Predicting Isotopologue Counts from Bulk Metabolomics Data

by

Ravindra Bisram

Thesis submitted in partial fulfillment of the
requirements for the degree of Master in Engineering

Department of Electrical Engineering
THE COOPER UNION FOR THE ADVANCEMENT OF SCIENCE AND ART
ALBERT NERKEN SCHOOL OF ENGINEERING

Advisor: Prof. Samuel Keene

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Metabolism describes the set of all enzyme-catalyzed reactions that transform nutrients into molecules that support biological function. Stable-isotope tracing is a method to measure intracellular metabolic pathway utilization by feeding a cellular system a stable-isotope-labeled tracer nutrient. In this study, we present a novel deep-learning model that predicts metabolic fluxes in the form of isotopologue counts from bulk metabolomics data. Our model was trained on a large dataset of stable-isotope tracing experiments using MALDI-MSI and was able to accurately predict the proportions of isotopologue counts in the absence of labeled tracers. By leveraging the power of deep learning, our model was able to capture complex relationships between metabolites and predict metabolic fluxes with state-of-the-art accuracy. We demonstrate the effectiveness of our approach by comparing our model's predictions to those obtained through traditional stable-isotope tracing experiments. Our method has the potential to revolutionize the field of metabolomics by providing a non-invasive and cost-effective alternative to traditional stable-isotope tracing methods for predicting metabolic fluxes.

Acknowledgements

Thank you to my primary thesis advisor, Professor of Electrical Engineering at The Cooper Union Sam Keene, for guiding me through the thesis process, as well as instructing many of my fundamental engineering classes during my time at Cooper.

Thank you to my advisors at Memorial Sloan Kettering Cancer Center, principal investigators Dr. Ed Reznik and Dr. Wesley Tansey for welcoming me into your labs and shaping this research. During my time here, I have been exposed to a fascinating new world of computational biology and gained a deeper understanding of the intersection between machine learning and cancer research. I am grateful to have had the opportunity to work with such knowledgeable and supportive mentors who have helped me hone my skills and provided guidance throughout the research process.

Thank you to Amy Xie, my co-worker at MSK and co-author of the paper also written on this thesis topic. Your background in biology and statistics was paramount for this research.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

Table of Contents

List of Tables **vii**

List of Figures **viii**

1 Introduction **2**

1.1 An Overview 2

2 Background **3**

2.1 Metabolomics 3

2.2 Isotope Tracing 4

2.3 Mass Spectrum Imaging 5

2.4 Prior Work 7

2.5 Dataset 8

2.5.1 Methodology 8

3 Preprocessing **10**

3.1 Loading Dataset 10

3.1.1 IsoScope 10

3.1.2 TIC Normalization 11

3.1.3 Data Separation 12

3.2 Ranking 12

3.3 Moran's I 13

3.4 Z-score 14

3.5 Train-Test Split 14

4 Methods **16**
4.1 Principal Component Analysis 16
4.2 Deep Learning Model 16
4.3 Quantifying Uncertainty 17
4.4 Evaluation Metrics 18
4.4.1 Mean Squared Error (MSE) 18
4.4.2 Root Mean Squared Error (RMSE) 18
4.4.3 Mean Absolute Error (MAE) 18
4.4.4 Coefficient of Determination (R^2) 19
4.4.5 Spearman's rank correlation coefficient 19
4.5 Baseline 19

5 Results **20**
5.1 Principal Component Analysis for Isotope Tracing Data 20
5.2 Predicting Isotopologues from Bulk Metabolomics Data 24
5.2.1 Glucose Labeled Brain Data 24
5.2.2 3HB Labeled Brain Data 25
5.2.3 Other Tracers here 25

References **38**

APPENDICES **38**

A Proof of Concept **42**

B Python Implementation **43**
B.1 Libraries 43
B.2 Code 43

List of Tables

5.1 Evaluation metrics for 30 best-predicted metabolites in glucose labeled brain data ranked by Spearman's coefficient. Only the best-predicted isotopologue for each metabolite was included. 28

List of Figures

2.1 Principles of isotope tracing. (a) Isotopes of an element are atoms of that element with different masses due to the number of neutrons they contain. (b) Schematic of possible TCA cycle metabolite labeling patterns due to the incorporation of ^{13}C from a [U-$^{13}C_5$]glutamine tracer. Reproduced from Jeong et al. [13] 4

2.2 The schematic diagram of MS. Reproduced from Chen et al. [6] 6

2.3 Principle of MALDI Imaging mass spectrometry. Reproduced from Aichler et al. [1] 6

2.4 Isoimaging pipeline. Reproduced from Wang et al. [23] 8

3.1 Total Ion Count (TIC) Normalization on mass spectrogram for metabolomics data [16] 11

3.2 The data used for this method takes the form of a bulk metabolomics matrix with a corresponding isotopologue matrix. 12

3.3 Plots for isotopologues that fell below required Moran's I score to be considered for prediction (Left) vs isotopologue plots that passed the Moran's I criteria (right). 14

4.1 Deep Learning Model Architecture 17

5.1 Principal Components diagram for isotope tracing data 20

5.2 Explained Variance of each component in PCA on Isotope Tracing 21

5.3 Correlation heatmap between principal components and metabolite phenotypes 22

5.4 Second Principal component plotted against single phenotype 23

5.5 Best predicted isotopologues for glucose labeled brain data 26

5.6 Scatter plots of the 25 best-predicted isotopologues in glucose labeled brain data based on Spearman's rank correlation coefficient. The x-axis is the actual proportion of the isotopologue in the holdout sample and the y-axis is the predicted proportion. 27

5.7 Bar plot of the median rho value (Spearman's coefficient) for each isotopologue in Glucose labeled brain data. Isotopologues are ordered from highest median rho to lowest, with the color of the bar corresponding do an adjusted p-value above a certain threshold. 29

5.8 (Left) Ratio of total isotopologues that were deemed invalid for prediction based on Moran's I autocorrelation metric to those that passed the cutoff threshold. (Right) Out of those valid isotopologues that passed the cutoff threshold, the proportion of those that were successfully predicted vs those that were not. 30

5.9 Ratio of successfully predicted isotopologues to unsuccessfully predicted within each metabolite. 31

5.10 Worst predicted isotopologues for glucose labeled brain data 32

5.11 Best predicted isotopologues for 3HB labeled brain data 33

5.12 Cross validation results for top predicted isotopologues in 3HB labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate. 34

5.13 Cross validation results for top predicted isotopologues in 15NLeu labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate. 35

5.14 Cross validation results for top predicted isotopologues in 15NNH4CL labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate. 36

5.15 Cross validation results for top predicted isotopologues in 15NGln labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate. 37

List of Abbreviations

AI	Artificial Intelligence.
3HB	3-Hydroxybutyrate
C57BL/6	often referred to as "C57 black 6", "C57" or "black 6", is a common inbred strain of laboratory mice. It is the most widely used "genetic background" for genetically modified mice for use as models of human disease.
KD	Ketogenic Diet (mainly fat with low protein and very low carbohydrate).
ND	Normal Diet
'metabolite' m+00	The unlabeled isotopologue for a given metabolite. Used as the reference point of measuring the relative abundances of the other isotopologue levels for said metabolite.

Chapter 1

Introduction

Memorial Sloan Kettering Cancer Center (MSK or MSKCC) is one of the largest and most respected cancer centers in the world. Computational oncology is an emerging subfield of cancer research. It is made possible by high-throughput data collection techniques which produce rich datasets and advanced computational methods which make the vast quantity of information interpretable. Computational oncology involves the use of data science, machine learning, and other computational approaches to analyze complex biological data and develop models that can help predict and understand the behavior of cancer cells, the response to different treatments, and the progression of the disease. This field has the potential to revolutionize the way cancer is diagnosed, treated, and managed by providing more personalized and effective approaches based on individual patient characteristics and cancer biology.

1.1 An Overview

Two separate methods were developed for this thesis, both incorporating metabolomics isotope-tracing data. The initial stages of this research involved becoming comfortable with the intricacies of isotope-tracing data, specifically with the compositional nature inherent in this sort of data. To this end, we developed a method of performing principal component analysis (PCA) on compositional isotope tracing data. Such a technique could then be used to identify the most important metabolic pathways that contribute to the observed isotopologue patterns. A supporting library was developed in tandem with this technique to highlight important phenotypes per metabolite given a phenotype matrix.

The second, and more significant method developed for this thesis is a method to predict isotopologue counts from bulk metabolomics data using deep learning. This technique uses bulk metabolomics data produced from MALDI iso-imaging experiments and predicts spatial isotopologue concentrations. This method is the first of its kind in this field, and therefore state-of-the-art, paving the way for metabolic flux analysis without the need for costly isotope-tracing experiments.

Chapter 2

Background

2.1 Metabolomics

Metabolomics is the large-scale study of the small molecules involved in metabolism. Metabolism describes the set of all enzyme-catalyzed reactions that transform nutrients into molecules that support biological function. [10] There are four essential cellular functions that are performed by metabolism: providing energy by creating ATP, converting nutrients into simpler structures (catabolism), converting simpler structures into macromolecules (anabolism), and participating in cellular functions such as cellular signaling and gene transcription. [5]

The substrates and products of these metabolic reactions, called metabolites, are organic compounds with low molecular weight. Common classes of metabolites include sugars, lipids, amino acids, and fatty acids. Knowledge of the metabolic state in tumor and normal tissue samples can provide insights into the development and potential therapeutic treatment options for cancer. The analysis of endogenous metabolite profiles in tissues can lead to a deeper understanding of disease-related mechanisms. [1] For example, increased metabolite levels can be due to either faster production or slower consumption. Differentiating these alternatives is often critical, an example being when the production of a metabolite is enhanced in a disease state, then it is logical to inhibit the pathway. [11]

Metabolic flux is the rate of turnover of molecules in a metabolic pathway. [5] In recent years, metabolic flux analysis has been widely used in bioprocess engineering to monitor cell viability and improve strain activity. Metabolic flux analysis refers to a methodology for investigating cellular metabolism whereby intracellular fluxes are calculated using a stoichiometric model for the major intracellular reactions and applying mass balances around intracellular metabolites. [15] Metabolic fluxes represent the final outcome of cellular regulation at many different levels, and hence they are an ultimate representation of the cellular phenotype expressed under certain conditions. [17] Analysis of metabolic fluxes is therefore an interesting approach to the functional analysis of cells.

2.2 Isotope Tracing

Stable-isotope tracing is a method to measure intracellular metabolic pathway utilization (metabolic flux) by feeding a cellular system a stable-isotope-labeled tracer nutrient. [13] The power of the method to resolve differential pathway utilization is derived from the enrichment of metabolites in heavy isotopes that are synthesized from the tracer nutrient. Stable isotopes are elements that occupy the same position in the periodic table, and are essentially "chemically and functionally identical" to their natural counterpart, but differ in mass due to a different number of neutrons within the atomic nucleus. This difference in mass makes these isotopes analytically distinguishable from each other; therefore, if introduced into a system the metabolic fate of these isotopes can be "traced." [24] Stable isotopes are particularly useful for metabolic tracing since they are chemically and functionally identical to their non-labeled counterparts, yet they do not significantly alter the behavior of the biological system being studied. For the vast majority of the biological processes probed using stable isotopes, the amounts used, and hence levels of enrichment reached, are so low they will have no effect on normal metabolism. [24] Typically, the tracer contains ^{13}C, ^{15}N, or ^{2}H atoms. These heavy atoms can either be uniformly enriched throughout the tracer molecule or enriched at specific positions. [22] The choice of tracer depends on the pathway being investigated.

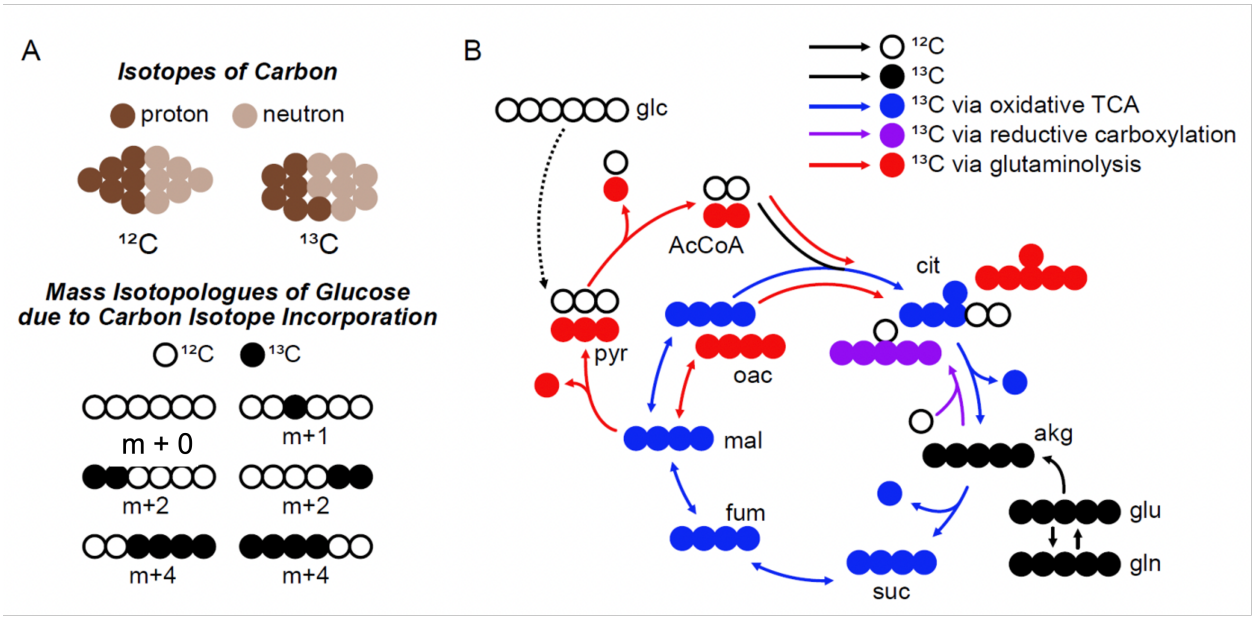


Figure 2.1: Principles of isotope tracing. (a) Isotopes of an element are atoms of that element with different masses due to the number of neutrons they contain. (b) Schematic of possible TCA cycle metabolite labeling patterns due to the incorporation of ^{13}C from a [U-$^{13}C_5$]glutamine tracer. Reproduced from Jeong et al. [13]

The principles of isotope tracing are shown in Figure 2.1. In Figure 2.1A, the stable isotope ^{13}C is created by adding an extra neutron to a carbon atom. The heavier atom can then be added to any of the six carbons of a glucose molecule to create a glucose tracer.

