Bachelor's Thesis

Machine Learning in Metabolomics

Satvik

11/12/2024

submitted to the Biological Sciences Department
at the BITS Pilani, Goa

Supervisors: Professor Rajesh Mehrotra
BITS Pilani
Goa

BITS Pilani
K K Birla Goa Campus

BITS Pilani, Goa

Sancoale, Goa
India

Satvik
2021A8TS3047G
Machine Learning in Metabolomics
11/12/2024

Contents

Abstract

In this research, I reviewed the existing literature about the application of deep learning techniques like convolutional neural networks (CNNs), graph neural networks (GNNs) and many others. These methods were investigated in order to reveal intricate correlations and patterns present in metabolomics datasets. I created the open-source package *mbSTATS* to expedite the analysis process and allow academics to efficiently conduct early data evaluations. The library facilitates accessible and reproducible research workflows by offering a full array of tools for normalizing, analyzing, and visualizing metabolomics data.

Applying this framework to experimental data from *Arabidopsis thaliana* seeds, I identified a batch effect in one of the samples, which was significantly influencing the results. After isolating and removing this sample, further analysis highlighted multiple enriched metabolic pathways. These pathways, validated through a review of existing literature, indicated that the overexpressed seeds were experiencing stress conditions. This observation aligns with known responses of *Arabidopsis thaliana* to environmental and physiological stressors.

The study demonstrates the potential of combining computational approaches with biological research to derive meaningful insights from complex datasets. By integrating cutting-edge deep learning techniques with tools like *mbSTATS*, this work not only provides a practical solution for metabolomics data analysis but also contributes to understanding stress responses in plant systems. You can find the code for the *mbSTATS* library at `https://github.com/Satvik713/mbSTATS.git`

Acknowledgement

I would like to express my gratitude to Prof. Rajesh Mehrotra for giving me the opportunity to work on my thesis at the BITS Pilani, Goa. His guidance and insightful discussions were invaluable in navigating this complex topic.

A special thanks goes to PhD students Arti Karamchandani and Yukti Singh, whose unwavering support, creative ideas, and weekly discussions significantly shaped this project. I learned a great deal from them. Lastly, I would like to acknowledge my colleagues at BITS for helping me throughout the semester, as well as my family and friends for their moral support.

1

Introduction

Metabolomics is the comprehensive study of small molecules, commonly referred to as metabolites, within a biological system. These metabolites represent the end products of cellular processes and provide a snapshot of the physiological state of an organism. Metabolomics is a powerful tool for understanding biochemical pathways and their regulation in health, disease, and environmental interactions. By analyzing metabolite profiles, researchers can gain insights into various biological processes, identify potential biomarkers for diseases, and explore the effects of genetic modifications, environmental changes, or pharmaceutical treatments. Applications of metabolomics span diverse fields, including medicine, agriculture, and environmental science. Metabolomics in plant biology, is used to study stress responses, and crop improvement.

1.1 Data Analysis Tools

The large and complex datasets generated by metabolomics experiment require sophisticated data analysis tools for preprocessing, normalization, statistical evaluation, and biological interpretation. Several tools and platforms have been developed to address these needs like MZmine [SHK+23], Metaboanalyst [PZE+22], XCMS [DMI+18].

While these tools provide essential capabilities, they often require expertise in bioinformatics and programming. Researchers often lack expertise in both biology and programming to seamlessly use these tools. Recognizing this gap, new tools like mbSTATS aim to simplify preliminary data analysis and make these workflows more accessible to researchers. By combining data preprocessing, statistical analysis, and visualization in a single platform, mbSTATS and similar tools contribute to enhancing metabolomics research efficiency.

1.2 Machine learning in Metabolomics

The application of machine learning (ML) in metabolomics have been used to address challenges such as high-dimensional data, noise, batch effects, pathway prediction, metabolite identification, feature extraction. Recent developments have seen a shift towards deep learning models like convolutional neural networks (CNNs) [ON15] and graph neural networks (GNNs)[ZCH+18], which offer superior performance in handling the complexity of metabolomics data. I have discussed three recent papers below that use multiple deep learning models. The details about the techniques are explained later in materials and methods section.

[TY23]

The paper introduces a deep learning framework meant for untargeted metabolomics data analysis, addressing the critical issue of matching uncertainty between data features and known metabolites. Traditional approaches rely on mass-to-charge ratio (m/z) matching, which often results in ambiguous feature-to-metabolite relationships due to shared molecular compositions and adduct ions. To overcome this, the proposed model incorporates a gradual sparsification neural network designed to reflect the modular structure of biological systems and the inherent annotation relationships. The framework simultaneously achieves three key objectives: evaluating metabolite importance, inferring feature-metabolite matching likelihood, and selecting disease sub-networks.

[DYW+24]

The paper presents DeepMSProfiler, a deep learning-based method designed for untargeted metabolomics data analysis, tackling challenges such as complex data processing, batch variability, and unidentified metabolites. DeepMSProfiler enables end-to-end analysis of raw mass spectrometry signals, providing highly accurate and reliable outputs. DeepMSProfiler's interpretability feature offers insights into disease-related metabolite-protein networks, paving the way for both disease diagnosis and mechanistic discoveries.

[AFJ+22]

This review paper discusses the growing role of network-based approaches in metabolomics, particularly for analyzing untargeted mass spectrometry (MS) data. Untargeted metabolomics enables comprehensive detection of metabolites, uncovering unexpected metabolic changes and novel compounds. However, there are difficulties in analyzing this enormous information, especially when it comes to biological insights and metabolite identification.

