A Permutation-Based Approach Applied to Biologically Informed Neural Networks for Ontological Pathway Analysis

by

Matthew Kraljevic

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies at

the University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biochemistry and Medical Genetics

Individual Interdisciplinary Studies

University of Manitoba

Winnipeg

Lay Abstract

Congenital diaphragmatic hernia is a deadly condition which kills approximately 1 child every 15-30 minutes. This is a condition which does not have any viable drug interventions. Despite its lethality, it is still a rare condition. To find medical interventions, samples of this condition must be studied. Yet scientists that seek a cure to this condition face a stark reality - it is difficult to obtain samples for rare diseases.

The goal of the interdisciplinary field of bioinformatics is to apply technological and statistical methods to extract as much information as possible from what few precious samples we have. A common technique to discover drug interventions with bioinformatics is known as **pathway analysis**. This is a diverse set of tools which take genetic expression data measured in wet lab research and determine which processes (or “pathways”) in biology have been disrupted by a disease.

Since 2018, a family of techniques has been developing which incorporates the methods and the philosophy of science underpinning pathway analysis into machine learning models, called **Biologically Informed Neural Networks** (BINNs). This approach may offer an alternate way to perform pathway analysis, and does not suffer faults that exist in current techniques.

However, the best practices for applying BINNs to perform pathway analysis have not been established. We evaluated BINNs by comparing their results to existing techniques, and applied statistical methods to better improve the capabilities of BINNs and lay the groundwork for future work into BINNs.

Abstract

Introduction: Pathway analysis is a widespread method for determining actionable insight from high-throughput sequencing experiments. Traditional techniques have limitations, which novel machine learning methods seek to address. Biologically Informed Neural Networks (BINNs) are an emerging technique which applies ontological information built from existing literature, and explainable artificial intelligence metrics to perform biomarker discovery (genetic material of relevance to disease) and pathway analysis. Best practices for the implementation and interpretation of BINNs do not currently exist, and as such, we compare the results produced from a BINN to a traditional approach as a form of validation. We also examine the impact of pre-clustering data, and performing a permutation test on the output of BINNs.

Methods: Using a Python library which implements a BINN, and a traditional hypergeometric overrepresentation analysis, we compare the output of these methods on bulk RNA-sequencing cancer data accessed from the Cancer Cell Line Encyclopedia. Weighted Gene Co-expression Network Analysis was used to pre-cluster the RNA-sequencing data, and a “point estimate” null probability was produced using the explainable AI metrics and a permutation test, was implemented using R and Python.

Results: Using a ROC curve between our approach and the hypergeometric, we are able to achieve a high (>0.80) area under the curve in predicting statistically significant terms using the point estimate metric. This showed an improvement over other interpretations of the BINN output.

Conclusion: Interpretation of BINNs may be improved with a point estimate when considering a traditional, hypergeometric approach as a ground truth. We also observed interpretations of BINNs that use purely Deep SHAP importances show a low correlation with an existing pathway analysis approach.

Acknowledgements

.

Deepest thank you to my Mom and Dad.

and all my friends and family.

(full list in supplemental data)

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AKI	Acute kidney injury
ANOVA	Analysis of Variance
AUC	Area under curve
BINN	Biologically Informed Neural Network
CCLE	Cancer Cell Line Encyclopedia
CDH	Congenital diaphragmatic hernia
COVID-19	Coronavirus disease 2019
CircRNA	Circular ribonucleic acid
DAG	Directed Acyclic Graph
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DE	Differentially expressed
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
SHAP	SHapley Additive exPlanations
DNA	Deoxyribonucleic acid

FCS	Functional class scoring
FKPM	Fragments per kilobase of transcript per million fragments mapped
GO	Gene ontology
GSEA	Gene set enrichment analysis
IG	Integrated Gradients
ISO	International Organization for Standardization
ML	Machine Learning
NA	Not available
NGS	Next generation sequencing
OBO	Open Biological and Biomedical Ontologies
ORA	Over representation analysis
OWL	Web Ontology Language
PTB	Pathway topology based
ROC	Receiver operating characteristic
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
scRNA	Since cell ribonucleic acid

TMM	Trimmed mean of M-value
TOM	Topological overlap matrix
WGCNA	Weighted Gene Correlation Network Analysis
XAI	Explainable artificial intelligence

Chapter 1. Background

1.1 Translational medicine

Translational medicine is a branch of research focussed on translating knowledge “from bench to bedside”. It is an interdisciplinary field concerned with making actionable insights from biological research, in developing drug interventions, procedures and diagnostics (Wehling, 2021).

