduke. "COBRApy: constraints-based reconstruction and analysis for python." *BMC systems biology* 7, no. 1 (2013): 1-6.

41. Abadi, Martin, et al. “TensorFlow: A System for Large-Scale Machine Learning.” 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–83. Google Research, https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

42. Khalil, Ahmad S., and James J. Collins. “Synthetic Biology: Applications Come of Age.” Nature Reviews Genetics, vol. 11, no. 5, May 2010, pp. 367–79. www.nature.com, https://doi.org/10.1038/nrg2775.

43. Beyer, Peter, Salim Al-Babili, Xudong Ye, Paola Lucca, Patrick Schaub, Ralf Welsch, and Ingo Potrykus. "Golden rice: introducing the β-carotene biosynthesis pathway into rice endosperm by genetic engineering to defeat vitamin A deficiency." The Journal of nutrition 132, no. 3 (2002): 506S-510S.

44. Mehta, Roshni A., et al. “Engineered Polyamine Accumulation in Tomato Enhances Phytonutrient Content, Juice Quality, and Vine Life.” *Nature Biotechnology*, vol. 20, no. 6, June 2002, pp. 613–18. *PubMed*, https://doi.org/10.1038/nbt0602-613.

45. Kitano, Hiroaki. “Systems Biology: A Brief Overview.” Science, vol. 295, no. 5560, Mar. 2002, pp. 1662–64. DOI.org (Crossref), https://doi.org/10.1126/science.1069492.

46. Wang, Jeffrey H., et al. “Analytical Approaches to Metabolomics and Applications to Systems Biology.” Seminars in Nephrology, vol. 30, no. 5, Sept. 2010, pp. 500–11. ScienceDirect, https://doi.org/10.1016/j.semnephrol.2010.07.007.

47. Zhang, Cheng, and Qiang Hua. "Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine." Frontiers in Physiology, vol. 6, 2016. Frontiers, https://www.frontiersin.org/articles/10.3389/fphys.2015.00413.

48. Machado, Daniel, et al. "Fast Automated Reconstruction of Genome-Scale Metabolic Models for Microbial Species and Communities." Nucleic Acids Research, vol. 46, no. 15, Sept. 2018, pp. 7542–53. PubMed, https://doi.org/10.1093/nar/gky537.

49. Samuel M D Seaver, Filipe Liu, Qizhi Zhang, James Jeffryes, José P Faria, Janaka N Edirisinghe, Michael Mundy, Nicholas Chia, Elad Noor, Moritz E Beber, Aaron A Best, Matthew DeJongh, Jeffrey A Kimbrel, Patrik D'haeseleer, Sean R McCorkle, Jay R Bolton, Erik Pearson, Shane Canon, Elisha M Wood-Charlson, Robert W Cottingham, Adam P Arkin, Christopher S Henry, The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D575–D588, https://doi.org/10.1093/nar/gkaa746

50. Lee, Kyung Yun, et al. "The Genome-Scale Metabolic Network Analysis of Zymomonas Mobilis ZM4 Explains Physiological Features and Suggests Ethanol and Succinic Acid Production Strategies." Microbial Cell Factories, vol. 9, no. 1, Nov. 2010, p. 94. BioMed Central, https://doi.org/10.1186/1475-2859-9-94.

51. Shameer, Sanu, et al. "Flux Balance Analysis of Metabolism during Growth by Osmotic Cell Expansion and Its Application to Tomato Fruits." The Plant Journal, vol. 103, no. 1, July 2020, pp. 68–82. DOI.org (Crossref), https://doi.org/10.1111/tpj.14707.

52. Clarke, Bruce L. "Stoichiometric Network Analysis." Cell Biophysics, vol. 12, no. 1, Jan. 1988, pp. 237–53. Springer Link, https://doi.org/10.1007/BF02918360.

53. Suthers, Patrick F., et al. "A Genome-Scale Metabolic Reconstruction of Mycoplasma Genitalium, IPS189." PLoS Computational Biology, vol. 5, no. 2, Feb. 2009, p. e1000285. PubMed Central, https://doi.org/10.1371/journal.pcbi.1000285.

54. Feist, Adam M., and Bernhard O. Palsson. "The Biomass Objective Function." Current Opinion in Microbiology, vol. 13, no. 3, June 2010, pp. 344–49. ScienceDirect, https://doi.org/10.1016/j.mib.2010.03.003.

55. Heineken, F. G., et al. "On the Mathematical Status of the Pseudo-Steady State Hypothesis of Biochemical Kinetics." Mathematical Biosciences, vol. 1, no. 1, Mar. 1967, pp. 95–113. ScienceDirect, https://doi.org/10.1016/0025-5564(67)90029-6.