The authors suggest using networks to address this, with nodes standing in for metabolites or attributes and edges denoting other kinds of interactions, such chemical similarities, metabolic pathways, or statistical correlations. By facilitating activities like detecting clusters of co-regulated metabolites or proposing potential metabolite identifications based on enzymatic processes, these networks act as tools for organizing complex metabolomics data.

1.3 Goal of this thesis

The primary goal of this thesis is to explore the integration of machine learning techniques and metabolomics for improved data analysis and gaining important biological insights. Firstly, I have done a comprehensive review of the recent state-of-the-art methods currently employed in metabolomics. I have also explained in detail the deep learning techniques used in the papers reviewed as a part of this thesis. Building on this foundation, the study introduces mbSTATS, an open-source library developed for the preliminary analysis of metabolomics data obtained from experimental workflows. This library facilitates processes like normalization, visualization, and statistical assessment, providing a user-friendly platform for researchers. Utilizing the outputs generated by mbSTATS on the data acquired from a metabolomics experiment at my university, the analysis identified biologically relevant pathways, revealing that overexpressed Arabidopsis thaliana seeds were subjected to stress conditions, a finding consistent with prior literature. Furthermore, the library proved instrumental in recognizing and addressing batch effects in the dataset, ensuring the reliability of subsequent analyses. This work emphasizes the synergy between computational tools and biological research, offering a pathway for improved data interpretation and enhanced understanding of metabolic processes.

2

Theoretical Background

2.1 Metabolomics

Metabolomics experiments aim to comprehensively profile small molecules (metabolites) in biological samples. The process typically begins with sample preparation, where biological specimens such as plant extracts are processed to extract metabolites using solvents like methanol or water. The extracted metabolites are then analyzed using high-throughput techniques such as mass spectrometry (MS) or nuclear magnetic resonance (NMR) spectroscopy. For MS-based metabolomics, the samples are ionized and introduced into the mass spectrometer, which separates metabolites based on their mass-to-charge ratio (m/z). Coupling MS with chromatography techniques like gas chromatography (GC) or liquid chromatography (LC) further enhances separation and detection, ensuring high sensitivity and specificity. LC-MS is particularly popular for its ability to handle diverse chemical properties of metabolites, while GC-MS is suited for analyzing volatile compounds. The raw data from MS instruments are typically stored in file formats like mzML or mzXML, which contain detailed information about detected peaks, including m/z values, retention times, and intensity. These formats are open and widely compatible with data processing tools. The data analysis workflow involves several key steps:

Peak Alignment

Since retention times of metabolites can vary slightly across samples, alignment ensures that identical peaks (representing the same metabolites) are matched. This step corrects retention time shifts and aligns the data across multiple samples for consistency.

Feature Extraction

Features are defined as unique m/z-retention time pairs representing potential metabolites. Algorithms deconvolute overlapping peaks and quantify the intensity of each feature, which correlates with the metabolite abundance.

Metabolite Identification

This stage entails correlating characteristics with identified metabolites in reference databases (e.g., HMDB, METLIN) according to their m/z values, retention times, and fragmentation patterns. Challenges such as isobaric compounds and adduct forms can hinder identification, frequently necessitating tandem mass spectrometry (MS/MS) for enhanced precision.

This method converts raw mass spectrometry data into a structured collection of metabolite characteristics, facilitating subsequent statistical and biological analysis. Instruments such as mzMine, XCMS, and OpenMS streamline these procedures by providing automated workflows for the management of raw data, peak detection, alignment, and annotation. These processed datasets serve as the foundation for pathway analysis and the creation of biological insights.

2.2 Deep Learning Algorithms

Graph neural networks

Graph Neural Networks (GNNs) are specialized deep learning frameworks engineered to analyze data structured as graphs. In contrast to conventional convolutional neural networks (CNNs), which are tailored for structured data such as images, graph neural networks (GNNs) address the distinct issues presented by graph data, including irregular node connection and the heterogeneous number of neighbors for each node. By applying the convolution idea to graphs, Graph Neural Networks (GNNs) revise node embeddings according to the attributes of neighboring nodes, facilitating the acquisition of intricate relationships within the graph framework. Diverse GNN architectures have been created, each designed for particular needs such as managing heterogeneous graphs or enhancing computing efficiency, rendering them adaptable instruments for tasks including social network analysis, chemical property prediction, and recommendation systems.

[AFJ+22] paper has designed a sparse neural network model to analyze untargeted metabolomics data while addressing the challenges of feature-metabolite matching uncertainty and leveraging known metabolic network structures.

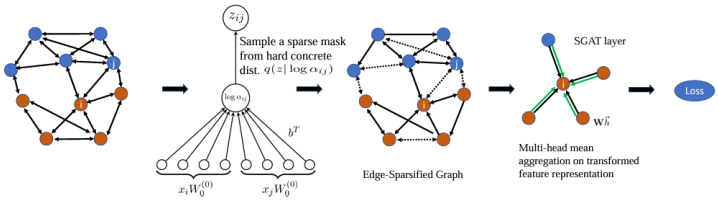

Figure 2.1: The overview of SGATs.

2.2.1 Model Architecture

The model consists of several structured layers designed to progressively integrate biological knowledge, enforce sparsity, and improve interpretability and efficiency. These layers are outlined below:

2.2.2 Input Layer: Feature Abundance Matrix

The input $\mathbf{X} \in \mathbb{R}^{n \times p}$ represents n samples and p features, where each feature corresponds to a detected ion in the metabolomics data. Features are connected to potential metabolites through an annotation matrix $\mathbf{M}$.