For many diseases, such as congenital diaphragmatic hernia (CDH), no drug interventions exist. This is a rare condition, yet accounts for approximately 1% of all infant mortality worldwide, with a death from CDH occurring approximately every 30 minutes (Schultz et al., 2007; World Bank Group, n.d.; World Health Organization, n.d.). Furthermore, this is a condition for which no drug interventions exist (Vallejo-Cremades et al., 2024). The role of translational medicine can be seen in the efforts for drug discovery through high throughput sequencing studies to find treatments for conditions such as CDH (Cannata et al., 2021).

1.2 Pathway analysis

Since the 1990s, the field of genetics has seen rapid development in sequencing technology, which has led to the rise of the omics fields (Kulski, 2016). Omics refers to the measurement of a complete set of genes, transcripts, proteins or other molecular entities that exist in an organism (Rogers, n.d.).

As these fields have developed, tools have emerged to grapple with the realities of processing vast amounts of data now accessible to researchers (Vitorino, 2024).

In their review, García-Campos et al. stated that given the high dimensionality of measurements, naïve approaches to analyzing even a single timepoint of the human genome

may take more time than has elapsed in the universe. By its very nature, omics research must face the phenomenon of “the curse of dimensionality”, the unique complexities which arise when the number of features measured in an analysis are much higher than the number of samples collected. Within omics, this has led to techniques in which prior biological information is integrated into analyses to better handle the complexity of data. This is implemented in a family of tools known as pathway analysis. (2015)

Pathway analysis, pathway enrichment analysis, or functional enrichment analysis, is a widespread bioinformatic method used to discover actionable insights from omic data and is widely adopted in translational medicine and other biological research (García-Campos et al., 2015; Wijesooriya et al., 2022). Pathway analysis is implemented in many commonplace tools to analyze high-throughput sequencing omics data. It refers to a family of analyses which identify biological processes and factors that differ between experimental phenotypes or conditions examined in an experiment (i.e. case versus control). Despite the name, these tools identify not only pathways, but also molecular and biological processes, or other elements of biology associated with differences in phenotype. This methodology has seen great success and widespread adoption. In a translational setting, pathway analysis has helped in designing novel cancer therapies, and in discovering changes in metabolic and signaling pathways, widely known as contributors to cancer progression (Folger et al., 2011; Hanahan & Weinberg, 2011).

While there are ongoing novel implementations of pathway analysis, some of the earliest implementations are still in common usage today. Through web portals and freely available software, tools such as the Database for Annotation, Visualization, and Integrated Discovery (DAVID) still see regular and recent updates from its initial release in 2003, where the original 2003 paper has been cited over 10,000 times, with 428 citations in 2025 as of writing (Dennis et al., 2003). Similarly, clusterProfiler, a library in the R programming language which performs pathway analysis, has had over 30,000 citations since its first release in 2012, with 4,981 in 2025 (Yu et al., 2012). GSEA is another ubiquitous tool in the family of pathway analysis with its

original manuscript having over 50,000 citations over 20 years and over 5000 citations this year (Subramanian et al., 2005).

1.2.1 Overview of Omics Data

In order to apply analyses, omics data from a high-throughput sequencer must go through a series of bioinformatic processes and mapping. What is produced and ultimately analyzed is a table or spreadsheet. Each row is identified by a gene identifier which corresponds to the gene associated with the molecular entity, followed by a numeric value of the expression data for each sample measured. An illustration of typical omics data can be seen in Figure 1.1.