Out of the six carbons in glucose, if none are the ^{13}C isotope, it is termed m+00 aka unlabeled glucose. If any single carbon is labeled, this would be glucose m+01, and this pattern continues with the isotopologue number corresponding to how many carbon atoms are the ^{13}C isotope. If all of the carbons in the molecule are labeled, this is considered fully labeled glucose. Figure 2.1B shows a diagram of the tricarboxylic acid cycle (TCA cycle) with possible labeling patterns given a fully labeled glutamine tracer. The heavier atoms can be tracked through the cycle as a way to measure metabolic flux.

Stable isotope tracers are now at the forefront of the majority of physiological/biological mechanistic studies performed *in vivo*, and are complimented by the wealth of methodologies and techniques for determining flux through metabolic pathways and rates of substrate/pool turnover that are available, alongside the numerous analytical platforms which are now commercially available for the measurement of isotopic abundance. [24]

As cell types within tissues have different rates of nutrient uptake, infusing stable-isotope tracers to an isotopic pseudo-steady-state enables a more direct comparison of metabolic activity between spatially distinct regions. Analysis of labeled metabolites with iso-imaging at the isotopic pseudo-steady-state facilitates quantitation of different nutrients' fractional contributions to downstream metabolites across tissue regions. [23] Though it is now the industry standard for collecting this type of data, isotope-tracing experiments do face limitations - specifically when it comes to the cost and amount of these tracers needed. It is incredibly expensive and tedious to perform such an experiment, while traditional MALDI-MSI is only a fraction of the price.

2.3 Mass Spectrum Imaging

Modern metabolomics technology, led by innovations in mass spectrometry, has allowed for systems biological understanding at the gene, RNA, protein, or metabolite level. [26] Recent advances in high-throughput metabolomics technology allow for the identification and quantification of hundreds to thousands of metabolites in a small sample. [2] The bulk of metabolomics data in biology research is now generated using mass spectrometry, a method to quantify and identify chemical compounds in a sample. [19, 3] Mass spectrometry, following liquid or gas chromatography, reports the number of measured ions of each unique metabolite in the biological sample. Figure 2.2 displays the schematic of a typical MS experiment, from sample vaporization to ion detection.

Among the several mass spectrometry ionization techniques that can be used to directly analyze tissues, Marix-Assisted Laser Desorption Ionization (MALDI) has led the way in the development of biological and clinical applications for Imaging mass spectrometry and is therein one of the most commonly used techniques. [18, 4, 21] The principal workflow of a MALDI-MSI experiment is as follows. Tissue sections are covered with a matrix for extracting molecules from the tissue specimen into the matrix which aids desorption/ionization for further analysis in the mass spectrometer. A laser is shot into the tissue, and the matrix absorbs the laser energy and transfers the analytes to the gas phase, promoting ionization in the process. [14, 12] In the mass spectrometer, the tissue specimen is then raster-scanned, generating a mass spectrum for each measuring spot. [18]

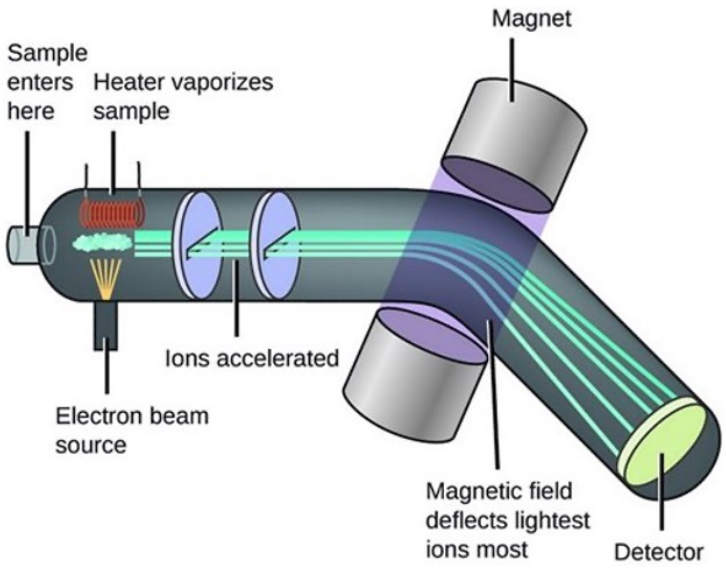


Figure 2.2: The schematic diagram of MS. Reproduced from Chen et al. [6]

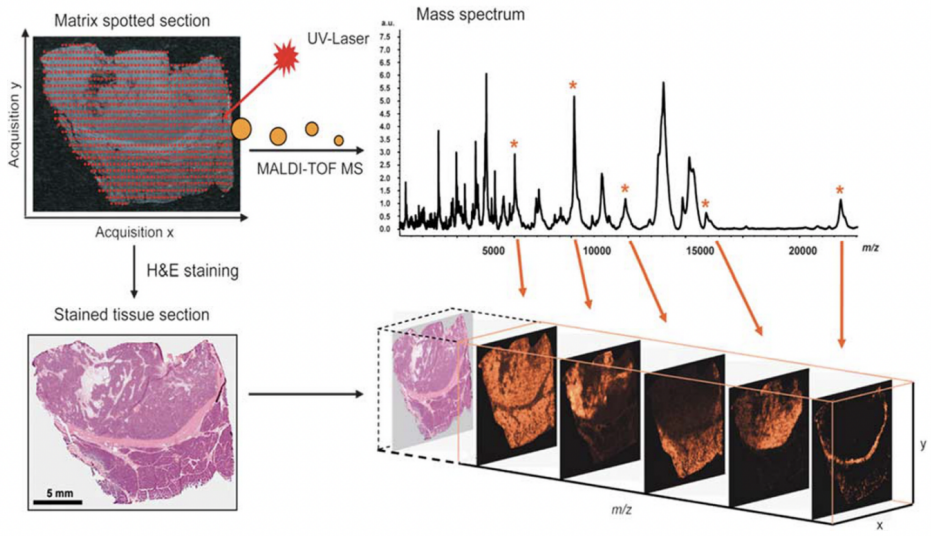


Figure 2.3: Principle of MALDI Imaging mass spectrometry. Reproduced from Aichler et al. [1]

The identification of molecules out of mass spectra is possible through databases containing the mass-to-charge (m/z) ratios of known metabolites. Especially the results from high-resolution mass spectrometers allow the identification of the peaks via those databases because they provide a high mass accuracy. [1] This was also confirmed by a validation approach in which liquid chromatography-tandem mass spectrometry (LC-MS/MS) was used to validate m/z-species from high-resolution mass spectrometry which were identified previously by database alignment. Once the spectrum of the sample is obtained, software is used to preprocess the data by identifying components, removing outliers, and resolving for batch-effects. [9]

2.4 Prior Work

The purpose of this review is to find and analyze any literature on the prediction of isotope-tracing data/metabolic flux analysis from bulk metabolomics data that exist in the context of mass spectrometry-based metabolomics data. We survey a collection of the most relevant work to the proposed model and highlight that there are no current methods close to our scope.

In their study 'Pseudo-transition Analysis Identifies the Key Regulators of Dynamic Metabolic Adaptations from Steady-State Data', Gerosa et al. aimed to identify the regulators responsible for metabolic adaptations in E. coli when transitioning between different carbon sources. [8] They developed a new approach called pseudo-transition analysis, an approach that uses multiple steady-state observations of 13C-resolved fluxes, metabolites, and transcripts to infer which regulatory events drive metabolic adaptations following environmental transitions. By applying this approach to eight different carbon sources, they found that the regulation of fluxes by transcription mainly occurs in the TCA cycle, while metabolites in the EMP pathway mainly regulate fluxes.

Heise et al. measure pool sizes of metabolites in order to estimate fluxes in plant cells. [10] Pool size measurements are necessary for estimating absolute intracellular fluxes in certain scenarios based on data from heavy carbon isotope experiments. The results show that accurate pool size measurements are necessary to improve the quality of flux estimates from non-stationary flux estimates in intact plant cells, and also identify specific metabolic pools that have a strong effect on the flux estimates of canonical pathways.

In terms of machine learning approaches to metabolism analysis, most work has been done in trying to discover or reconstruct pathways. Shah et al. uses machine learning to try and predict key missing enzymes in metabolic pathways. [20] This team used various machine learning algorithms such as SVM, KNN, and Bayesian models coupled with clustering models to predict gene networks and therefore missing enzymes. This is just one of many examples of machine learning being used to predict metabolic pathways or metabolite levels, but does not begin to touch metabolic flux.

Any prior works in the field of isotope-tracing prediction are inherently limited in scope due to the nature of data available up until recently. The dataset used for our method is the first of its kind, developed by a team of metabalomists specifically for the purpose

of isotope-tracing visualization. Prior software for MSI was developed for visualizing the spatial distribution of ion intensities but not labeling patterns. [23] This new imaging software enables the spatial display of the fraction of different isotopic forms of a metabolite (relative to all detected forms), the average extent of labeling (such as the average number of labeled carbon atoms in the molecule) and labeling normalized to the infused circulating tracer's labeling measured based on LC–MS analysis of serum. [23] Resultantly, we are in the unique position to have a vast amount of spatial data where we can process hundreds of metabolites at once at the pixel level. The proposed model is therefore state-of-the-art in this field.

2.5 Dataset

The isotope-tracing dataset used for this research was the open-source data from "Spatially resolved isotope tracing reveals tissue metabolic activity" by Wang et al. [23] This team not only performed isotope tracing experiments but more importantly developed the iso-imaging software isoScope. Iso-imaging builds upon traditional MSI by visualizing isotope labeling patterns, reflective of metabolic flux, in addition to ion intensities reflective of metabolite abundances. We establish a pipeline from isotope tracer infusion through data visualization, with a focus on steady-state infusions and labeling quantitation in carefully selected reported metabolites.

2.5.1 Methodology

The iso-imaging pipeline (depicted in Figure Figure 2.4) begins with the introduction of stable-isotope-labeled substrates. This was carried out by the infusion of ^{13}C- or ^{15}N-labeled nutrients into the right jugular vein of a fasted C57BL/6N mouse. An intermediate infusion rate resulting in 20–40% tracer labeling was selected, which was sufficient to facilitate downstream analysis with modest perturbation of endogenous metabolite levels and circulatory flux. [23]

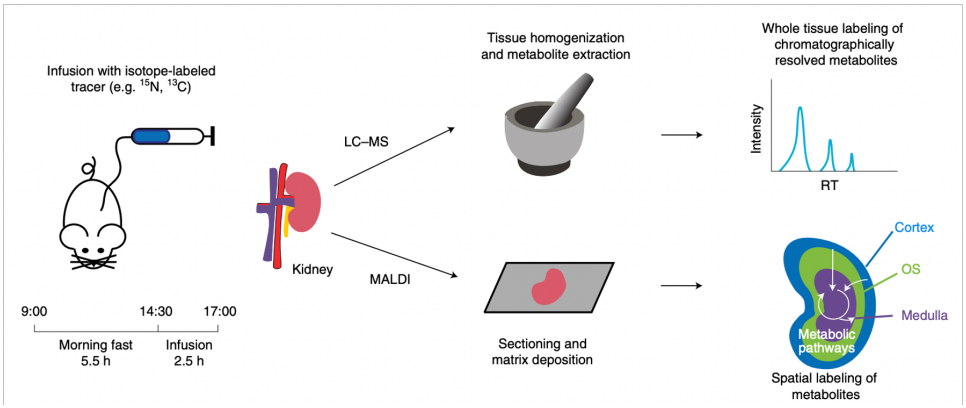


Figure 2.4: Isoimaging pipeline. Reproduced from Wang et al. [23]

Mice were killed and tissues were rapidly collected (< 2 min) and frozen using powdered dry ice. Tissues were then cryosectioned, the matrix was deposited and MALDI MSI was carried out on a 9.4 Tesla magnetic resonance cyclotron mass spectrometer at an average of 120,000 mass resolving power. By grinding a tissue slice and analyzing it with LC–MS in negative ionization mode, more than 200 canonical metabolites were identified based on m/z value and retention time. A standard approach for ion identification is tandem mass spectrometry (MS/MS) and 19 metabolite ions were validated in this manner. As a complementary method, isotopic labeling patterns achieved by MSI were compared to those resulting from LC–MS following the infusion of [U-13C] tracers. [23]

Up to this point, all experiments used mice that were fed a standard carbohydrate-rich chow. These mice are termed 'normal diet (ND)' mice. Given the dominance of glucose as a brain substrate under these conditions and the associated lack of clear spatial patterns in TCA cycle substrate choice within the brain, mice that were fed a ketogenic diet (mainly fat with low protein and very low carbohydrate) were also examined and termed 'KD'. Biochemically, a ketogenic diet induces high circulating levels of 3-hydroxybutyrate, which becomes one of the major brain fuels. [23] As expected, in the ketogenic diet condition, the contribution of glucose to brain glutamate and TCA cycle carbon fell, and that of 3-hydroxybutyrate increased.

Chapter 3

Preprocessing

The following sections detail the steps taken to import, preprocess, and alter the data to be used for a machine learning model.

3.1 Loading Dataset

3.1.1 IsoScope

The raw data was processed using the IsoScope software for MATLAB, developed by the same research team that the data was sourced from. The IsoScope developers describe it as an open-source, user-friendly software package to visualize and perform data analysis of mass spectrometry imaging (msi) data, with a focus on isotope labeling. It is specifically designed to make the analysis of isotopologues convenient and automatic. IsoScope speeds data handling for standard MSI functions, including generating ion images. More distinctively, it enables the spatial display of the fraction of different isotopic forms of a metabolite (relative to all detected forms), the average extent of labeling (such as the average number of labeled carbon atoms in the molecule), and labeling normalized to the infused circulating tracer's labeling measured based on LC–MS analysis of serum. [**?**]

IsoScope takes the '.mat' data of MATLAB structure as direct input, which is already processed after peak picking, much smaller in size, and contains centroid mass spectra peaks only. Given a predefined list of metabolites and their corresponding peaks (m/z), IsoScope finds it on the mass spectra and gets the intensity in each pixel, therefore generating an image. There is a tolerance setting (ppm) included in the software so that peaks of m/z within the tolerance are considered the same.