56. Japhalekar, Kshitija, et al. "Flux Balance Analysis for Overproduction of Organic Acids by Synechocystis Sp. PCC 6803 under Dark Anoxic Condition." Biochemical Engineering Journal, vol. 178, Jan. 2022, p. 108297. ScienceDirect, https://doi.org/10.1016/j.bej.2021.108297.

57. da Veiga Moreira, Jorgelindo, et al. "Fine-Tuning Mitochondrial Activity in Yarrowia Lipolytica for Citrate Overproduction." Scientific Reports, vol. 11, no. 1, Jan. 2021, p. 878. www.nature.com, https://doi.org/10.1038/s41598-020-79577-4.

58. Gianchandani, Erwin P., et al. "The Application of Flux Balance Analysis in Systems Biology." WIREs Systems Biology and Medicine, vol. 2, no. 3, May 2010, pp. 372–82. DOI.org (Crossref), https://doi.org/10.1002/wsbm.60.

59. Goodfellow, Ian J., et al. Generative Adversarial Networks. arXiv, 10 June 2014. arXiv.org, https://doi.org/10.48550/arXiv.1406.2661.

60. Lan, Lan, et al. "Generative Adversarial Networks and Its Applications in Biomedical Informatics." Frontiers in Public Health, vol. 8, 2020. Frontiers, https://www.frontiersin.org/articles/10.3389/fpubh.2020.00164.

61. Ghahramani, Arsham, et al. Generative Adversarial Networks Simulate Gene Expression and Predict Perturbations in Single Cells. bioRxiv, 30 July 2018. bioRxiv, https://doi.org/10.1101/262501.
62. De Cao, Nicola, and Thomas Kipf. MolGAN: An Implicit Generative Model for Small Molecular Graphs. arXiv, 27 Sept. 2022. arXiv.org, https://doi.org/10.48550/arXiv.1805.11973.
63. Arjovsky, Martin, et al. *Wasserstein GAN*. arXiv, 6 Dec. 2017. *arXiv.org*, https://doi.org/10.48550/arXiv.1701.07875.
64. Wiback, Sharon J., et al. "Monte Carlo Sampling Can Be Used to Determine the Size and Shape of the Steady-State Flux Space." Journal of Theoretical Biology, vol. 228, no. 4, June 2004, pp. 437–47. ScienceDirect, https://doi.org/10.1016/j.jtbi.2004.02.006.
65. Sarıyar, Berna, et al. "Monte Carlo Sampling and Principal Component Analysis of Flux Distributions Yield Topological and Modular Information on Metabolic Networks." Journal of Theoretical Biology, vol. 242, no. 2, Sept. 2006, pp. 389–400. ScienceDirect, https://doi.org/10.1016/j.jtbi.2006.03.007.
66. Charles J Norsigian, Neha Pusarla, John Luke McConn, James T Yurkovich, Andreas Dräger, Bernhard O Palsson, Zachary King, BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D402–D406, https://doi.org/10.1093/nar/gkz1054.
67. Rahuman S Malik-Sheriff, Mihai Glont, Tung V N Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, Emma L Fairbanks, Johannes P Meyer, Chinmay Arankalle, Thawfeek M Varusai, Vincent Knight-Schrijver, Lu Li, Corina Dueñas-Roca, Gaurhari Dass, Sarah M Keating, Young M Park, Nicola Buso, Nicolas Rodriguez, Michael Hucka, Henning

Hermjakob, BioModels—15 years of sharing computational models in life science, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D407–D415, https://doi.org/10.1093/nar/gkz1055.

68. Wang, Zheng-Yu, et al. "The Human T Cell Proliferative Response to Triple-Knockout Pig Cells in Mixed Lymphocyte Reaction." Xenotransplantation, vol. 27, no. 5, Sept. 2020, p. e12619. PubMed Central, https://doi.org/10.1111/xen.12619.
69. Goldstein, Yaron AB, and Alexander Bockmayr. "Double and Multiple Knockout Simulations for Genome-Scale Metabolic Network Reconstructions." Algorithms for Molecular Biology, vol. 10, no. 1, Jan. 2015, p. 1. BioMed Central, https://doi.org/10.1186/s13015-014-0028-y.
70. Deutscher, David, et al. "Multiple Knockout Analysis of Genetic Robustness in the Yeast Metabolic Network." Nature Genetics, vol. 38, no. 9, Sept. 2006, pp. 993–98. www.nature.com, https://doi.org/10.1038/ng1856.
71. Segrè, Daniel, et al. "Analysis of Optimality in Natural and Perturbed Metabolic Networks." Proceedings of the National Academy of Sciences, vol. 99, no. 23, Nov. 2002, pp. 15112–17. DOI.org (Crossref), https://doi.org/10.1073/pnas.232349399.
72. Wang, Zhi, and Jianzhi Zhang. "Abundant Indispensable Redundancies in Cellular Metabolic Networks." Genome Biology and Evolution, vol. 1, Apr. 2009, pp. 23–33. PubMed, https://doi.org/10.1093/gbe/evp002.
73. Behre, Jörn, et al. "Structural Robustness of Metabolic Networks with Respect to Multiple Knockouts." Journal of Theoretical Biology, vol. 252, no. 3, June 2008, pp. 433–41. PubMed, https://doi.org/10.1016/j.jtbi.2007.09.043.
74. Mahadevan, R., and D. R. Lovley. "The Degree of Redundancy in Metabolic Genes Is Linked to Mode of Metabolism." Biophysical Journal, vol. 94, no. 4, Feb. 2008, pp. 1216–20. PubMed Central, https://doi.org/10.1529/biophysj.107.118414.