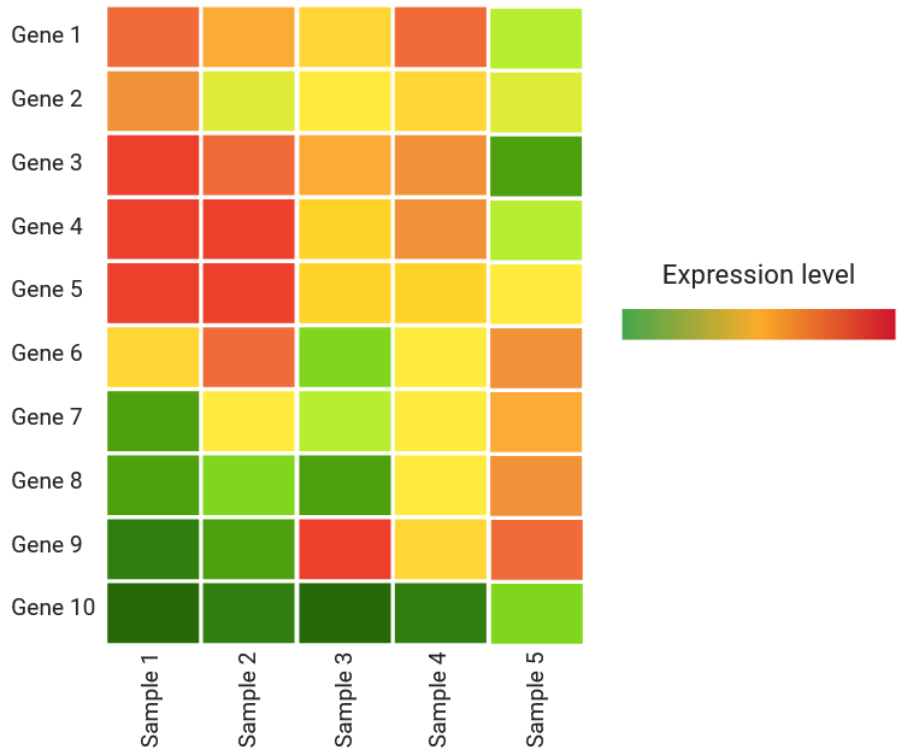

Figure 1.1 An illustration of typical omics data showing the expression level for 10 genes measured in 5 different samples, although human data will typically has in excess of 14,000 rows.

1.2.2 Ontologies

To perform pathway analysis, information must be formalized to explain the relationships between genetic entities and pathways, as well as the relationships between pathways themselves. To form the prior information used by pathway analysis tools, biological information has been formalized in ontologies. Ontologies refer to controlled language in a specific field of knowledge (Gruber, 1995; Lambrix, 2014). An ontology will formally define terms and their relationships to one another. Functionally, this allows for drawing conclusions based on the axioms of a specific domain (Tõnisson & Preden, 2024). In biological research, ontologies are generally used to describe genes, gene products and what they are associated with as determined by domain experts. Figure 1.2 shows example terms, which describe biological processes, from the Gene Ontology and their relations to one another accessed via QuickGO (Huntley et al., 2014).