A short MATLAB script was written to extract the data from IsoScope while it is running and save it to a '.csv' formatted as x, y, metab_1 m+00, metab_1 m+01, etc. This data could then be loaded and processed further with Python. Any metabolites that were unlabeled (consisting of only an m+00 isotopologue) were dropped during this step.

It was observed that across replicates of a given tracer, the number of metabolites was not consistent. This is due to the fact that for some replicates, certain metabolites' peaks

were not detected at all in the spectrogram. These metabolites were programmatically identified and removed from consideration for the machine learning model. The names of these metabolites were also saved to be removed from testing data.

3.1.2 TIC Normalization

Most metabolomics data is normalized to reduce systematic bias or technical variations from non-biological sources, such as batch effects, while maintaining biological variation. [25] Various effects other than the distribution of endogenous proteins in a tissue sample can influence the intensity of signals in MALDI imaging datasets. In matrix-assisted laser desorption/ionization (MALDI) imaging, normalization is used to remove systematic artifacts that affect mass spectral intensity. [7] This avoids the misinterpretation of significant differences between sample sets that are due to non-biological factors. [3] Normalization attempts to adjust the ion counts in different batches to each other.

Total Ion-count Normalization (TIC Normalization) is one of the most common normalization procedures applied to mass spectrometry metabolomics data. [7][25] Here, all mass spectra are divided by their TIC so that all spectra in a dataset have the same integrated area under the spectrum. For each observation (pixel/row) of the data, each entry (metabolite) is divided by the sum of that row. This normalization approach is based on the assumption that there are comparable numbers of signals present in each spectrum. TIC is susceptible to being overly influenced by a small number of features with very high ion observations.

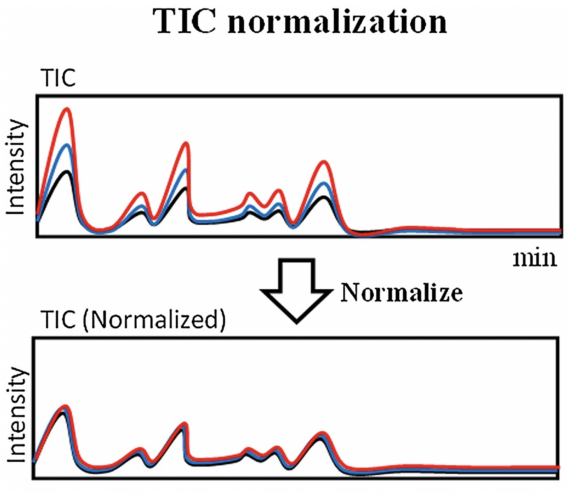


Figure 3.1: Total Ion Count (TIC) Normalization on mass spectrogram for metabolomics data [16]

There are several variations on the premise of TIC normalization, including median absolute deviation (MAD), mass spectrometer total useful signal (MSTUS), median normalization, and probabilistic quotient normalization (PQN). [25] These methods divide the ion intensity of each feature in a sample by a function of the sample spectrum. MSTUS

for instance computes the normalizer using the features common to all samples while PQN normalizes to a calculated control spectra. These methods operate under the assumption that the average ion count should be equal when batch effects are removed. [25]

3.1.3 Data Separation

The data file is then looped through, summing the isotopologues for each metabolite and saving the sum in a separate file of total ion counts. The resulting dataset consists of two files with the same number of observations (pixels): a total ion-count file with metabolites as the columns and an isotopologue count file. A log transformation of the total ion-counts is performed to make the data appear normal. The additive log-ratio transformation is performed on the isotopologue data to convert from raw ion counts to the proportion of that isotopologue in the metabolite.

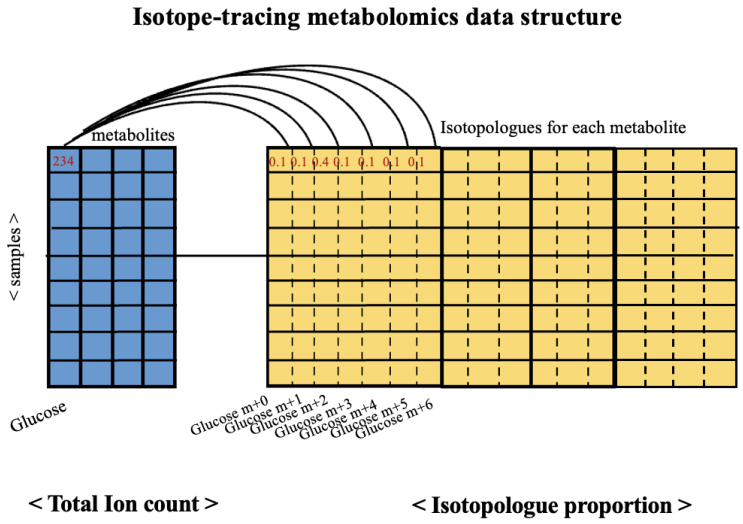


Figure 3.2: The data used for this method takes the form of a bulk metabolomics matrix with a corresponding isotopologue matrix.

3.2 Ranking

Although multiple replicates (e.g. different mice infused with the same tracer) were used in training the model, these replicates had varying percentages of the tracer in their bloodstream. Each mouse was injected with the same amount of tracer, but variances in their size and amount of blood led to different concentrations. This variance resulted in msi data that was inconsistent across replicates, with some replicates haven higher numerical values per pixel across the board, but showing the same relative patterns when the brain is plotted.

To normalize across replicates, each metabolite and isotopologue is rank transformed across the entire sample. For a given isotopologue in a sample, the pixel with the highest

amount of the isotopologue present is assigned a value of one, the pixel with the least amount a value of 0, and the remaining pixels are transformed accordingly between these two values. Following this process, a model can be trained on a series of replicates and tested on a held-out replicate with no issue, as the model will predict relative isotopologue density, not the absolute amount present in the pixel. The ranking was implemented using SciPy's 'rankdata' function from the stats library. It was empirically observed that the ranking of data did not alter the properties of the brain that are of interest, specifically the areas of high isotopologue concentration.

3.3 Moran's I

The Moran's I global spatial autocorrelation statistic measures the degree to which nearby observations in a dataset are similar to one another. This metric serves to identify potential patterns/clustering within spatial data. Moran's I is calculated by looking at the difference between the value of an observation and the average value of all nearby observations and compares this to the overall variation in the dataset. The value for Moran's I can range from negative one to one, with a negative value indicating that dissimilar values are likely to be clustered (the data is perfectly dispersed), a value of zero indicating that there are no correlations at all (random noise), and a positive value indicating that similar values tend to be clustered. The Moran's I metric is calculated as:

$$\mathcal{I} = \frac{N}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where y is the variable of interest, $\bar{y}$ is the mean of y, N is the number of spatial units indexed by i and j, $w_{i,j}$ is the spatial weight between feature i and j, and S_0 is the aggregate of all the spatial weights ($\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$). The spatial weights are the elements of a matrix with zeros along the diagonal (i.e. $w_{ii} = 0$).

This spatial autocorrelation metric is typically used in conjunction with a z-score and p-value to evaluate the significance of the index. However, for the purposes of this research, the team empirically selected a Moran's I cutoff score for predictable isotopologues by visually examining different brain images and comparing the images to their corresponding Moran's I value. If the brain plot for a certain isotopologue appeared to be noisy/static-like it was an indication that this isotopologue would be impossible to predict with any accuracy (Figure 3.3 left). On the other hand, isotopologues that displayed clear areas of a high metabolite concentration (through clustering in the brain plot) were deemed feasible for prediction (Figure 3.3 right). A cutoff value of 0.6 was chosen, with any isotopologues falling under that threshold being removed from consideration. When training the model across multiple replicates, an isotopologue had to exceed the cutoff amount for a majority of the replicates to be eligible for prediction.

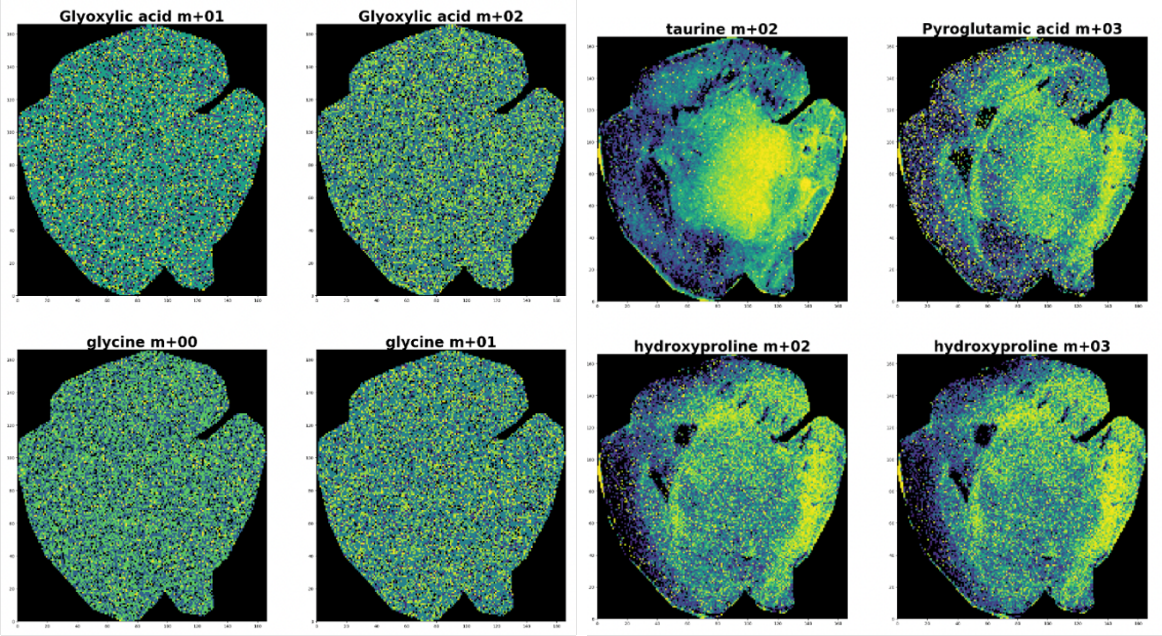


Figure 3.3: Plots for isotopologues that fell below required Moran's I score to be considered for prediction (Left) vs isotopologue plots that passed the Moran's I criteria (right).

3.4 Z-score

Both feature and target data for the model were normalized by z-scoring such that each metabolite had a mean of 0 and a standard deviation of 1. The formula to convert the data to z-scored is

$$z = \frac{x - \mu}{\sigma}$$

where x is the original data point, μ is the mean of that metabolite (column) and σ is the standard deviation of the metabolite. The benefit of this type of normalization is that any outliers will no longer have as large an influence when training the machine learning model. The model was trained on the dataset with and without the z-score, and it was found to not have much of a difference outside of stripping the plotted brains down to areas of high isotopologue concentration only. As a result, we opted to use the non-zscored results for discussion.

3.5 Train-Test Split

For a given tracer, the dataset used consisted of six replicates - three with a normal diet (ND) and three with a ketogenic diet (KD). To observe how the model performs across replicates, the training set for a given tracer consists of five of the replicates (all normal diet and two ketogenic diets) and is tested on the entirety of the holdout replicate. This process

is repeated as many times as there are replicates, with a different replicate being held out each iteration. The results are then analyzed as a whole. The majority of method testing was conducted on brain data for mice that were injected with glucose-labeled tracer.

At training time, the training data is randomly split into a training test and validation set using sklearn's 'train_test_split' functionality. A testing split of 20% was used for the validation set.

Chapter 4

Methods

4.1 Principal Component Analysis

Compositional data is defined as vectors of proportions that add up to a constant, typically representing the relative abundance of different components in a mixture, such as the relative abundance of different metabolite isotopologues in a biological sample. These proportions are inherently constrained and cannot be treated as independent variables. PCA is not suitable for compositional data because it assumes that the variables are independent and normally distributed, and it can generate misleading results when these assumptions are violated.

4.2 Deep Learning Model

The model selected for the task of predicting isotope-tracing data in the form of relative isotopologue abundances from bulk metabolomics data (total ion count) was a supervised deep-learning model. Deep learning is a subfield of machine learning that leverages artificial neural networks with many hidden layers (greater than or equal to 2). One of the main advantages of deep learning lies in its ability to solve complex problems that require discovering hidden patterns in the data and/or a deep understanding of intricate relationships between a large number of interdependent variables. Isotope-tracing data is inherently full of interdependent variables with metabolites having various effects on each other biologically, determining proportions of their own isotopologue counts, and the isotopologues themselves being a compositional set. Deep learning algorithms are able to learn hidden patterns from the data by themselves (no feature extraction required), combine them together, and build much more efficient decision rules.

We experimented with several different sets of parameters for our model, including the number of layers, the number of neurons per layer, activation functions, regularization techniques, training parameters, optimization algorithms, and learning rates.

The primary component of this model is a fully connected layer followed by a batch normalization layer. There are seven of these blocks put together, with a dropout layer every three blocks.

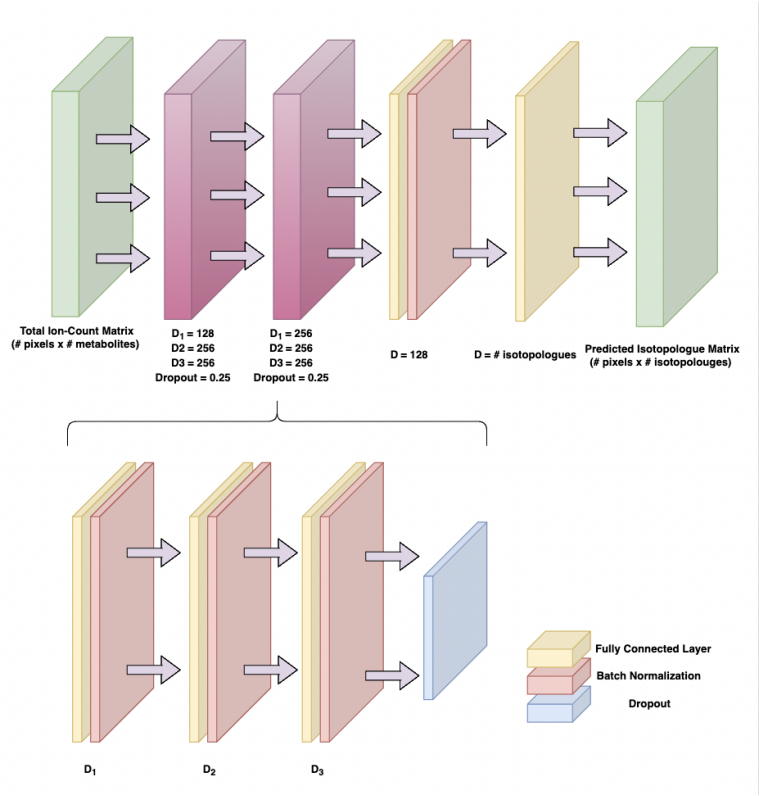


Figure 4.1: Deep Learning Model Architecture

4.3 Quantifying Uncertainty

Training a "bag of neural networks" (BoNN) is a technique used to quantify uncertainty in deep learning models. The idea behind this approach is to train multiple instances of the same neural network architecture with different initial conditions, such as different random weight initializations. Each instance of the network will produce slightly different predictions due to the random initialization, but by averaging the predictions of all the instances, the model's overall prediction can be obtained. This can provide a more robust and reliable estimate of the model's prediction compared to relying on just one instance of the network.