2.2.3 Matching-Embedding Layer

This is the first hidden layer with neurons corresponding one-to-one to metabolites. The connections between features and metabolites are encoded by the matching matrix $\mathbf{M} \in \mathbb{R}^{p \times m}$, where:

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{if feature } x_i \text{ matches metabolite } v_j \\ 0 & \text{otherwise} \end{cases}$$

The activation $\mathbf{Z}_1$ of this layer is computed as:

$$\mathbf{Z}_1 = \sigma\left(\mathbf{X} \odot \mathbf{W}_1 \cdot \mathbf{M} + \mathbf{b}_1\right)$$

where σ is the activation function, $\odot$ denotes element-wise multiplication, and $\mathbf{W}_1$ and $\mathbf{b}_1$ are trainable weights and biases.

2.2.4 Graph-Embedding Layer

This second hidden layer embeds the structure of the metabolic network represented by the graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ with nodes $\mathbf{V} = \{v_1, v_2, \ldots, v_m\}$ and edges $\mathbf{E}$. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ encodes metabolite connections:

$$\mathbf{A}_{i,j} = \begin{cases} 1 & \text{if } v_i \text{ and } v_j \text{ are connected in } \mathbf{G} \\ 0 & \text{otherwise} \end{cases}$$

The activation $\mathbf{Z}_2$ of this layer is computed as:

$$\mathbf{Z}_2 = \sigma\left(\mathbf{Z}_1 \cdot \mathbf{W}_2 \cdot \mathbf{A} + \mathbf{b}_2\right)$$

2.2.5 Gradual Sparsification Layers

Following the graph-embedding layer, the network applies sparsification to reduce computational complexity and enhance robustness. For each sparsified layer l, only neurons corresponding to metabolites with a degree exceeding a threshold μ in $\mathbf{G}$ are retained. The updated adjacency matrix $\tilde{\mathbf{A}}_l$ reflects these changes:

$$\tilde{\mathbf{A}}_l = \begin{cases} 1 & \text{if the degree of } v_i \text{ exceeds } \mu \\ 0 & \text{otherwise} \end{cases}$$

The activation $\mathbf{Z}_{l+1}$ in a sparsified layer is:

$$\mathbf{Z}_{l+1} = \sigma\left(\mathbf{Z}_l \cdot \mathbf{W}_{l+1} \cdot \tilde{\mathbf{A}}_l + \mathbf{b}_{l+1}\right)$$

2.2.6 Fully Connected Layers and Output

After sufficient sparsification, fully connected layers convert the network into a dense representation for prediction. The final layer uses a softmax activation function to output class probabilities:

$$f(\mathbf{X}) = \text{softmax}\left(\mathbf{Z}_L \cdot \mathbf{W}_L + \mathbf{b}_L\right)$$

2.2.7 Optimization

The model parameters (weights $\mathbf{W}$, biases $\mathbf{b}$) are optimized by minimizing a cross-entropy loss function:

$$L = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{i,k} \log(f(\mathbf{X}_{i,k}))$$

where $y_{i,k}$ is the true label for sample i in class k, and $f(\mathbf{X}_{i,k})$ is the predicted probability.

The optimization process uses the Adam optimizer, which dynamically adjusts learning rates to ensure efficient convergence.

2.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are deep learning models designed to process data with a grid-like structure, such as images, videos, or spatial data. CNNs are composed of several key components: convolutional layers, pooling layers, and fully connected layers.

2.3.1 Convolutional Layer

The convolution operation in a CNN is defined mathematically as:

$$(f * g)(x) = \int f(t)g(x - t)\, dt$$

where f is the input signal, and g is the kernel or filter. In discrete terms, this operation becomes:

$$(f * g)(i, j) = \sum_m \sum_n f(i + m, j + n)g(m, n)$$

In CNNs, filters g slide over the input f, performing element-wise multiplications and summing the results. This operation helps capture localized patterns such as edges, textures, and features at various spatial scales.

2.3.2 Activation Function

Activation functions introduce non-linearity into the network, enabling it to learn more complex patterns. A common activation function used in CNNs is the Rectified Linear Unit (ReLU), defined as:

$$\text{ReLU}(x) = \max(0, x)$$

ReLU allows the network to model non-linear relationships, enhancing the expressiveness of the model.

2.3.3 Pooling Layer

Pooling layers reduce the spatial dimensions of feature maps, which decreases computational complexity and introduces invariance to minor translations in the input. The most commonly used pooling operation is max-pooling, which selects the maximum value from a set of values in a local neighborhood:

$$p_{i,j} = \max\{f(x, y) \mid (x, y) \in \text{receptive field of } (i, j)\}$$

This process helps retain the most important features while reducing the size of the data.

2.3.4 Fully Connected Layer

After feature extraction through convolution and pooling, the network uses fully connected layers to perform classification or regression. In these layers, the features are mapped to output probabilities using:

$$y = \sigma(Wx + b)$$

where W is the weight matrix, x is the input vector, b is the bias term, and σ is an activation function, such as softmax, used for classification tasks.

2.3.5 Advantages of CNNs

Convolutional Neural Networks (CNNs) excel in tasks related to visual data, including picture recognition and object detection, owing to their capacity to autonomously learn hierarchical features and capture spatial hierarchies. The convolutional operation enables CNNs to achieve computational efficiency by applying the same filter over various parts of the input. The implementation of weight sharing decreases the parameter count relative to fully linked networks, hence enhancing the scalability of CNNs.