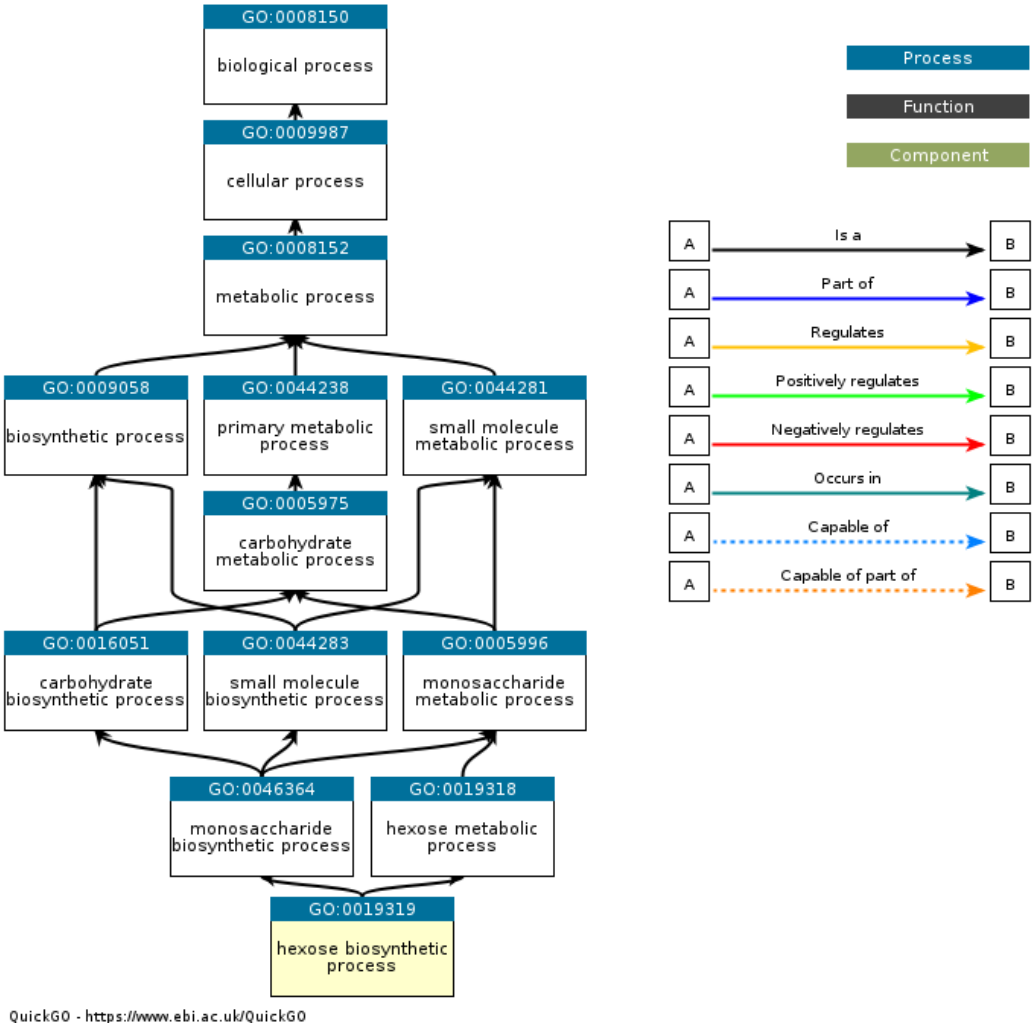

Figure 1.2 Sample terms from the Gene Ontology and their relations to each other accessed from QuickGO (Huntley et al., 2014).

This is an important part of pathway analysis because it allows for terms to be associated with genes, in “gene annotations”. The development and maintenance of ontologies allow for statistical and topology based analyses to be applied (Mazandu & Mulder, 2012; Khatri et al., 2012). This formalized information will be used to create a digital artifact which may be accessed and used in analyses (Silva et al., 2022). Ontologies have many definitions but generally all ontologies share the ability to be processed computationally (Gruber, 1995; Bernabé et al., 2023).

1.2.3 The Gene Ontology

Many ontologies exist within bioinformatics, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) or Reactome. But perhaps the most ubiquitous ontology within bioinformatics is The Gene Ontology (GO). The gene ontology was cited in peer-reviewed literature just over 9 times a day in 2024. First constructed in 1998, it is curated by the Gene Ontology Consortium (Ashburner et al., 2000). It aims to cover terms associated with genes and gene products within three functional domains: biological processes, cellular components, and molecular function. Their work applies to a vast array of organisms, and represents the experimental findings from over 180,000 publications . The Gene Ontology Resource state that the GO is currently composed of 39,354 terms and there are 972,445 annotations to the human genome (The Gene Ontology Resource, 2025).

1.2.4 Directed Acyclic Graphs

Directed Acyclic Graphs (DAGs) are a concept most commonly used in mathematics and computer science. They are defined as a graph with edges that are connected to vertices in a topological format, where no vertices and edges create a loop. DAGs are leveraged in certain implementations of pathway analysis. These techniques are common, as many popular ontologies take the form of a DAG, such as the GO mentioned previously. The combination of ontological information and genetics have allowed for the application of practices out of graph theory to understand the relationships between omics data and known biology (Yinyin Yuan & Chang-Tsun Li, 2008).

1.2.5 Web Ontology Language

Ontologies generally take the form of a digital artifact which may be accessed, shared and processed by people and machines (Silva et al., 2022; Bernabé et al., 2023). One common form of representing an ontology is in the Web Ontology Language (OWL) format, which is a

human and machine-readable file format for ontologies (W3C OWL Working Group, 2012). The ISO states the OWL format as one of the required file options for storing a top-level ontology, as defined by the ISO in their ISO/IEC 21838 standard (2021). Part of this standard refers to the principles of the Open Biological and Biomedical Ontologies (OBO) Foundry, an organization that promotes the usage of ontologies within the life sciences. This group electronically provides access to hundreds of ontologies which cover a wide range of topics including genetics, anatomy and agriculture. They lay out best practices and requirements for forming ontologies, as well as define the OBO file format. OBO Foundry manages the stewardship of ontologies, ensuring that and they are developed by diverse subject matter experts and have a central manager. They also ensure the ontology is useful, and contains unique information. (Smith et al., 2007; Jackson et al., 2021)