A realized coverage of 24.37% for the 95th percentile means that, for the test set used to evaluate the model, only 24.37% of the true observations fell within the 95% confidence interval predicted by the model. This indicates poor calibration performance, as the model is underestimating the true uncertainty and its predicted intervals are too narrow.

4.4 Evaluation Metrics

Metrics for regression involve calculating an error score to summarize the predictive skill of a model. The deep learning model has been trained to optimize mean-squared error (MSE), while we also report other popular evaluation metrics.

4.4.1 Mean Squared Error (MSE)

The mean squared error (MSE) assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. Squaring the residual (difference between predicted and actual value) eliminates negative values for the differences and ensures that the mean squared error is always greater than or equal to zero. Additionally, squaring increases the impact of larger errors. These calculations disproportionately penalize larger errors more than smaller errors. This property is essential when you want your model to have smaller errors.

The formula to calculate MSE is:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the i^{th} observed value, $\hat{y_i}$ is the corresponding prediction, and N is the number of samples. The mean squared error can be reported on a per-target (isotopologue) basis, or for the average of all the regression targets. This measure can be thought of as the variance of the residuals.

4.4.2 Root Mean Squared Error (RMSE)

The root mean squared error (RMSE) is calculated as the square root of the MSE, and can be thought of as a good estimate for the standard deviation of a typical observed value from the model's prediction. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

4.4.3 Mean Absolute Error (MAE)

The mean absolute error (MAE) is the average of the absolute values of the residuals. Absolute error preserves the same units of measurement as the data under analysis and gives all individual errors the same weights (as compared to squared error), which makes it less skewed towards outliers. The formula for MAE is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |(y_i - \hat{y}_i)|$$

4.4.4 Coefficient of Determination (R^2)

The coefficient of determination($\hat{R2}$) is a goodness-of-fit measure for regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between a model and the dependent variable on a convenient 0 – 1 scale.

4.4.5 Spearman's rank correlation coefficient

The Spearman rank-order correlation coefficient is a nonparametric measure of the monotonicity of the relationship between two datasets. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Spearman correlation at least as extreme as the one computed from these datasets. Although the calculation of the p-value does not make strong assumptions about the distributions underlying the samples, it is only accurate for very large samples (¿500 observations).

If the p-value for Spearman's correlation coefficient is exactly 0, it means that the observed correlation between the two variables being compared is extremely unlikely to have occurred by chance alone. This suggests that there is a strong relationship between the two variables and that the direction and strength of the relationship are unlikely to be explained by random chance.

4.5 Baseline

It is a good idea to first establish a baseline MAE for a dataset using a naive predictive model, such as predicting the mean target value from the training dataset. A model that achieves an MAE better than the MAE for the naive model has skill. The proposed model outperformed the naive predictive model by a significant margin. This was to be expected given that a prediction of the mean everywhere is useless when trying to predict isotope-labeling patterns within spatial tissue data. Results from the naive model showed an 'average' concentration of the isotopologue uniformly distributed throughout the tissue, which is both biologically and mathematically meaningless as we were working with ranked data.

Chapter 5

Results

5.1 Principal Component Analysis for Isotope Tracing Data

A novel method for conducting Principal Component Analysis was developed to work with isotope tracing data.

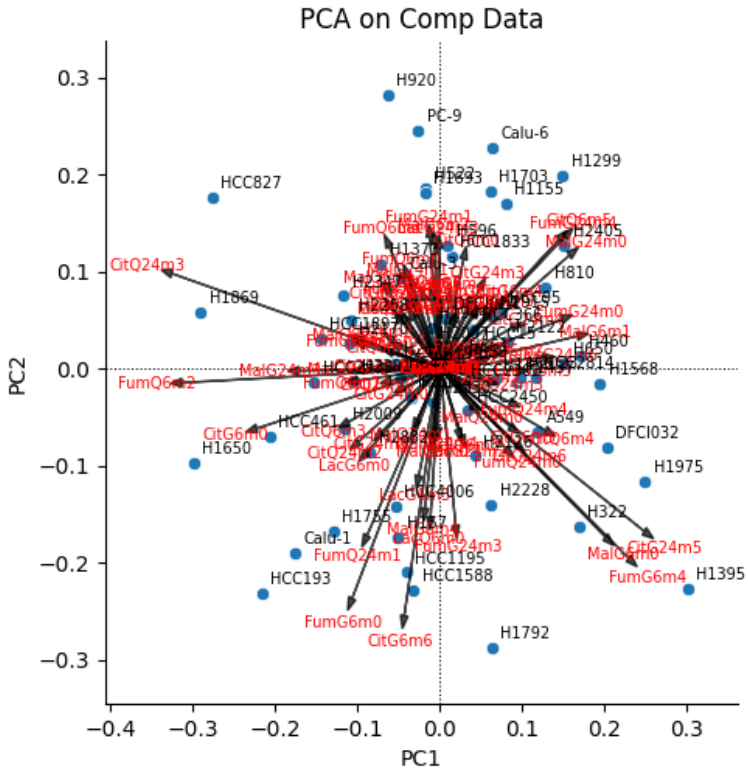


Figure 5.1: Principal Components diagram for isotope tracing data

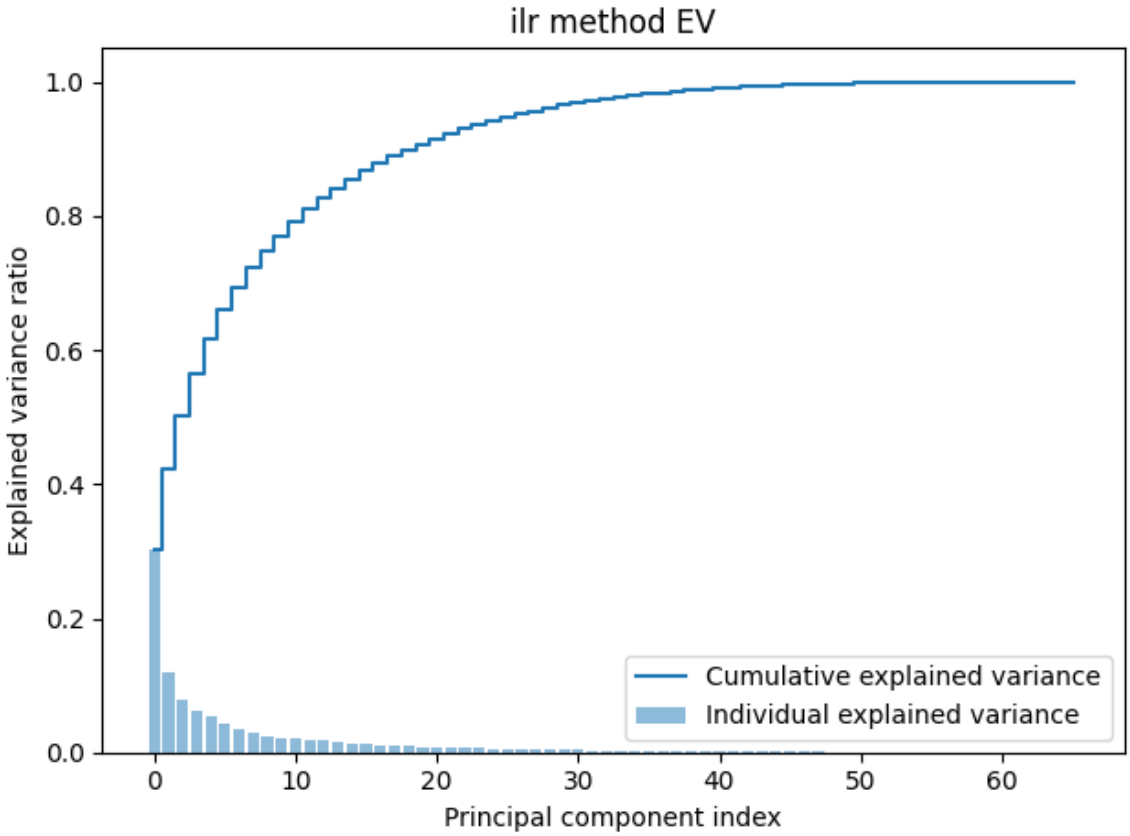


Figure 5.2: Explained Variance of each component in PCA on Isotope Tracing

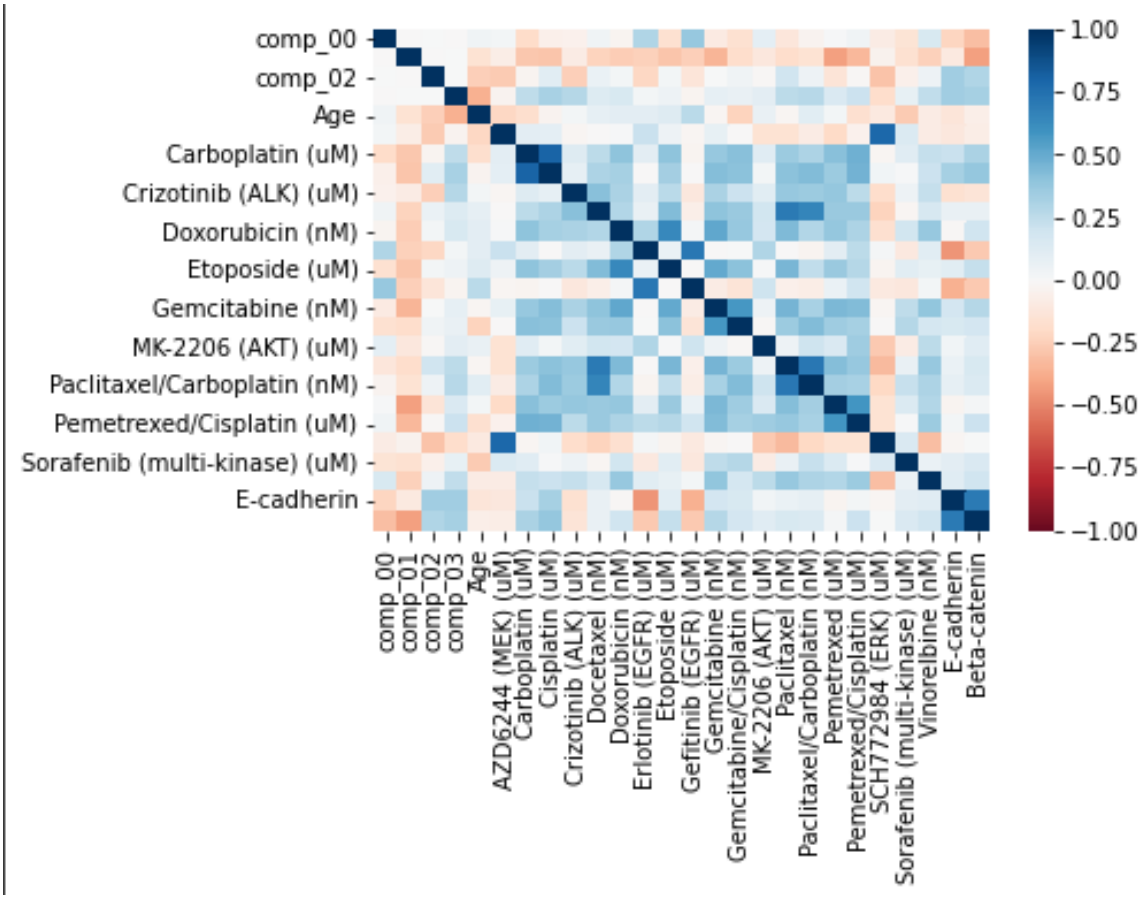


Figure 5.3: Correlation heatmap between principal components and metabolite phenotypes

