Master Thesis Summary
Inferring metabolic states from omics datasets with graph neural networks

Fabio Marini 851977

1 Introduction

Understanding cellular metabolism is a fundamental challenge in computational biology. Metabolic flux estimation provides insights into how biochemical pathways operate under different conditions, yet direct measurement of intracellular fluxes remains impractical. Consequently, computational models are employed to infer metabolic activity based on available biological data. *Flux Balance Analysis* (FBA) is one of the most widely used methods for predicting metabolic flux distributions [1]. It formulates metabolism as a constrained optimization problem, leveraging stoichiometric constraints, which enforce mass balance by ensuring that the production and consumption of each metabolite at steady state sum to zero. This is expressed as $S \cdot v = 0$, where S is the stoichiometric matrix and v is the vector of reaction fluxes. Additionally, flux bounds, $LB \leq v \leq UB$, impose physiological constraints on reaction rates. A predefined objective function, such as biomass or ATP production, is optimized to compute feasible fluxes. However, FBA relies on strong assumptions, such as the steady-state hypothesis and the maximization of a single metabolic objective, which may not fully capture the complexity and heterogeneity of real metabolic systems. Furthermore, constraint-based models often struggle to incorporate high-dimensional biological data, such as gene expression profiles, in a flexible manner.

The primary objective of this work is to investigate whether deep learning approaches can provide accurate and biologically consistent predictions of metabolic fluxes. Recent advances in deep learning, particularly *Graph Neural Networks* (GNNs), offer a promising alternative for modeling structured biological data. GNNs are well suited for metabolic networks, as they can leverage the graph-like topology of biochemical pathways to learn complex dependencies between reactions. Additionally, GNNs do not require explicit assumptions about cellular objectives. This thesis explores the application of GNNs for metabolic flux prediction, aiming to develop a framework that overcomes the limitations of classical methods.

2 Methods

The study is based on two primary sources of biological information: a *metabolic network* model and *transcriptomic data*. To represent metabolic pathways, we adopted ENGRO [2], a manually curated metabolic model that captures the core central carbon metabolism and essential amino acid metabolism. This model is available in JSON and SBML (XML) formats and consists of a structured set of metabolites, reactions, and genes, where genes are associated with reactions via gene-protein-reaction (GPR) rules.

In addition to the metabolic model, we utilized gene expression data from a spatial transcriptomics dataset obtained from a kidney sample affected by Clear Cell Renal Cell Carcinoma (ccRCC). This dataset includes biological samples (spatial spots) and a set of genes with their respective expression values per sample. Such datasets are particularly useful for studying metabolic heterogeneity, allowing for the identification of metabolic differences between tumor and healthy tissue regions.

During the evaluation of different methodologies for predicting metabolic fluxes, one method that attracted particular interest was *single-cell Flux Estimation Analysis* (scFEA) [3]. This approach circumvents the constraints of FBA by implementing an unsupervised flux prediction framework based solely on transcriptomic data and the implementation of a biological loss function to guide predictions. scFEA models metabolism at the level of predefined metabolic modules, which are aggregations of reactions derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [4]. However, since ENGRO represents metabolism at the reaction level rather than as aggregated modules, applying scFEA required a significant restructuring of the input data.

To adapt scFEA to ENGRO, we tested two different approaches. In the first approach, we retained only the reactions that had a direct gene association, filtering out all reactions that were not explicitly linked to enzymatic activity. This ensured compatibility with scFEA's neural network architecture but resulted in a reduced metabolic network, limiting its biological relevance. The main drawback was the exclusion of exchange reactions, which are crucial for metabolite transport and nutrient uptake, and pseudo-reactions, which represent aggregate processes such as biomass production. In the second approach, to preserve all reactions in the ENGRO model, we artificially assigned gene associations to reactions that originally lacked them. Specifically, reactions without direct gene regulation were linked to the genes of neighboring reactions, allowing them to be processed by scFEA. However, this modification introduced biological inaccuracies, as gene regulation does not directly control the activity of these reactions, but their fluxes often depend on diffusion mechanisms. Despite these modifications, the reliance of scFEA on gene expression as the sole input feature, combined with its neural network architecture that does not employ explicit message passing mechanisms, restricted its flexibility. Additionally, scFEA does not account for reaction bounds or directionality, whether a reaction produces or consumes a metabolite, further limiting its accuracy when applied to ENGRO.

To overcome these challenges, we developed a novel Graph Neural Network (GNN)-based framework called *Metabolic Flux Predictor* (MFP). Unlike scFEA, the developed model allows for a fully flexible metabolic network representation, integrating all reactions from ENGRO without artificial constraints. By leveraging a Graph Attention Network (GAT) architecture, which employs a message-passing scheme to dynamically weigh the importance of each reaction's neighbors [5], the network propagates biochemical information across the metabolic network, learning relationships from the graph structure. To evaluate the effectiveness of this approach, we tested different model configurations. In one setup, GNN-generated node (reaction) embeddings were concatenated before being passed through a Multi-Layer Perceptron (MLP), allowing the model to learn metabolic interactions that may not be explicitly captured in the metabolic network due to approximation errors and pathway reconstruction limitations. In an alternative configuration, embeddings per reaction were processed individually by the MLP, testing the hypothesis that avoiding direct interaction between reactions during the final prediction step could return refined results. However, the former configuration was ultimately selected as the final one, as it provides a more accurate representation of the biological behavior. Figure 1 shows an overview of the chosen architecture.