2.4 DenseNet121: A Specialized CNN Architecture

DenseNet121, or Dense Convolutional Network with 121 layers [HLMW16], introduces a unique connectivity pattern where each layer receives inputs from all preceding layers. This design has several advantages:

2.4.1 Dense Connections

For layer l, the input is a concatenation of feature maps from all previous layers:

$$x_l = H_l\left([x_0, x_1, \ldots, x_{l-1}]\right)$$

where H_l is a composite function of batch normalization (BN), ReLU activation, and convolution.

2.4.2 Efficient Parameter Usage

Instead of learning redundant features, DenseNet promotes feature reuse. The number of parameters is reduced significantly compared to traditional CNNs.

2.4.3 Improved Gradient Flow

Dense connectivity mitigates the vanishing gradient problem, as gradients flow directly through the network via shortcut paths.

2.4.4 Feature Extraction

DenseNet121 consists of densely connected blocks separated by transition layers. The transition layers use convolution and pooling to compress feature maps, reducing computational cost while preserving essential features.

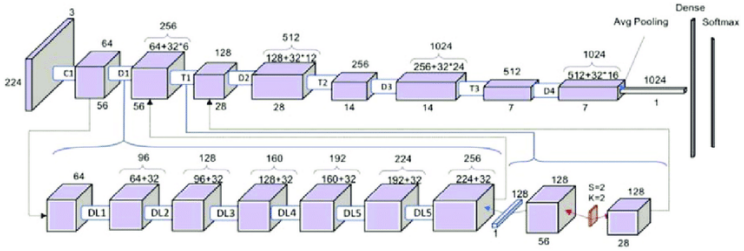

Figure 2.2: DenseNet Architecture. This network improves image classification accuracy by using dense connections between layers.

2.5 Experimental networks in metabolomics

Experimental networks are important tools created from untargeted metabolomics data. They provide insights into compound relationships and transformations by analyzing several features of mass spectrometry (MS) data, such as MS1, MS2, and MSn. These networks provide complementary techniques to deciphering metabolic linkages and are classified according to their focus: mass differences, adducts and characteristics, structural similarities, and correlation data.

Mass difference networks

These networks concentrate on the metabolic changes that manifest as variations in molecular ion masses. They are predicated on the idea that physiological events entail predictable atom gains or losses, which alter molecular formulas and, in turn, vary precise mass. Nodes in the MS data represent features, while edges show particular mass differences that correspond to recognized metabolic reactions.

Adducts and feature networks

During MS analysis, these networks deal with non-biological mass differences brought on by physicochemical processes. For instance, changes in observed m/z values are caused by isotopes, ion adducts, and in-source fragments. To increase the accuracy of feature annotation, these artifacts can be aggregated and deconvoluted.

Structure similarity networks

These networks, which are based on MS2 fragmentation data, link metabolites according to their chemical similarity. Common substructures or fragmentation patterns are frequently used to infer structural similarity, which may indicate metabolic connections between different molecules.

Correlation networks

These networks utilize abundance patterns to elucidate causal links among metabolites. Correlated variations in metabolite levels across experimental circumstances may indicate common routes or regulatory mechanisms.

The integration of network analysis and multi-layer networks in metabolomics improves comprehension by utilizing the complementing characteristics of various network types. The integration of spectral similarity networks with chemical ontologies and mass difference networks enhances metabolite annotations. ChemRICH associates metabolic structures with classifications that offer greater breadth than conventional approaches, whereas MolNetEnhancer amalgamates molecular networks with ontologies and computational methodologies to furnish more comprehensive annotations, even in the absence of prior library correspondences. FT-BLAST and iMet manage unidentified metabolites by employing fragmentation trees and integrating mass discrepancies with spectral similarity networks. Correlating metabolite abundances enhances validation, as metabolites associated through biological events frequently exhibit significant spectrum similarities and correlations. Genome-scale metabolic networks (GSMNs) improve interpretation by validating linkages and elucidating pathways. Untargeted metabolomics addresses deficiencies in Genome-Scale Metabolic Networks (GSMNs) by detecting absent metabolites and pathways. Improvements in structural curation enable the integration of metabolomics data, guaranteeing dependable biological context. This comprehensive method enhances annotation precision and enriches biological understanding.

2.5.1 Machine Learning tools in mbSTATS

mbSTATS presently facilitates normalization and visualization using many plots, including volcano plots, p-value plots, PCA plots, and HCA plots, among others. I have attempted to elucidate these plots to assist others in their interpretation.

P-value plot

A p-value graphic depicts the statistical significance of characteristics across different situations or groups. The graphic illustrates the p-values from hypothesis testing (e.g., t-tests, ANOVA) for each compound, signifying the likelihood that the observed data is due to random chance. In metabolomics, p-value plots aid in identifying molecules that exhibit significant changes between experimental conditions. A low p-value (often < 0.05) signifies a substantial association with the examined conditions, implying that the metabolite may be essential in the biological process. This figure aids in the identification of critical biomarkers.

Volcano plot

A volcano plot combines statistical significance (p-value) and fold change into a unified representation. The x-axis represents the magnitude of the fold change (e.g., log2 transformed), whereas the y-axis depicts the negative logarithm of the p-value. Metabolites situated in the upper right or left quadrants that demonstrate considerable fold changes and low p-values are considered the most significantly different among groups. This plot is essential for rapidly finding chemicals of considerable biological significance, often employed in biomarker identification within metabolomics research.