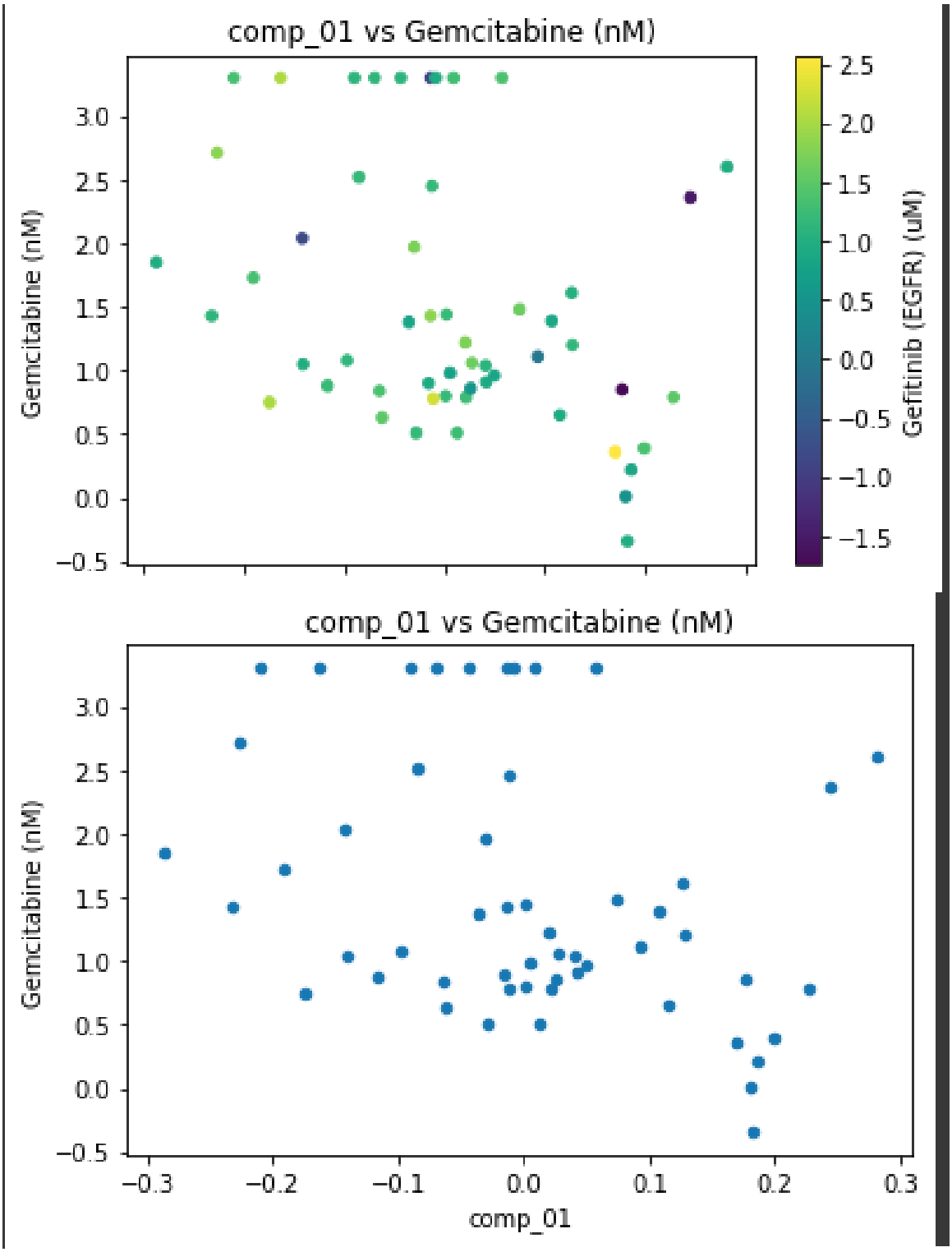


Figure 5.4: Second Principal component plotted against single phenotype

5.2 Predicting Isotopologues from Bulk Metabolomics Data

5.2.1 Glucose Labeled Brain Data

We applied our deep learning model to the MALDI-MSI Isotope-Tracing datasets from mouse brains and kidneys collected by Wang et al (carbon-13 labeled glucose was injected into the mouse). MALDI-MSI methods can measure the spatial distribution of metabolites at single-cell resolution, thus capturing cell-specific metabolic dynamics. The majority of early testing for this method was benchmarked on glucose-labeled brain data, which was tested on the hold-out mouse ND-M3. K-fold cross-validation was then later done with a different mouse being held out each time.

For each isotopologue, predicted values were then compared with their ground truth values to measure accuracy. We use the median Spearman's correlation coefficient between true values and predicted values as the metric to evaluate the performance of the model. Among all valid isotopologues that pass our Moran's I criterion, (specify percentage) isotopologues that had an R squared value larger than 0.3 were considered "well-predicted". A large subset of isotopologues had a very high median rho. For instance, Linoleic acid m+03 and Docosahexaenoic acid (DHA) m+04 were reproducibly well-predicted with an average rho of 0.8 and 0.78 respectively, demonstrating that our model accurately predicts isotopologue levels from total ion counts in bulk metabolomics data. In addition, our prediction preserved the spatial context of isotopologue abundances.

One of the primary purposes for isotope-tracing experiments relating to metabolic flux understanding is to provide a glimpse into the potential utility of spatially resolving 13C-labeling patterns with MSI. As cell types within tissues have different rates of nutrient uptake, infusing stable-isotope tracers to an isotopic pseudo-steady-state enables a more direct comparison of metabolic activity between spatially distinct regions. A significant result of our method is simply analyzing tissue plots empirically to identify the preservation of metabolite patterns. Figure 5.5a, shows a side-by-side comparison of the ground truth and predicted plots for the best-predicted isotopologue (as determined by Spearman's coefficient), Linoleic acid m+03, for the glucose labeling brain dataset. It is clearly observed that the pattern of high concentration forming a thick border around the top three-quarters of the brain is perfectly preserved while maintaining the region of low concentrations in the middle and bottom. Figure 5.5b shows another top predicted isotopologue, Docosahexaenoic acid (DHA) m+04, that was able to preserve a completely different high-concentration labeling pattern just as well (or arguably even better based on smoothing).

It was empirically observed that though the labeling patterns of a majority of isotopologues were preserved, the overall images were slightly smoothed out. This may be an indication that the model is slightly biased toward the mean value of this isotopologue across the testing replicates. It is worth noting that there are still isotopologues that this model is unable to predict with any degree of accuracy. Figure 5.9 displays the portion of isotopologues that could be predicted well and those that could not with respect to the

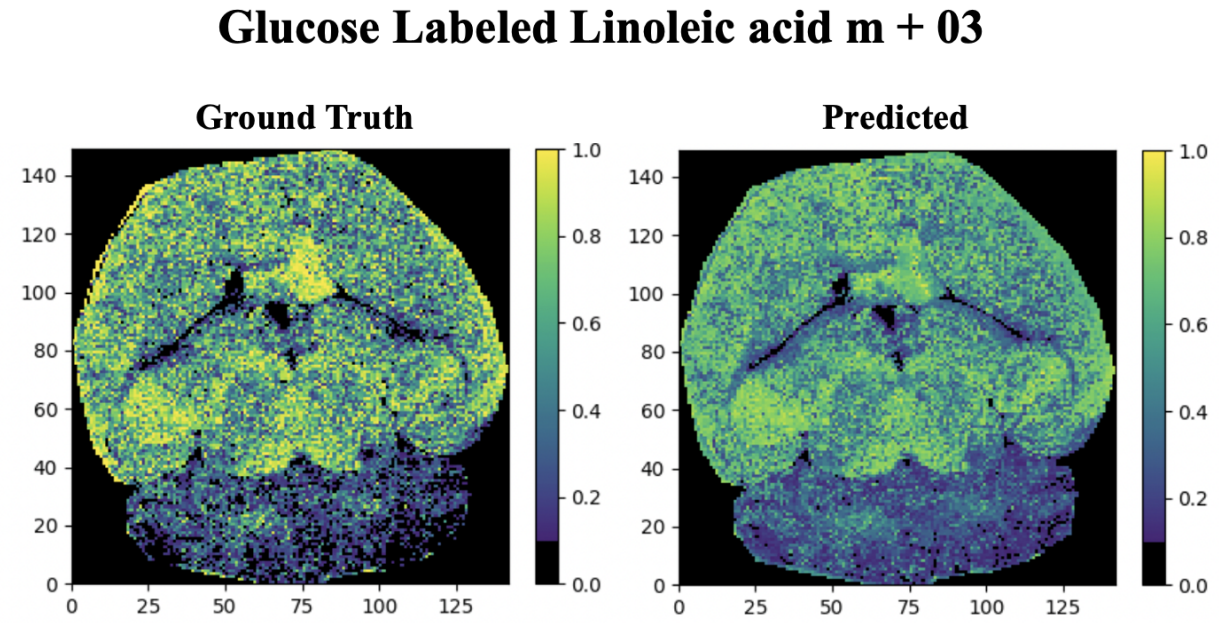
number of valid isotopologues that passed the Moran's I threshold. Roughly a fifth of the valid isotopologues could not be predicted, but when examining them empirically, such as in Figure 5.10a, a few trends are evident. A majority of the unpredicted isotopologues were themselves very noisy and likely did not pass the Moran's I test for this replicate. The only reason they would not have been removed from consideration is due to them being more strongly clustered in the other replicates for the same tracer. Both Glucose labeled homocystine m+06 and Glucose labeled UDP m+06 have very similar predictions, though their ground truths are vastly different. This points to some overfitting of the training data as this pattern is found in a large number of other isotopologues.

5.2.2 3HB Labeled Brain Data

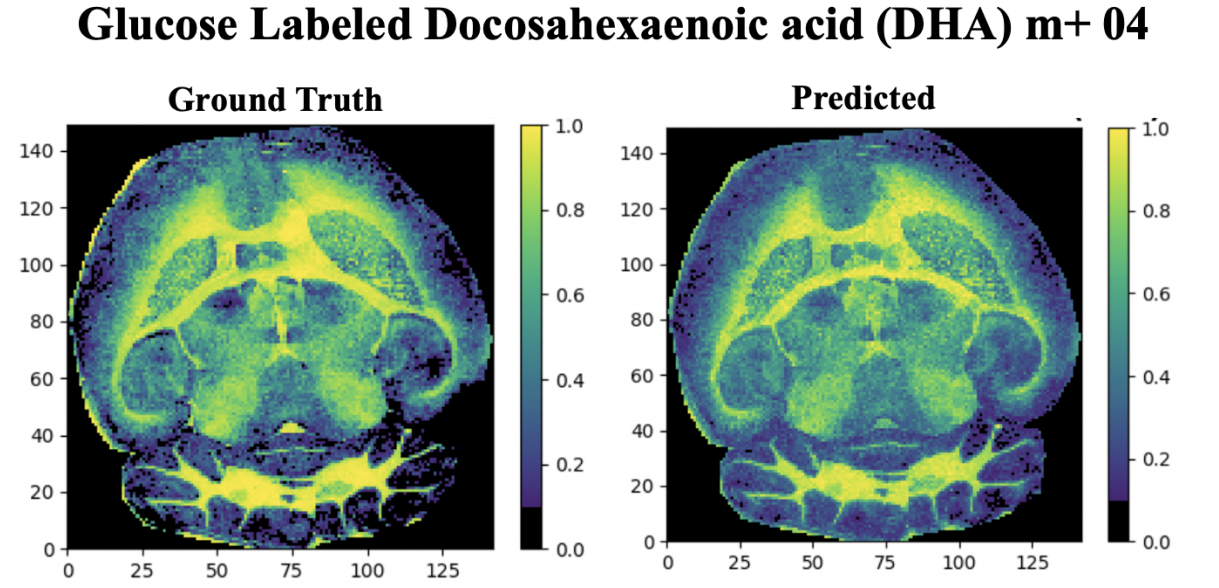
This method is not only proven to work on glucose-labeled brain data, but also displays equally promising results on different tracers. Figure 5.11a and Figure 5.11b show a side-by-side plot of the top two predicted isotopologues on replicate 'ND-M3' for the 3HB-tracing experiment.

5.2.3 Other Tracers here

This method will be tested on 4 other tracer datasets for brain data, as well as a series of kidney data. The results will be inserted here.



(a)



(b)

Figure 5.5: Empirical comparison between ground truth and predicted isotopologues in glucose labeled brain data. The top two best-predicted isotopologues based on Spearman's rank correlation coefficient are depicted. (a) Glucose labeled Linoleinc acid m + 03. (b) Glucose labeled Docosahexaenoic acid (DHA) m+04.

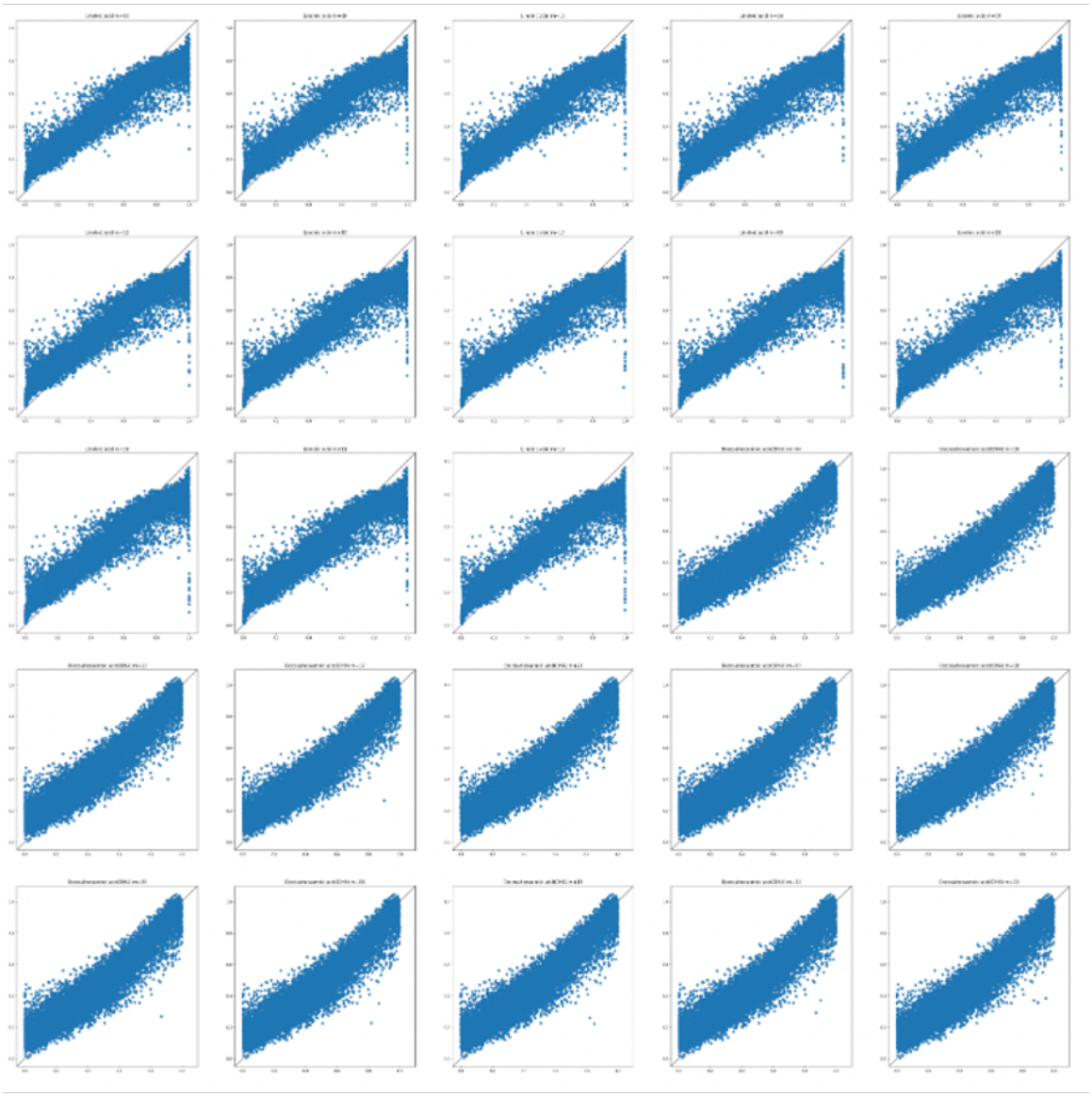


Figure 5.6: Scatter plots of the 25 best-predicted isotopologues in glucose labeled brain data based on Spearman's rank correlation coefficient. The x-axis is the actual proportion of the isotopologue in the holdout sample and the y-axis is the predicted proportion.