75. Almaas, Eivind, et al. "The Activity Reaction Core and Plasticity of Metabolic Networks." PLOS Computational Biology, vol. 1, no. 7, Dec. 2005, p. e68. PLoS Journals, https://doi.org/10.1371/journal.pcbi.0010068.

76. Burgard, A. P., et al. "Minimal Reaction Sets for Escherichia Coli Metabolism under Different Growth Requirements and Uptake Environments." *Biotechnology Progress*, vol. 17, no. 5, Oct. 2001, pp. 791–97. *DOI.org (Crossref)*, https://doi.org/10.1021/bp0100880.

77. Jonnalagadda, Sudhakar, and Rajagopalan Srinivasan. "An Efficient Graph Theory Based Method to Identify Every Minimal Reaction Set in a Metabolic Network." BMC Systems Biology, vol. 8, no. 1, Mar. 2014, p. 28. BioMed Central, https://doi.org/10.1186/1752-0509-8-28.

78. Lugar, Daniel J., et al. "NetRed, an Algorithm to Reduce Genome-Scale Metabolic Networks and Facilitate the Analysis of Flux Predictions." Metabolic Engineering, vol. 65, May 2021, pp. 207–22. ScienceDirect, https://doi.org/10.1016/j.ymben.2020.11.003.

79. Sambamoorthy, Gayathri, and Karthik Raman. "MinReact: A Systematic Approach for Identifying Minimal Metabolic Networks." *Bioinformatics (Oxford, England)*, vol. 36, no. 15, Aug. 2020, pp. 4309–15. *PubMed*, https://doi.org/10.1093/bioinformatics/btaa497.

80. Lewis, Nathan E., et al. "Omic Data from Evolved E. Coli Are Consistent with Computed Optimal Growth from Genome-scale Models." Molecular Systems Biology, vol. 6, no. 1, Jan. 2010, p. 390. DOI.org (Crossref), https://doi.org/10.1038/msb.2010.47.

81. Kingma, Diederik P., and Jimmy Ba. Adam: A Method for Stochastic Optimization. arXiv, 29 Jan. 2017. arXiv.org, https://doi.org/10.48550/arXiv.1412.6980.

82. Srivastava, Nitish, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research*, vol. 15, no. 56, 2014, pp. 1929–58. *jmlr.org*, http://jmlr.org/papers/v15/srivastava14a.html.

83. Feist, Adam M., and Bernhard Ø. Palsson. "The Growing Scope of Applications of Genome-Scale Metabolic Reconstructions Using Escherichia Coli." *Nature Biotechnology*, vol. 26, no. 6, June 2008, pp. 659–67. *www.nature.com*, https://doi.org/10.1038/nbt1401.

84. Feist, Adam M., et al. "A Genome-Scale Metabolic Reconstruction for Escherichia Coli K-12 MG1655 That Accounts for 1260 ORFs and Thermodynamic Information." Molecular Systems Biology, vol. 3, 2007, p. 121. PubMed, https://doi.org/10.1038/msb4100155.

85. Roberts, Seth B., et al. "Proteomic and Network Analysis Characterize Stage-Specific Metabolism in Trypanosoma Cruzi." *BMC Systems Biology*, vol. 3, no. 1, May 2009, p. 52. *BioMed Central*, https://doi.org/10.1186/1752-0509-3-52.

86. Shiratsubaki, Isabel S., et al. "Genome-Scale Metabolic Models Highlight Stage-Specific Differences in Essential Metabolic Pathways in Trypanosoma Cruzi." *PLOS Neglected Tropical Diseases*, vol. 14, no. 10, Oct. 2020, p. e0008728. *PLoS Journals*, https://doi.org/10.1371/journal.pntd.0008728.

87. Dumoulin, Peter C., and Barbara A. Burleigh. "Metabolic Flexibility in Trypanosoma Cruzi Amastigotes: Implications for Persistence and Drug Sensitivity." *Current Opinion in Microbiology*, vol. 63, Oct. 2021, pp. 244–49. *ScienceDirect*, https://doi.org/10.1016/j.mib.2021.07.017.

88. Zhuang, Fuzhen, et al. *A Comprehensive Survey on Transfer Learning*. arXiv, 23 June 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.1911.02685.