PCA plot

The Principal Component Analysis (PCA) visualization reduces the dimensionality of complex metabolomics data while maintaining maximum variance. The first two or three principal components are depicted to highlight the distribution of samples based on metabolic traits. Principal Component Analysis (PCA) in metabolomics enhances the display of sample clusters, the detection of potential outliers, and the understanding of the overall data structure. It is advantageous for evaluating data quality and discerning patterns in categories such as wild-type versus overexpressed or stressed versus control conditions.

PLS-DA plot

Partial Least Squares Discriminant Analysis (PLS-DA) is an effective supervised dimensionality reduction method commonly employed in metabolomics to improve the interpretability of intricate data while preserving predictive accuracy. In contrast to PCA, which operates in an unsupervised manner, PLS-DA integrates class labels (e.g., wild-type versus overexpressed, or stressed versus control) to optimize the differentiation across established groups. This strategy emphasizes essential metabolites that account for group differences, rendering it crucial for biomarker identification and comprehension of biological variability. PLS-DA is proficient in visualizing sample clusters, evaluating data quality, detecting outliers, and distinguishing unique metabolic processes, especially in datasets with nuanced or intricate group differences.

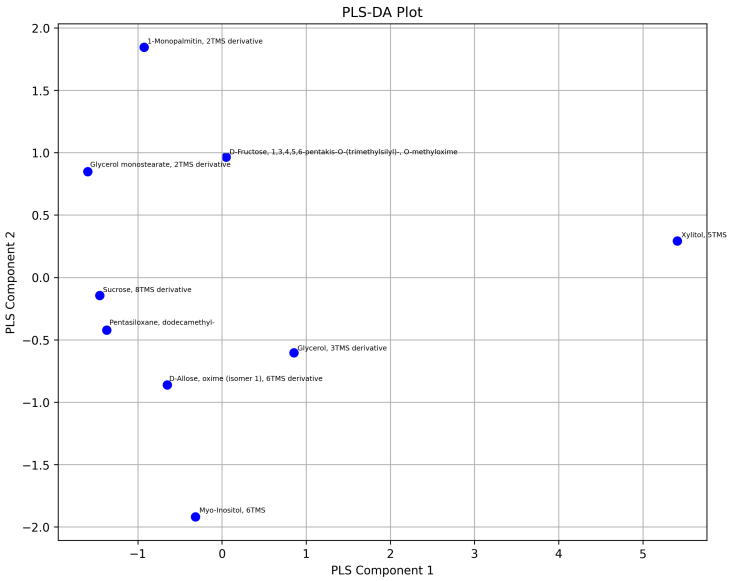

Figure 2.3: PLS-DA plot for the common compounds

HCA plot

Hierarchical Cluster Analysis (HCA) is a method utilized to classify similar samples or metabolites into clusters based on their proximity or dissimilarity. The resulting dendrogram or heatmap depicts these clusters, enabling a visual examination of the relationships among samples or features. In metabolomics, HCA plots enable the discovery of metabolite clusters displaying similar behavior across samples, hence clarifying probable biological pathways or metabolic conditions. The hierarchical structure enables the identification of primary categories, which may then be analyzed in further depth.

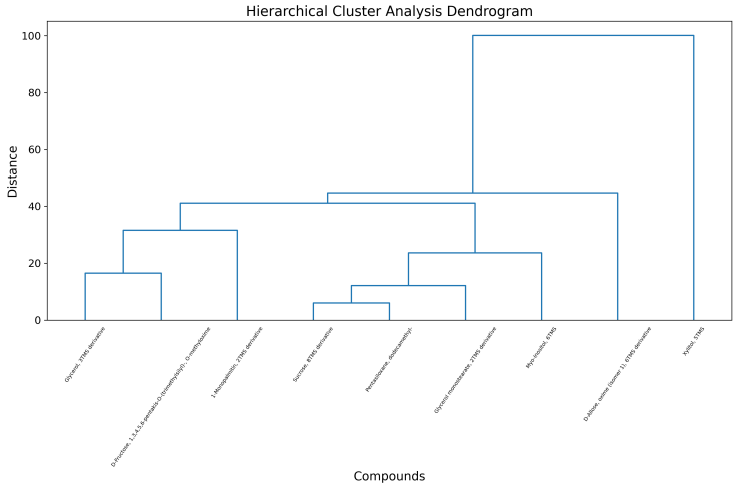

Figure 2.4: HCA Plot

Correlation matrix plot

A correlation matrix plot demonstrates the associations among metabolites by representing their correlation coefficients. Each cell in the matrix represents the relationship between two metabolites, with colors indicating the strength of the association (positive or negative). In metabolomics, such plots aid in identifying metabolites that may be co-regulated or have similar biological activities. Strong correlations may indicate that the metabolites are involved in the same metabolic route or biological activity, while weak or negative correlations suggest separate pathways.

Violin plot

The violin plot depicts the distribution and density of data points (e.g., metabolite concentrations) across multiple groups or circumstances. The plot combines features of a box plot with a kernel density plot, depicting the distribution's range, median, and potential multimodal traits. In metabolomics, violin charts proficiently depict the distribution and central tendency of metabolite concentrations across experimental groups. Differences in the shape or width of the violins among groups may indicate significant changes in metabolite levels, providing insights on biological variability and experimental conditions.