Isotopologue	Median Rho	P-value	MSE	MAE	R2
Linoleic acid m+03	0.9623	0.0	0.0094	0.0737	0.8871
Docosahexaenoic acid(DHA) m+04	0.9575	0.0	0.0083	0.0706	0.9009
pantothenate m+09	0.9509	0.0	0.0102	0.0786	0.8778
DL-Benzylsuccinic acid m+01	0.9453	0.0	0.011	0.0837	0.8679
UDP m+09	0.943	0.0	0.0099	0.0763	0.8809
GMP m+10	0.9415	0.0	0.0141	0.0973	0.831
Taurodeoxycholic acid m+25	0.938	0.0	0.0109	0.0808	0.8696
ADP m+07	0.9351	0.0	0.0105	0.079	0.8745
GDP m+09	0.9344	0.0	0.0113	0.0837	0.864
Carnosine m+08	0.9308	0.0	0.0169	0.1056	0.7976
DL-4-Hydroxy-3-methoxymandelic acid m+00	0.927	0.0	0.0149	0.0968	0.8216
AMP m+10	0.9154	0.0	0.015	0.0886	0.8202
dGMP m+10	0.9154	0.0	0.015	0.0886	0.8202
UMP m+02	0.9076	0.0	0.0164	0.0899	0.803
Cyclic-ADP-ribose m+14	0.9051	0.0	0.0152	0.098	0.818
cyclic-AMP m+10	0.9042	0.0	0.0164	0.103	0.8028
N-Acetylneuraminic acid m+11	0.9033	0.0	0.0168	0.1008	0.7978
Anserine m+06	0.9032	0.0	0.0156	0.0984	0.8132
Stearic acid m+11	0.9032	0.0	0.0187	0.1105	0.7754
Quinaldic acid m+07	0.9009	0.0	0.0178	0.1075	0.7867
fructose-1-6-bisphosphate m+05	0.8951	0.0	0.0192	0.1132	0.7691
aconitate m+01	0.8948	0.0	0.0197	0.1153	0.764
cystathionine m+00	0.8909	0.0	0.0251	0.1312	0.6989
glutathione m+10	0.8831	0.0	0.0199	0.1109	0.7607
Xanthurenic acid m+07	0.878	0.0	0.0211	0.1143	0.7472
N-Acetyl-L-asparagine m+00	0.8757	0.0	0.0239	0.1278	0.7131
riboflavin m+04	0.8753	0.0	0.0204	0.1095	0.755
CMP m+03	0.8717	0.0	0.0216	0.1199	0.7406
IMP m+10	0.8704	0.0	0.0246	0.1281	0.7051
oxoadipate m+04	0.8686	0.0	0.0247	0.1234	0.703
2-Hydroxy Hippuric Acid m+07	0.8557	0.0	0.023	0.1107	0.7245
D-Pantethine m+00	0.8555	0.0	0.0286	0.1414	0.657
S-Adenosylmethionine m+00	0.8504	0.0	0.0251	0.1269	0.6987
dAMP m+03	0.849	0.0	0.0245	0.1256	0.7059

Table 5.1: Evaluation metrics for 30 best-predicted metabolites in glucose labeled brain data ranked by Spearman's coefficient. Only the best-predicted isotopologue for each metabolite was included.

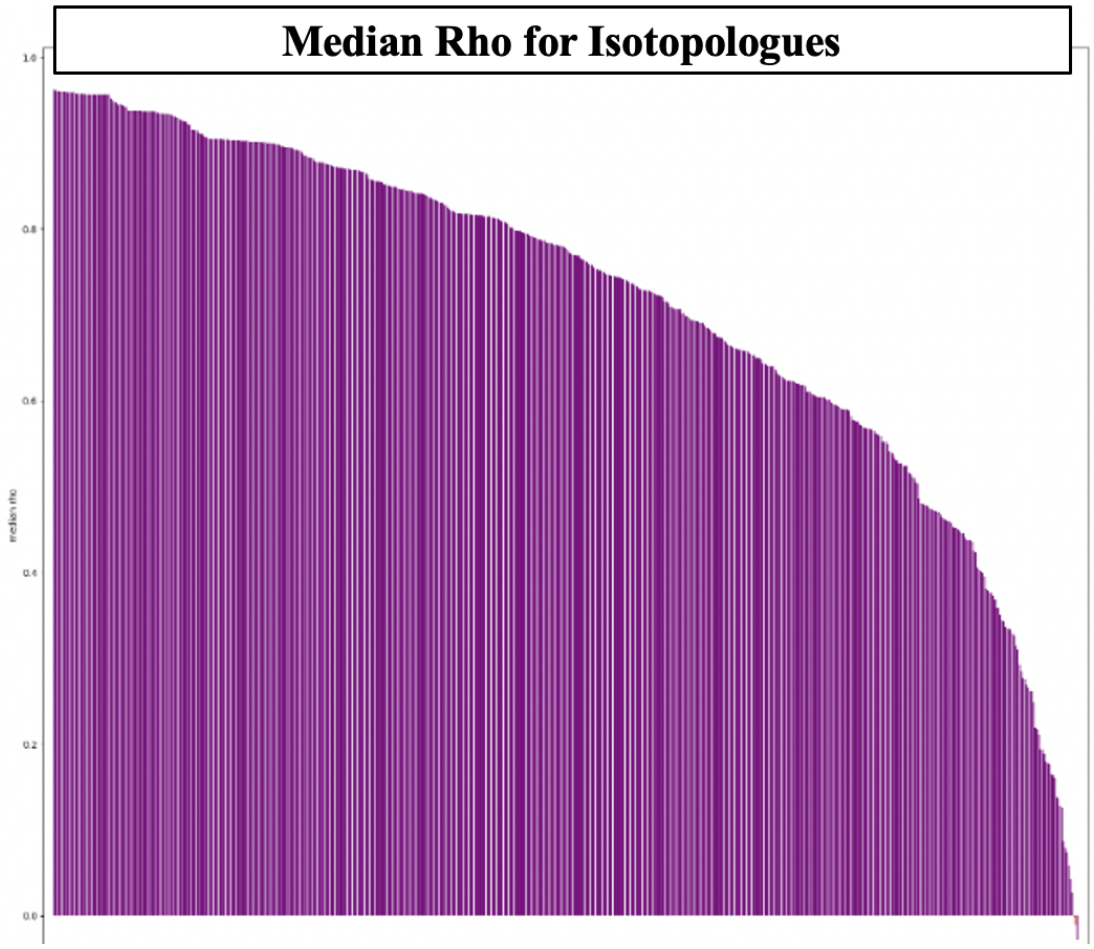


Figure 5.7: Bar plot of the median rho value (Spearman's coefficient) for each isotopologue in Glucose labeled brain data. Isotopologues are ordered from highest median rho to lowest, with the color of the bar corresponding do an adjusted p-value above a certain threshold.

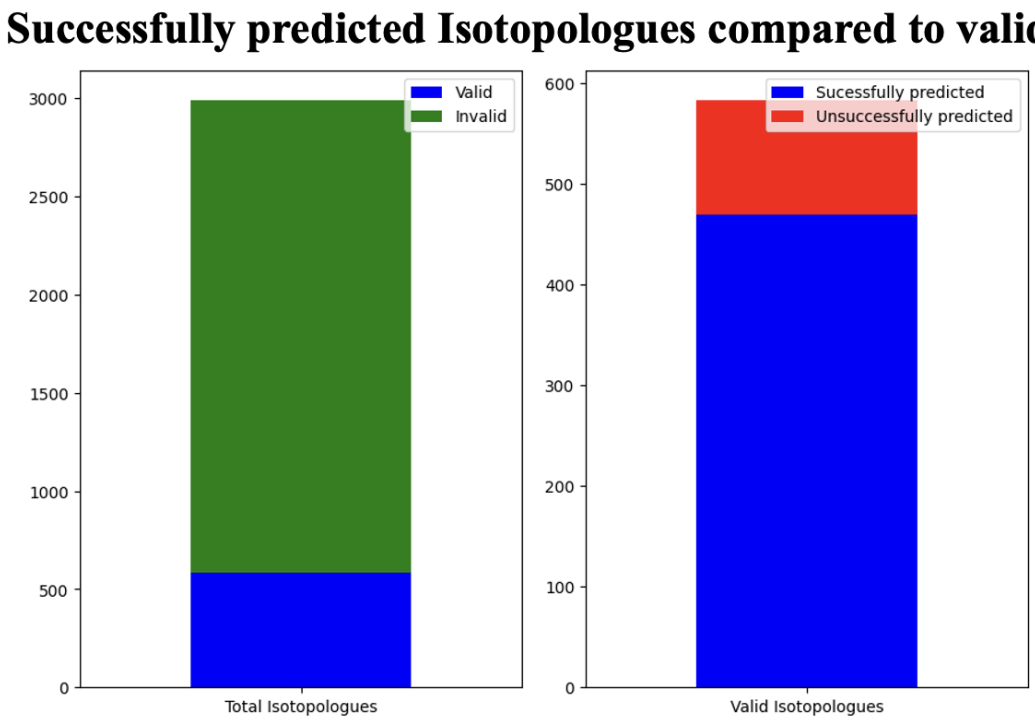


Figure 5.8: (Left) Ratio of total isotopologues that were deemed invalid for prediction based on Moran's I autocorrelation metric to those that passed the cutoff threshold. (Right) Out of those valid isotopologues that passed the cutoff threshold, the proportion of those that were successfully predicted vs those that were not.

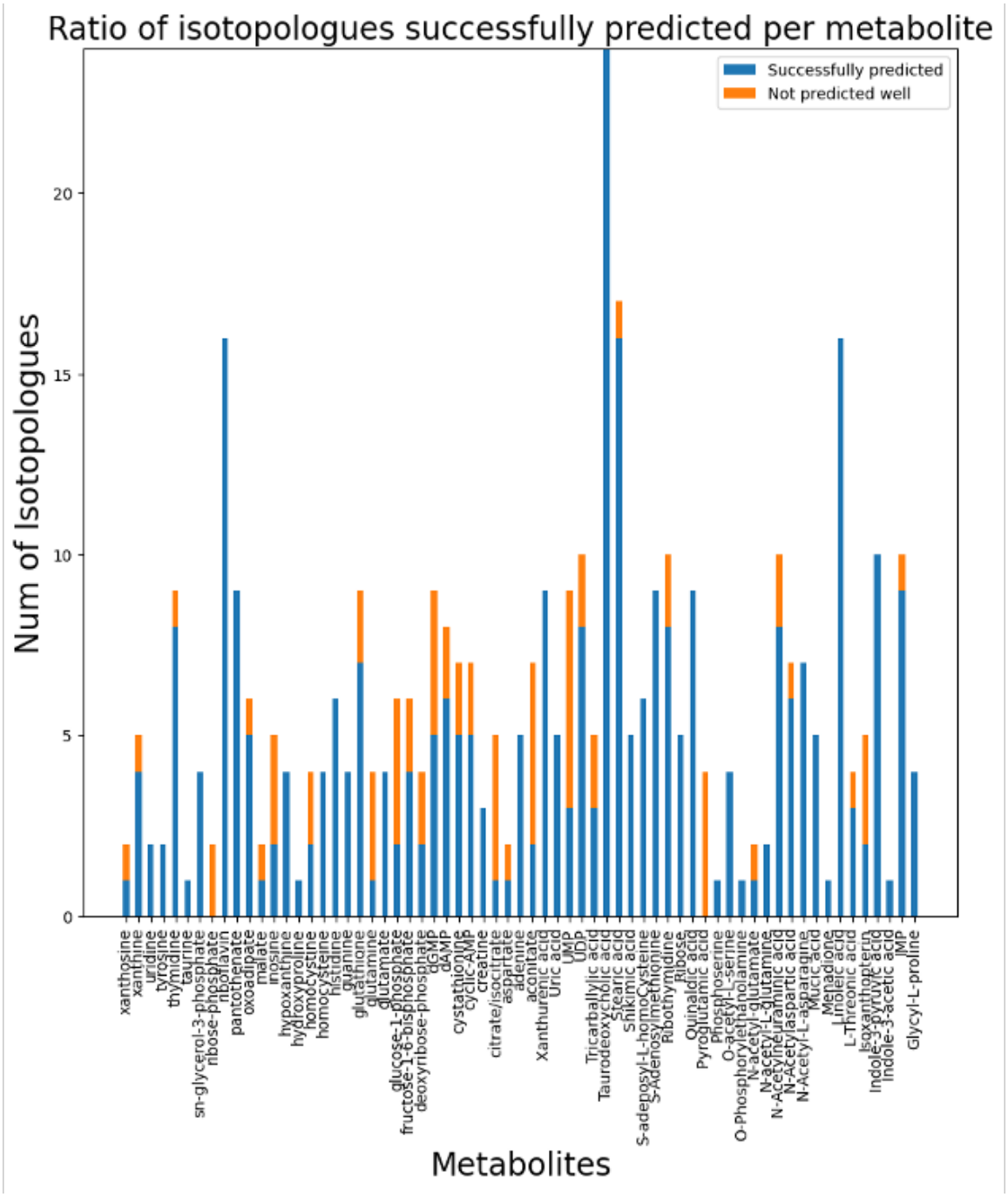
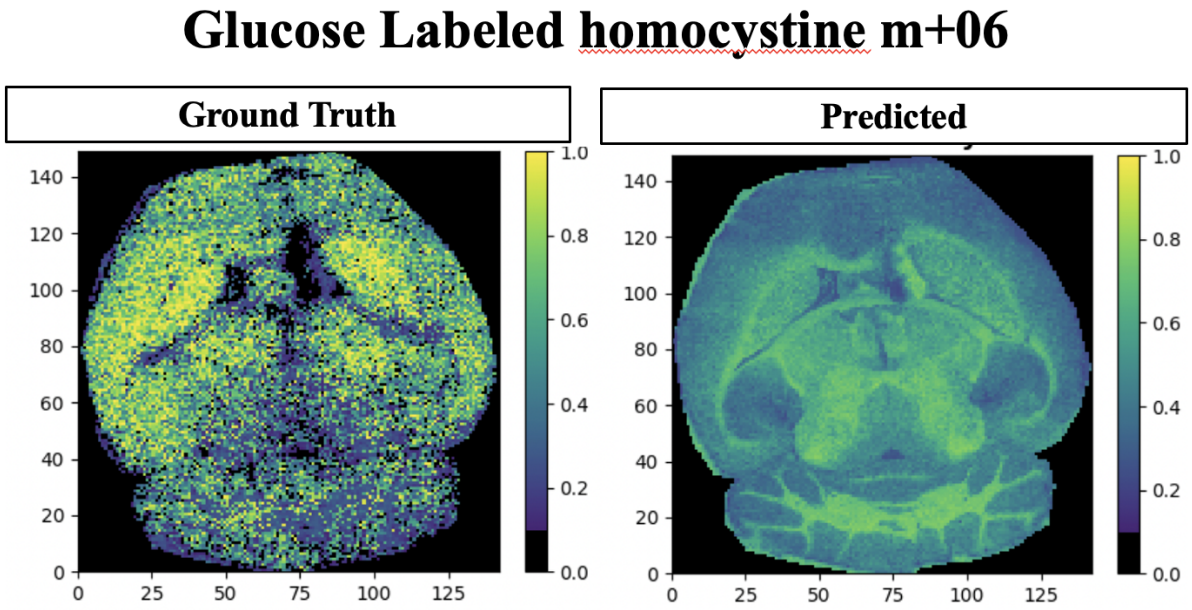
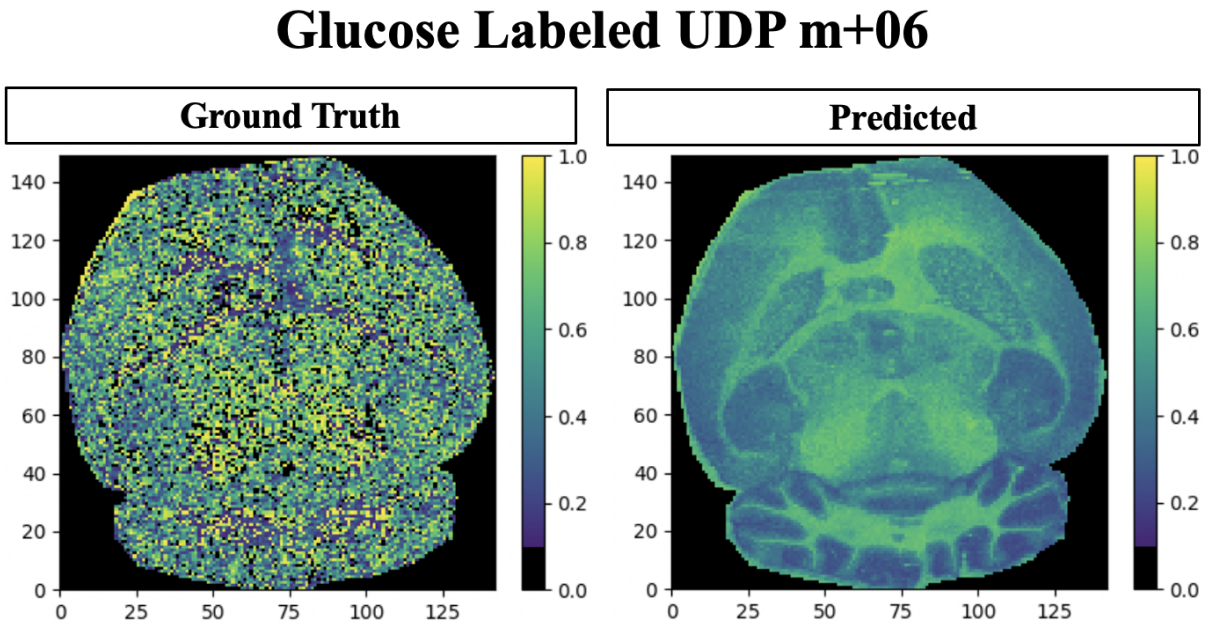