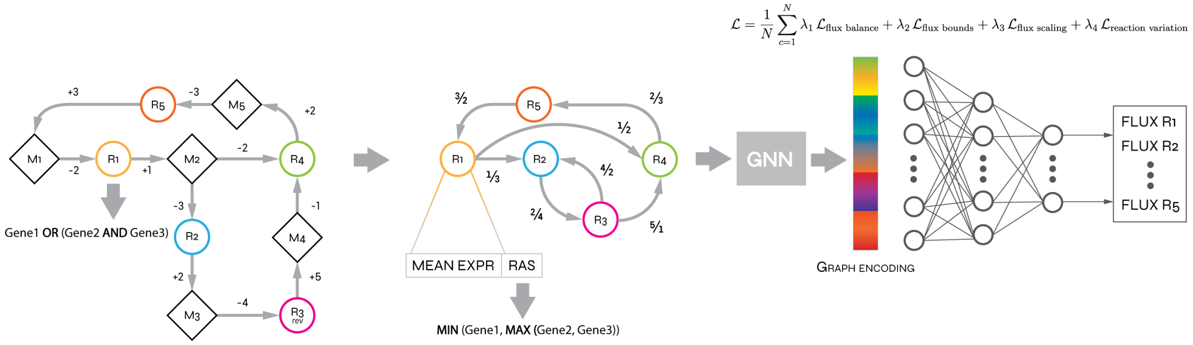
$\mathcal{L} = \frac{1}{N}\sum_{c=1}^{N} \lambda_1 \mathcal{L}_{\text{flux balance}} + \lambda_2 \mathcal{L}_{\text{flux bounds}} + \lambda_3 \mathcal{L}_{\text{flux scaling}} + \lambda_4 \mathcal{L}_{\text{reaction variation}}$
+3
R5
-3
M5
+2
M1
-2
R1
+1
M2
-2
R4
-3
-1
Gene1 OR (Gene2 AND Gene3)
R2
M4
+2
+5
M3
-4
R3 rev
3/2
R5
2/3
1/2
R1
R2
R4
1/3
4/2
2/4
R3
5/1
MEAN EXPR
RAS
MIN (Gene1, MAX (Gene2, Gene3))
GNN
GRAPH ENCODING
FLUX R1
FLUX R2
FLUX R5

Figure 1: Toy example of the Metabolic Flux Predictor architecture. The metabolic network is simplified into a set of reaction nodes, where edges represent the exchange rates of metabolites. Aggregated gene expression values are used as node features, and the entire graph is then passed through the neural network architecture.

3 Results

Evaluating the performance of metabolic flux prediction models presents a significant challenge due to the absence of ground truth flux measurements. Nevertheless, one of the main achievements of this work is the theoretical development of a flexible framework capable of incorporating gene expression information even for reactions without direct genetic associations. However, to assess the practical validity of the proposed approach, several evaluation strategies were employed, focusing on the ability of the model to *differentiate tumor and healthy tissues*, its *correlation with transcriptomic data*, and its adherence to established metabolic principles such as the *Warburg Effect* [6]. This phenomenon refers to the metabolic adaptation of cancer cells, which preferentially consume *glucose* to produce *ATP* and *lactate* rather than relying on *oxygen*-dependent mitochondrial respiration. Although energetically less efficient, this metabolic shift supports increased *biomass* synthesis, thereby facilitating rapid cellular proliferation.

The first tested approach, scFEA applied to the ENGRO network with reactions lacking gene associations removed, was evaluated for its ability to cluster metabolic states. The results showed that flux-based clustering identified three distinct clusters within tumor and healthy tissues, aligning well with transcriptomic clustering and highlighting the method's biological relevance. However, excluding key reactions, such as exchanges, limited its capacity to fully capture metabolic interactions.

To address this, a modified version of scFEA artificially assigned gene associations to previously unsupported reactions, enabling the prediction of exchange fluxes. While this improved clustering performance metrics, the three-cluster structure persisted. Additionally, while the model accurately predicted an increase in ATP flux in tumor regions and a reduction in glucose consumption, a substantial increase in lactate production, expected under the Warburg Effect, was not observed. More critically, the model incorrectly predicted higher biomass production in healthy regions, contradicting the well-established biological expectation that tumor cells should exhibit increased biomass synthesis. Furthermore, the flux distributions exhibited high variance and numerical instability, suggesting fundamental issues in the method's formulation, particularly in its loss function, which aggregates fluxes rather than enforcing biologically meaningful constraints.