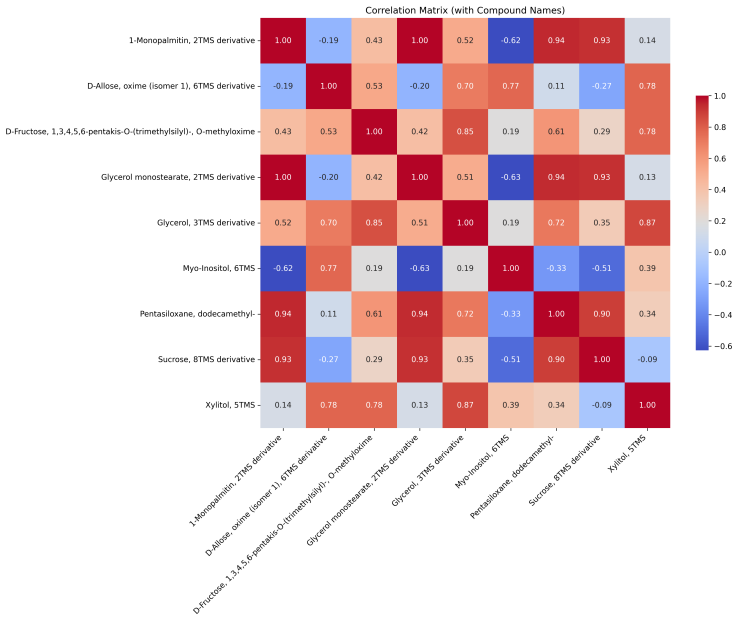

Figure 2.5: Correlation Matrix Plot

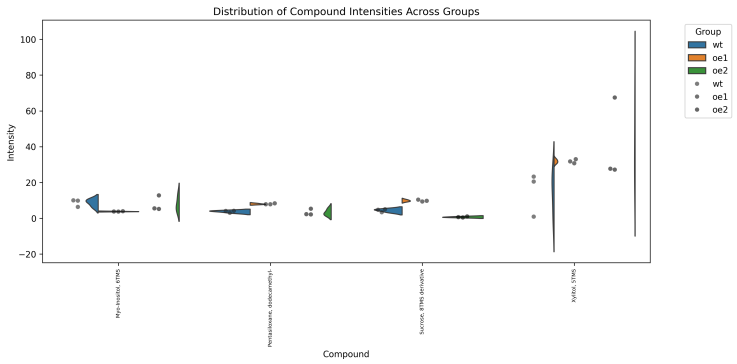

Figure 2.6: Violin plot representing the variation in density

3

Results and Discussion

Principal Component Analysis (PCA) is an effective method for detecting batch effects in metabolomics data. A separate cluster in the PCA plot frequently indicates the existence of non-biological variability influencing particular data. Batch effects, resulting from technical irregularities in sample preparation or measurement, might hide authentic biological differences and result in data misinterpretation. The elimination of a sample classified as an outlier due to batch effects markedly improved the results, augmenting the number of statistically significant metabolites and elucidating the metabolic differences across experimental groups. This underscores the imperative for robust preprocessing techniques, such as PCA, to guarantee data reliability and meaningful biological conclusions.

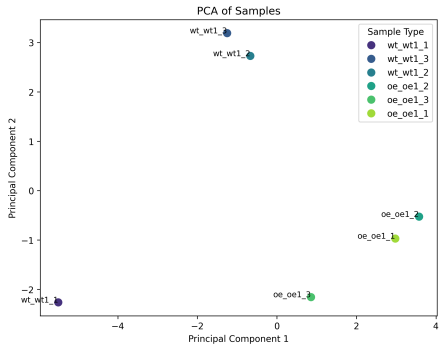

Figure 3.1: PCA plot of samples without batch effects

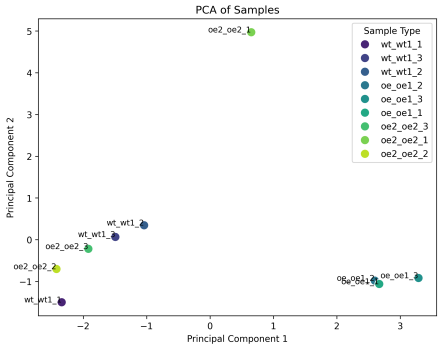

Figure 3.2: PCA plot of samples with batch effects

Following the application of significant filters via p-value plots, eight metabolites were identified: Propanediol, Monopalmitin, Trehalose, Dodecane, Glycerol Monostearate, Lactic Acid, Myo-inositol, and Sucrose. Subsequent

pathway analysis with MetaboAnalyst identified three essential pathways: starch and sucrose metabolism, galactose metabolism, and ascorbate and aldarate metabolism. Previous findings ([ZS22] and [HBP+20]) underscore the significance of two principal antioxidants that are activated in these pathways to safeguard plants against reactive oxygen species (ROS). These findings emphasize the significance of these metabolites in protecting plants from oxidative stress, hence reinforcing the physiological importance of the discovered pathways.