Figure 5.9: Ratio of successfully predicted isotopologues to unsuccessfully predicted within each metabolite.

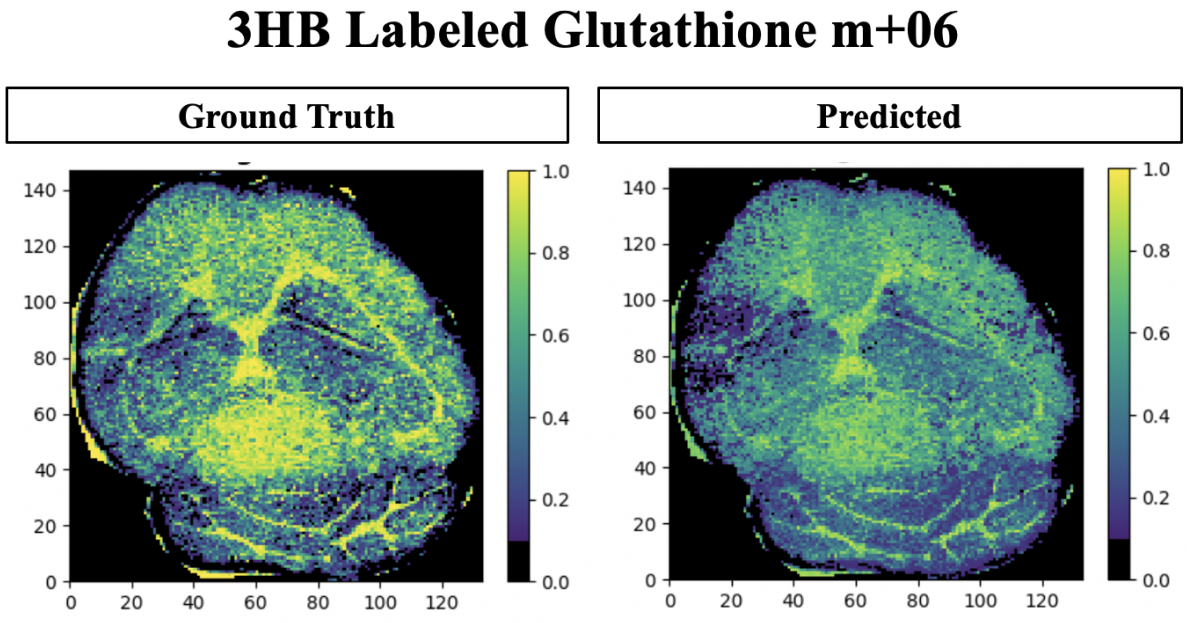


(a)

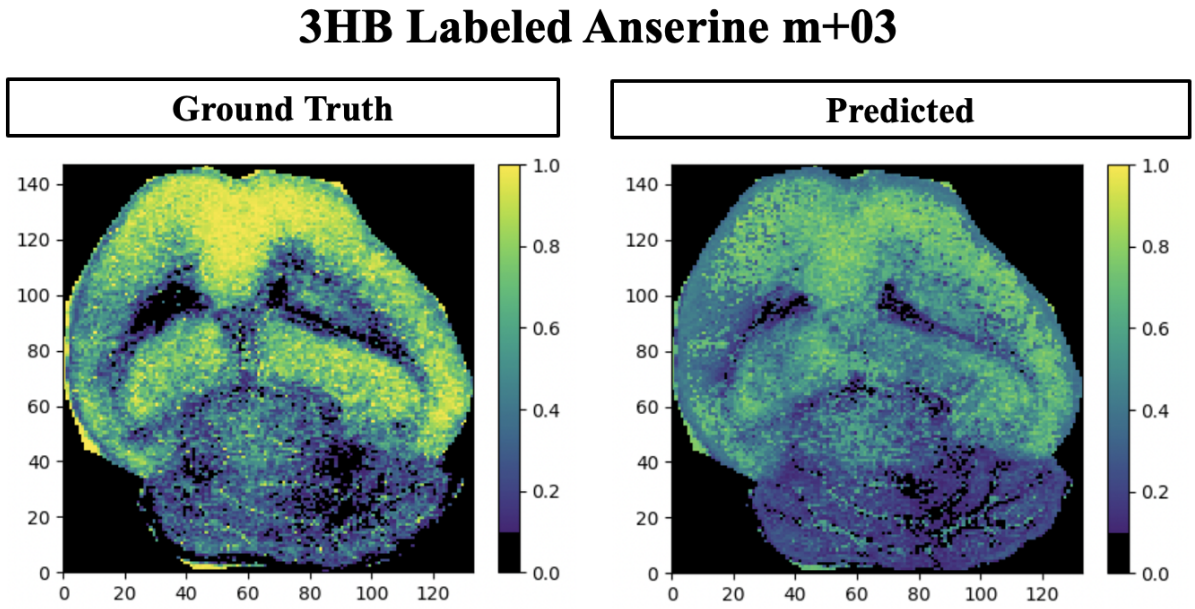


(b)

Figure 5.10: Empirical comparison between ground truth and predicted isotopologues in glucose labeled brain data. Two of the worst-predicted isotopologues based on Spearman's rank correlation coefficient are depicted. (a) Glucose labeled homocystine m+06. (b) Glucose labeled UDP m+06.



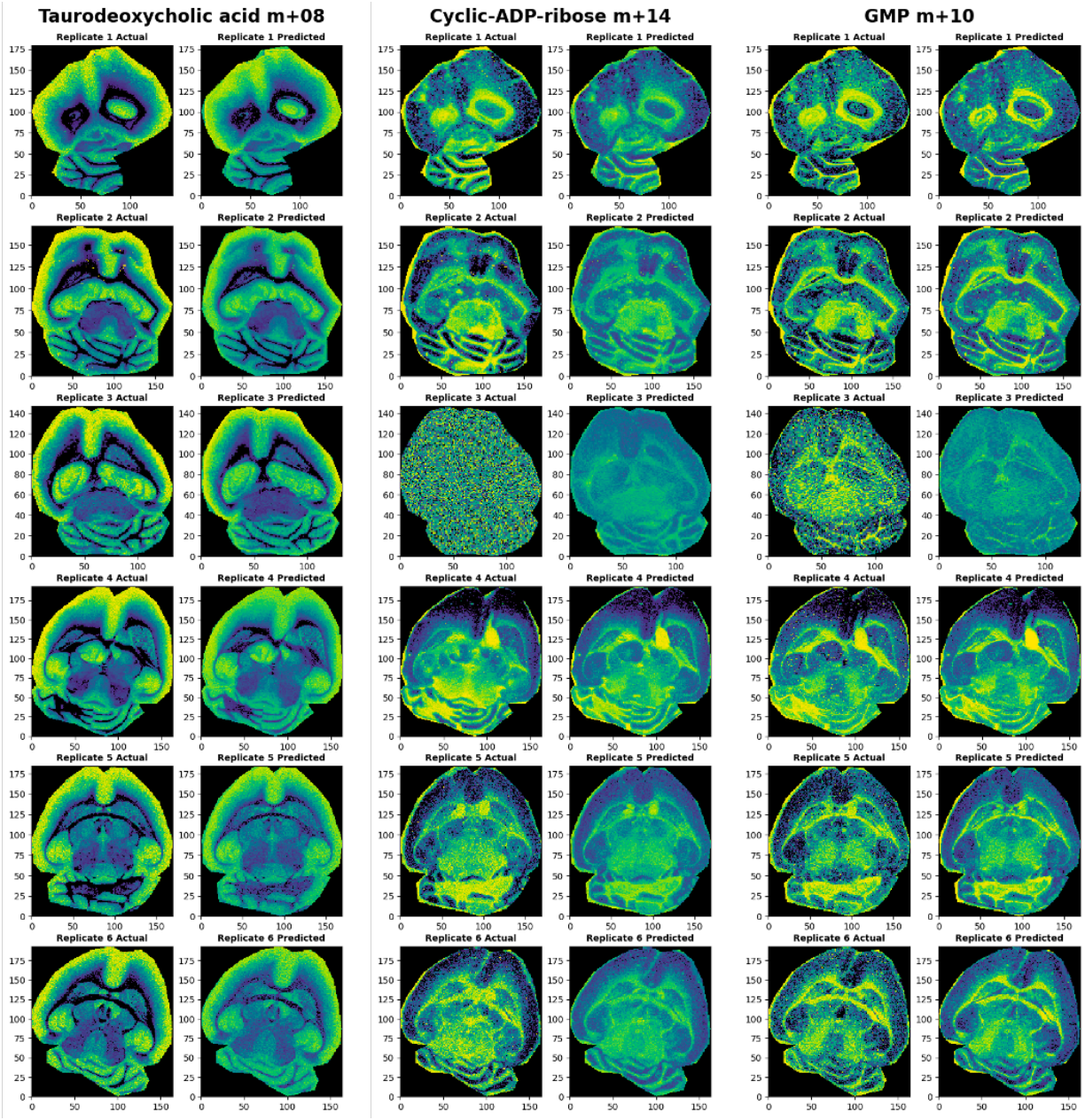
(a)



(b)

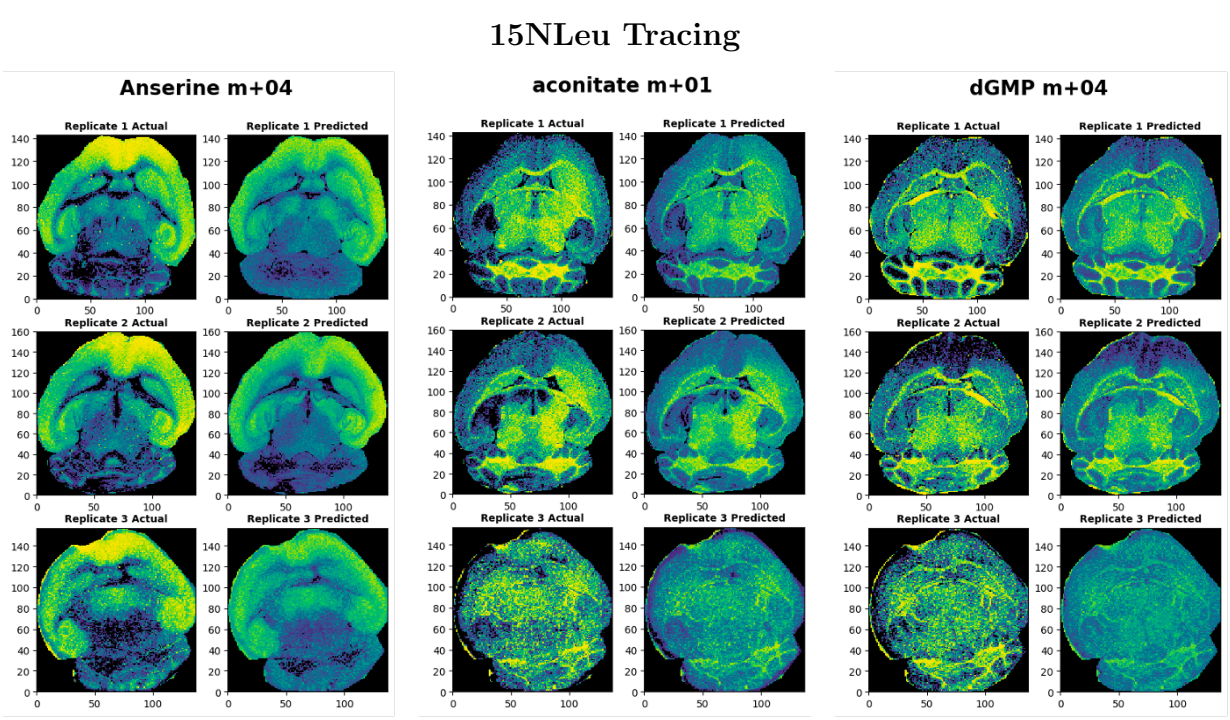
Figure 5.11: Empirical comparison between ground truth and predicted isotopologues in 3HB (3-Hydroxybutyrate) labeled brain data. The top two best-predicted isotopologues based on Spearman's rank correlation coefficient are depicted. (a) 3HB labeled Glutathione m+06. (b) 3HB labeled Anserine m+03.