The Metabolic Flux Predictor (MFP) developed in this thesis was evaluated to determine whether it addressed these shortcomings. The model effectively distinguished tumor from healthy metabolic states while producing more stable and interpretable flux distributions than scFEA. Significant metabolic shifts were captured, particularly in alignment with the Warburg Effect, with tumor cells exhibiting increased ATP and lactate production, increased glucose uptake, and lower oxygen consumption. However, biomass synthesis remained a challenge, showing only a slight, non-significant increase in tumors, likely due to its large connectivity and complex regulatory dependencies.

The alternative MFP configuration, where node encodings were processed individually rather than concatenated, was tested to assess the impact of removing global dependencies between reactions. This modification improved biomass differentiation and clustering quality but failed to fully capture the metabolic rewiring expected in tumors, emphasizing the need to balance local and global metabolic information. These findings suggest that while reducing reaction interdependencies can enhance certain aspects of metabolic modeling, a complete decoupling may hinder the accurate prediction of flux patterns within the broader network.

4 Conclusions and Future Directions

The objective of this thesis was to explore deep learning-based approaches for metabolic flux analysis and assess their ability to overcome the restrictions of traditional constraint-based methods such as FBA. This work investigated deep learning methodologies applied to metabolic flux estimation, with a focus on scFEA and a novel GNN-based approach. The evaluation of scFEA highlighted both its strengths and limitations: while it successfully integrated transcriptomic data into flux estimation, it was constrained by its reliance on a predefined reduced metabolic network and its inability to natively handle exchange and pseudo-reactions. To overcome these issues, this thesis introduced a GNN-based model designed to provide a more flexible and biologically grounded approach to flux prediction. In contrast to the scFEA, the proposed framework permits the incorporation of arbitrary metabolic networks, incorporates reaction-specific characteristics such as reversibility and flux constraints, and utilizes message passing to capture both local and global metabolic interactions. Through multiple validation strategies, the model demonstrated its ability to differentiate tumor from healthy metabolic states while capturing biologically relevant metabolic shifts, including key aspects of the Warburg Effect. Clustering analysis confirmed that flux predictions maintained strong spatial coherence and were closely aligned with transcriptomic profiles. However, the comparison between different model configurations revealed trade-offs between clustering accuracy and biological consistency, emphasizing the need to balance local metabolic interactions with global network-wide dependencies. Despite the challenges of modeling cellular metabolism without ground-truth flux measurements, this thesis establishes a foundation for deep learning-based metabolic modeling. The proposed framework serves as a starting point for future research, providing a flexible and extensible approach that can be adapted to different metabolic networks and input data types.

Looking ahead, three main directions can be explored to further improve deep learning approaches for metabolic flux estimation. First, enhancing model architecture by incorporating temporal dependencies through Temporal GNNs to capture metabolic dynamics [7], self-attention mechanisms to refine metabolic interactions [8], and graph pooling techniques to improve scalability while preserving biological structure. Second, integrating additional biological context by incorporating more multi-omics data or prior biochemical knowledge to constrain some predictions and ensure biological consistency. Third, exploring alternative modeling strategies, such as decoupling flux predictions at the node level, variational inference for uncertainty estimation, or pretraining strategies to better capture information for reactions lacking direct expression data.

References

[1] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010.

[2] Marzia Di Filippo, Dario Pescini, Bruno Giovanni Galuzzi, Marcella Bonanomi, Daniela Gaglio, Eleonora Mangano, Clarissa Consolandi, Lilia Alberghina, Marco Vanoni, and Chiara Damiani. Integrate: Model-based multi-omics data integration to characterize multi-level metabolic regulation. *PLoS Comput Biol*, 18(2):e1009337, 2022.

[3] N Alghamdi, W Chang, P Dang, X Lu, C Wan, S Gampala, Z Huang, J Wang, Q Ma, Y Zang, M Fishel, S Cao, and C Zhang. A graph neural network model to estimate cell-wise metabolic flux using single-cell rna-seq data. *Genome Research*, 31(10):1867–1884, Oct 2021. Epub 2021 Jul 22.

[4] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 01 2000.

[5] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[6] Maria V Liberti and Jason W Locasale. The warburg effect: How does it benefit cancer cells? *Trends in Biochemical Sciences*, 41(3):211–218, Mar 2016. Received: 2015/09/11, Revised: 2015/11/24, Accepted: 2015/12/04, Published online: 2016/01/05.

[7] Antonio Longa, Veronica Lachi, Gabriele Santin, Monica Bianchini, Bruno Lepri, Pietro Lio, Franco Scarselli, and Andrea Passerini. Graph neural networks for temporal graphs: State of the art, open challenges, and opportunities, 2023.

[8] Mengying Jiang, Guizhong Liu, Yuanchao Su, and Xinliang Wu. Self-attention empowered graph convolutional network for structure learning and node embedding, 2024.