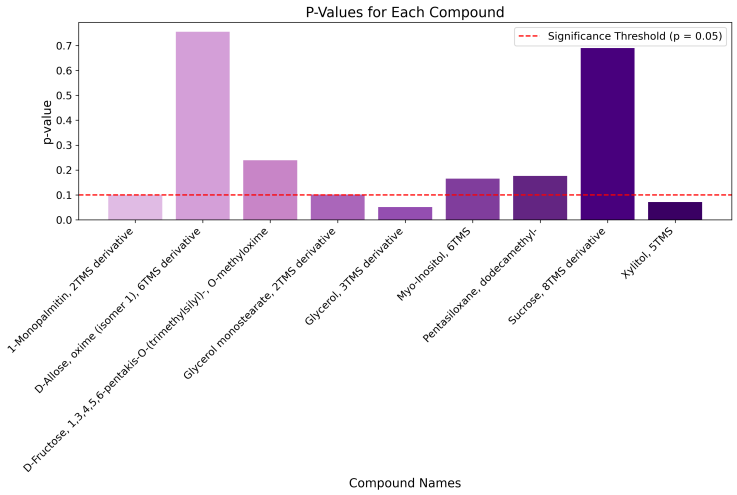

Figure 3.3: P value of the common metabolites from all the samples

The deficiency of three amino acids—Cysteine, Leucine, and Glycine—in overexpressed samples further substantiates stress conditions. Cysteine and Glycine are crucial for the formation of Glutathione (GSH), a process that is triggered to detoxify reactive oxygen species (ROS) during stress ([XO98] and [ZSKO13]. Analysis of overexpressed amino acids indicated enhanced pathways, including arginine biosynthesis, biosynthesis of alanine, aspartate, and glutamate (ALN, ASP, GLU), and lysine biosynthesis. Research indicates the roles of these pathways during stress: arginine biosynthesis is upregulated under stress [SM20], proline (derived from ALN, ASP, GLU biosynthesis) facilitates stress survival ([PLZ10] and [LZNB13], and aspartate metabolism improves drought response while acting as a precursor for lysine biosynthesis [LRH22]. These discoveries underscore the metabolic alterations suggestive of stress resilience systems in overexpressed plants.

The biologically significant findings of this investigation, along with the broader ramifications of the detected metabolic alterations, will be detailed in an upcoming article that is now in development.

4

Conclusion

4.1 Summary

This research employed metabolomics analysis to identify significant metabolites and amino acids that clarify essential metabolic pathways in *Arabidopsis thaliana* seed development. By synthesizing these findings with insights from the available literature, it was deduced that the overexpressed plants were undergoing stress conditions. These discoveries offer significant insights into the molecular foundations of plant stress resilience, which will be elaborated upon in a forthcoming paper.

This work's notable contribution is the creation of the open-source library *mbSTATS*. *mbSTATS* is an indispensable instrument for initial metabolomics data processing, facilitating the identification of batch effects, detection of important compounds, and examination of sample clustering. This library facilitates accessible and reproducible metabolomics analysis, enabling researchers to extract significant insights from their datasets.

Nonetheless, a shortcoming of the study was the inability to employ state-of-the-art (SOTA) deep learning techniques owing to the dataset's limited size. Although these methodologies frequently excel at detecting intricate patterns in extensive datasets, their utility is constrained by the restricted scope of smaller datasets, such as the one utilized in this study. Future study may investigate increasing dataset size or utilizing hybrid approaches to leverage deep learning capabilities, thereby augmenting the analytical strength of metabolomics studies.

4.2 Future Work

Future efforts will concentrate on augmenting the *mbSTATS* library through the use of sophisticated deep learning techniques to enhance its functionality in metabolomics data analysis. Efforts will focus on developing deep learning models tailored for small-scale datasets, mitigating existing constraints in data availability. A comprehensive benchmarking system will be established to assess and contrast cutting-edge deep learning methodologies. This system will guarantee the dependability, repeatability, and efficacy of various methodologies, fostering innovation and setting a new benchmark in metabolomics research.

List of Figures

Bibliography

[AFJ+22] Adam Amara, Clément Frainay, Fabien Jourdan, Thomas Naake, Steffen Neumann, Elva María Novoa-del-Toro, Reza M Salek, Liesa Salzer, Sarah Scharfenberg, and Michael Witting: *Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation*. In *Frontiers in Molecular Biosciences*, volume 9, 2022. DOI: 10.3389/fmolb.2022.841373. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2022.841373/full.

[DYW+24] Yongjie Deng, Yao Yao, Yanni Wang, Tiantian Yu, Wenhao Cai, Dingli Zhou, Feng Yin, Wanli Liu, Yuying Liu, Chuanbo Xie, Jian Guan, Yumin Hu, Peng Huang, and Weizhong Li: *An end-to-end deep learning method for mass spectrometry data analysis to reveal disease-specific metabolic profiles*. In *Nature Communications*, volume 15, 2024. DOI: 10.1038/s41467-024-35124-6. URL: https://www.nature.com/articles/s41467-024-35124-6.

[DMI+18] Xavier Domingo-Almenara, J. Rafael Montenegro-Burke, Julijana Ivanisevic, Aurelien Thomas, Jonathan Sidibé, Tony Teav, Carlos Guijas, Aries E. Aisporna, Duane Rinehart, Linh Hoang, Anders Nordström, María Gómez-Romero, Luke Whiley, Matthew R. Lewis, Jeremy K. Nicholson, H. Paul Benton, and Gary Siuzdak: *XCMS-MRM and METLIN-MRM: A Cloud Library and Public Resource for Targeted Analysis of Small Molecules*. In *Nature Methods*, volume 15, 2018, pages 681–684. DOI: 10.1038/s41592-018-0112-5.

[HBP+20] Mirza Hasanuzzaman, M. H. M. Borhannuddin Bhuyan, Khursheda Parvin, Tasnim Farha Bhuiyan, Taufika Islam Anee, Kamrun Nahar, Md. Shahadat Hossen, Faisal Zulfiqar, Md. Mahabub Alam, and Masayuki Fujita: *Regulation of ROS Metabolism in Plants under Environmental Stress: A Review of Recent Experimental Evidence*. In *International Journal of Molecular Sciences*, volume 21 (22), 2020. Submission received: 1 September 2020 / Revised: 14 November 2020 / Accepted: 17 November 2020 / Published: 18 November 2020, pages 8695. DOI: 10.3390/ijms21228695. URL: https://doi.org/10.3390/ijms21228695.