3HB Tracing



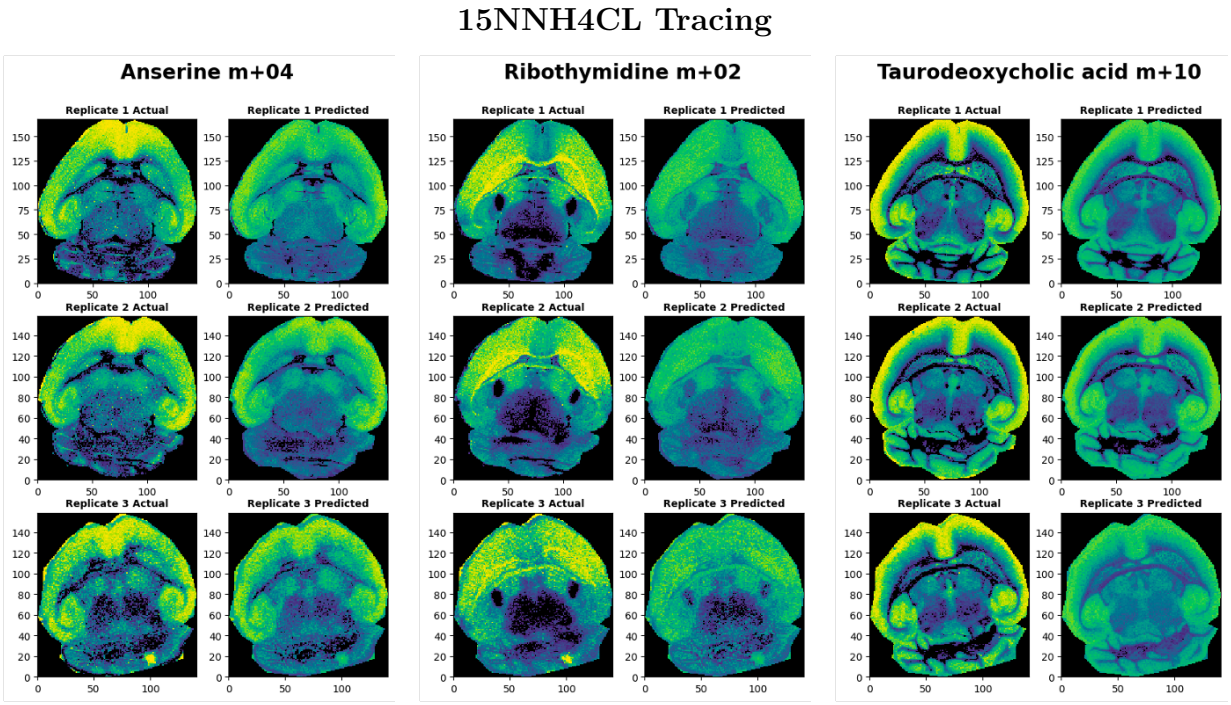
(a) Taurodeoxycholic acid m+8 (b) Cyclic-ADP-ribose m+14 (c) GMP m+10

Figure 5.12: Cross validation results for top predicted isotopologues in 3HB labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate.



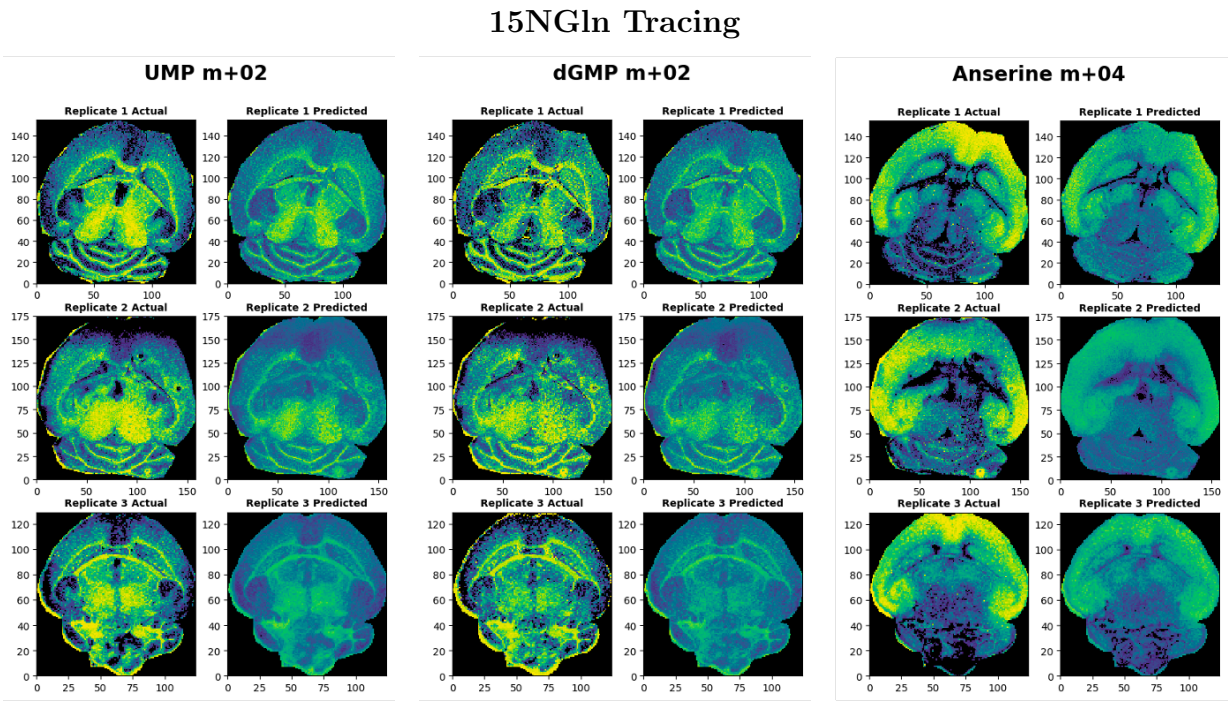
(a) (b) (c)

Figure 5.13: Cross validation results for top predicted isotopologues in 15NLeu labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate.



(a) (b) (c)

Figure 5.14: Cross validation results for top predicted isotopologues in 15NNH4CL labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate.



(a) (b) (c)

Figure 5.15: Cross validation results for top predicted isotopologues in 15NGln labeled isotope tracing brain data. For each replicate (row), the model was trained on the other 5 replicates and then tested on this holdout replicate.

References

[1] Michaela Aichler and Axel Walch. MALDI imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory Investigation*, 95(4):422–431, April 2015.

[2] D. Camacho, P. Mendes, and A. de la Fuente. Modelling and simulation for metabolomics data analysis. *Biochemical Society Transactions*, 33(6):1427, December 2005.

[3] Alice Cambiaghi, Manuela Ferrario, and Marco Masseroli. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Briefings in Bioinformatics*, page bbw031, April 2016.

[4] Richard M. Caprioli, Terry B. Farmer, and Jocelyn Gile. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry*, 69(23):4751–4760, December 1997.

[5] Navdeep Chandel. *Navigating Metabolism*. Cold Spring Harbor Laboratory Press, New York, NY, December 2014.

[6] Guilin Chen, Minxia Fan, Ye Liu, Baoqing Sun, Meixian Liu, Jianlin Wu, Na Li, and Mingquan Guo. Advances in MS based strategies for probing ligand-target interactions: Focus on soft ionization mass spectrometric techniques. *Frontiers in Chemistry*, 7, October 2019.

[7] Sören-Oliver Deininger, Dale S. Cornett, Rainer Paape, Michael Becker, Charles Pineau, Sandra Rauser, Axel Walch, and Eryk Wolski. Normalization in maldi-tof imaging datasets of proteins: practical considerations. *Analytical and Bioanalytical Chemistry*, 401(1):167–181, Jul 2011.

[8] Luca Gerosa, Bart R.B. Haverkorn van Rijsewijk, Dimitris Christodoulou, Karl Kochanowski, Thomas S.B. Schmidt, Elad Noor, and Uwe Sauer. Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data. *Cell Systems*, 1(4):270–282, October 2015.

[9] Ilya Gertsman and Bruce A. Barshop. Promises and pitfalls of untargeted metabolomics. *Journal of Inherited Metabolic Disease*, 41(3):355–366, March 2018.

[10] Robert Heise, Alisdair R. Fernie, Mark Stitt, and Zoran Nikoloski. Pool size measurements facilitate the determination of fluxes at branching points in non-stationary metabolic flux analysis: the case of arabidopsis thaliana. *Frontiers in Plant Science*, 6, 2015.

[11] Cholsoon Jang, Li Chen, and Joshua D. Rabinowitz. Metabolomics and isotope tracing. *Cell*, 173(4):822–837, May 2018.

[12] Thorsten W. Jaskolla and Michael Karas. Compelling evidence for lucky survivor and gas phase protonation: The unified MALDI analyte protonation mechanism. *Journal of the American Society for Mass Spectrometry*, 22(6):976–988, March 2011.

[13] Heesoo Jeong, Yan Yu, Henrik J. Johansson, Frank C. Schroeder, Janne Lehtiö, and Nathaniel M. Vacanti. Correcting for naturally occurring mass isotopologue abundances in stable-isotope tracing experiments with polymid. *Metabolites*, 11(5), 2021.

[14] Michael Karas and Ralf Krüger. Ion formation in MALDI: the cluster ionization mechanism. *Chemical Reviews*, 103(2):427–440, January 2003.

[15] Kyongbum Lee, Francois Berthiaume, Gregory N. Stephanopoulos, and Martin L. Yarmush. Metabolic flux analysis: A powerful tool for monitoring tissue function. *Tissue Engineering*, 5(4):347–368, August 1999.

[16] Hajime Mizuno, Kazuki Ueda, Yuta Kobayashi, Naohiro Tsuyama, Kenichiro Todoroki, Jun Zhe Min, and Toshimasa Toyo'oka. The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics. *Biomed. Chromatogr.*, 31(1):e3864, January 2017.

[17] Jens Nielsen. It is all about MetabolicFluxes. *Journal of Bacteriology*, 185(24):7031–7035, December 2003.

[18] Jeremy L. Norris and Richard M. Caprioli. Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chemical Reviews*, 113(4):2309–2342, February 2013.

[19] Daniel R. Schmidt, Rutulkumar Patel, David G. Kirsch, Caroline A. Lewis, Matthew G. Vander Heiden, and Jason W. Locasale. Metabolomics in cancer research and emerging applications in clinical oncology. *CA: A Cancer Journal for Clinicians*, May 2021.

[20] Hayat Ali Shah, Juan Liu, Zhihui Yang, and Jing Feng. Review of machine learning methods for the prediction and reconstruction of metabolic pathways. *Frontiers in Molecular Biosciences*, 8, June 2021.

[21] Markus Stoeckli, Pierre Chaurand, Dennis E. Hallahan, and Richard M. Caprioli. Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, 7(4):493–496, April 2001.

[22] Sara Violante, Mirela Berisa, Tiffany H Thomas, and Justin R Cross. Stable isotope tracers for metabolic pathway analysis. *Methods Mol. Biol.*, 1978:269–283, 2019.

[23] Lin Wang, Xi Xing, Xianfeng Zeng, S. RaElle Jackson, Tara TeSlaa, Osama Al-Dalahmah, Laith Z. Samarah, Katharine Goodwin, Lifeng Yang, Melanie R. McReynolds, Xiaoxuan Li, Jeremy J. Wolff, Joshua D. Rabinowitz, and Shawn M. Davidson. Spatially resolved isotope tracing reveals tissue metabolic activity. *Nature Methods*, 19(2):223–230, Feb 2022.

[24] Daniel James Wilkinson. Historical and contemporary stable isotope tracer approaches to studying mammalian protein metabolism. *Mass Spectrometry Reviews*, 37(1):57–80, 2016.

[25] Jacob E. Wulff and Matthew W. Mitchell. A comparison of various normalization methods for LC/MS metabolomics data. *Advances in Bioscience and Biotechnology*, 09(08):339–351, 2018.

[26] Aihua Zhang, Hui Sun, Ping Wang, Ying Han, and Xijun Wang. Modern analytical techniques in metabolomics analysis. *Analyst*, 137(2):293–300, January 2012.

APPENDICES

Appendix A

Proof of Concept

Appendix B

Python Implementation

B.1 Libraries

B.2 Code

```
import pandas as pd
import numpy as np
import tensorflow as tf
from tensorflow.keras.layers import Conv2D, Dropout, MaxPool2D, Flatten, Add, Dense, Ac
```