[HLMW16] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger: *Densely Connected Convolutional Networks*. In *arXiv preprint arXiv:1608.06993*, volume v5, 2016. CVPR 2017. DOI: 10.48550/arXiv.1608.06993. URL: https://doi.org/10.48550/arXiv.1608.06993.

[LRH22] Shuhan Lei, Stephanie Rossi, and Bingru Huang: *Metabolic and Physiological Regulation of Aspartic Acid-Mediated Enhancement of Heat Stress Tolerance in Perennial Ryegrass*. In *Plants*, volume 11 (2), 2022. Published: 13 January 2022, pages 199. DOI: 10.3390/plants11020199. URL: https://doi.org/10.3390/plants11020199.

[LZNB13] Xinwen Liang, Lu Zhang, Sathish Kumar Natarajan, and Donald F. Becker: *Proline Mechanisms of Stress Survival*. In *Antioxidants Redox Signaling*, 2013. Published Online: 29 August 2013. DOI: 10.1089/ars.2012.507. URL: https://doi.org/10.1089/ars.2012.507.

[ON15] Keiron O'Shea and Ryan Nash: *An Introduction to Convolutional Neural Networks*. In *arXiv preprint arXiv:1511.08458*, 2015. Submitted on 26 Nov 2015 (v1), last revised 2 Dec 2015 (v2). arXiv: 1511.08458 [cs.NE]. URL: https://arxiv.org/abs/1511.08458.

[PZE+22] Zhiqiang Pang, Guangyan Zhou, Jessica Ewald, Le Chang, Orcun Hacariz, Niladri Basu, and Jianguo Xia: *Using MetaboAnalyst 5.0 for LC–HRMS Spectra Processing, Multi-Omics Integration, and Covariate Adjustment of Global Metabolomics Data*. In *Nature Protocols*, volume 17, 2022, pages 1735–1761. DOI: 10.1038/s41596-022-00694-5.

[PLZ10] James M. Phang, Wei Liu, and Olga Zabirnyk: *Proline Metabolism and Microenvironmental Stress*. In *Annual Review of Nutrition*, volume 30, 2010. First published as a Review in Advance on April 23, 2010, pages 441–463. DOI: 10.1146/annurev.nutr.012809.104638. URL: https://doi.org/10.1146/annurev.nutr.012809.104638.

[SHK+23] Robin Schmid, Steffen Heuckeroth, Ansgar Korf, Aleksandr Smirnov, Owen Myers, Thomas S. Dyrlund, Roman Bushuiev, Kevin J. Murray, Nils Hoffmann, Miaoshan Lu, Abinesh Sarvepalli, Zheng Zhang, Markus Fleischauer, Kai Dührkop, Mark Wesner, Shawn J. Hoogstra, Edward Rudt, Olena Mokshyna, Corinna Brungs, Kirill Ponomarov, Lana Mutabdžija, Tito Damiani, Chris J. Pudney, Mark Earll, and Tomáš Pluskal: *Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3*. In *Nature Biotechnology*, volume 41, 2023, pages 447–449. DOI: 10.1038/s41587-023-01681-0.

[SM20] Shiva Siddappa and Gopal Kedihithlu Marathe: *What we know about plant arginases?* In *Plant Physiology and Biochemistry*, volume 156, 2020, pages 600–610. DOI: 10.1016/j.plaphy.2020.10.002. URL: https://doi.org/10.1016/j.plaphy.2020.10.002.

[TY23] Leqi Tian and Tianwei Yu: *An integrated deep learning framework for the interpretation of untargeted metabolomics data*. In *Briefings in Bioinformatics*, volume 24 (4), 2023, pages bbad244. DOI: 10.1093/bib/bbad244. URL: https://doi.org/10.1093/bib/bbad244.

[XO98] Chengbin Xiang and David J. Oliver: *Glutathione Metabolic Genes Coordinately Respond to Heavy Metals and Jasmonic Acid in Arabidopsis*. In *The Plant Cell*, volume 10 (9), 1998. Published: 01 September 1998, pages 1539–1550. DOI: 10.1105/tpc.10.9.1539. URL: https://doi.org/10.1105/tpc.10.9.1539.

[ZSKO13] Lyuben Zagorchev, Charlotte E. Seal, Ilse Kranner, and Mariela Odjakova: *A Central Role for Thiols in Plant Tolerance to Abiotic Stress*. In *International Journal of Molecular Sciences*, volume 14 (4), 2013. Submission received: 4 February 2013 / Revised: 28 February 2013 / Accepted: 14 March 2013 / Published: 2 April 2013, pages 7405–7432. DOI: 10.3390/ijms14047405. URL: https://doi.org/10.3390/ijms14047405.

[ZS22] Peiman Zandi and Ewald Schnug: *Reactive Oxygen Species, Antioxidant Responses and Implications from a Microbial Modulation Perspective*. In *Biology*, volume 11 (2), 2022. Submission received: 15 December 2021 / Revised: 14 January 2022 / Accepted: 17 January 2022 / Published: 18 January 2022, pages 155. DOI: 10.3390/biology11020155. URL: https://doi.org/10.3390/biology11020155.

[ZCH+18] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun: *Graph Neural Networks: A Review of Methods and Applications*. In *arXiv preprint arXiv:1812.08434*, 2018. Published at AI Open 2021. Submitted on 20 Dec 2018 (v1), last revised 6 Oct 2021 (v6). arXiv: 1812.08434 [cs.LG]. URL: https://arxiv.org/abs/1812.08434.