1.2.6 Differential expression analysis

Covered in a review by Rosati et al., to generate a list of differentially expressed genes, traditional statistical techniques are applied to omic data. Briefly, these techniques determine if differences in the expression level of a genetic entity exist between treatment levels. This may be performed with parametric or non-parametric tests, and is performed on data which is preprocessed with some form of normalization applied to reduce the systematic or technical variation. These biases may enter the data through the preparation and handling of samples (Chua et al., 2023). Common normalization within bioinformatics includes trimmed mean of M-value (TMM), which is implemented in tools such as edgeR (Robinson et al., 2009), or fragments per kilobase of transcript per million fragments mapped (FPKM) (Rosati et al., 2024). Figure 1.3 shows an illustration of how differential expression analysis determines differential expression.

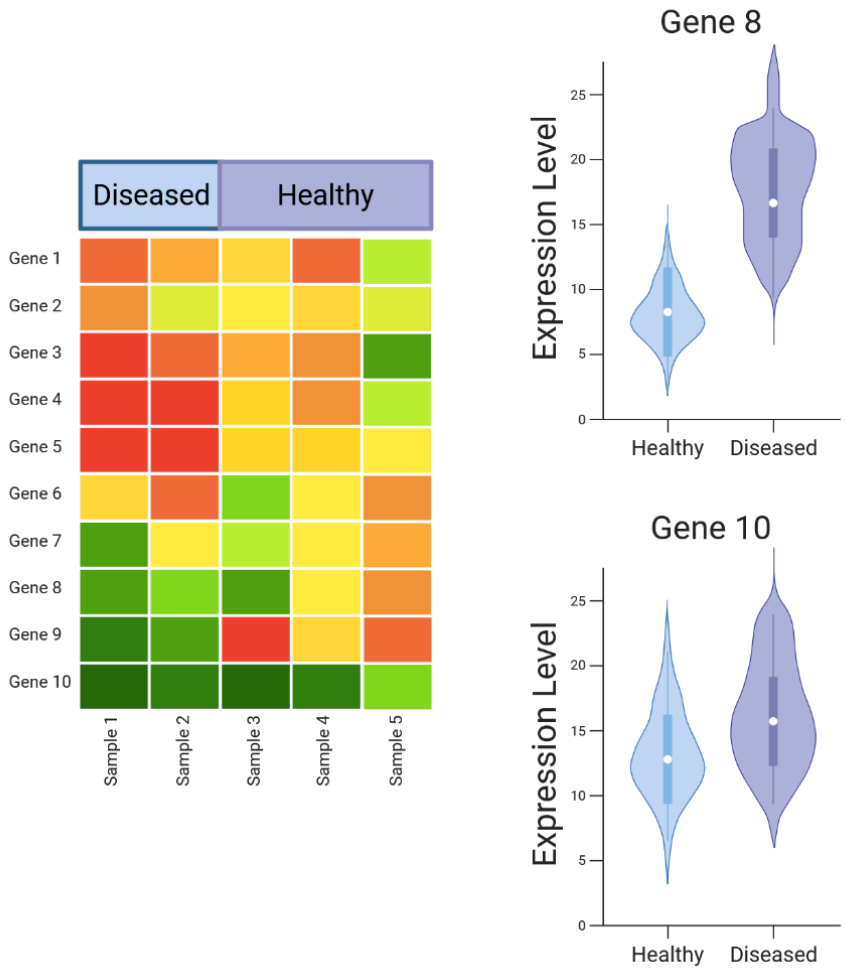

Figure 1.3 An illustration of typical differential expression analysis. A test such as a t-test is applied to determine if a gene showed a statistically significant change in expression between the phenotypes. In this hypothetical example, Gene 10 shows similar expression between the healthy and diseased samples, but Gene 8 shows different expression levels. In this case gene 8 may be considered “differentially expressed” for downstream analyses.

1.2.7 Overrepresentation analysis

One of the most popular methodologies of pathway analysis is referred to as overrepresentation analysis (ORA) (Xie et al., 2021). Briefly, ORA is an analysis which takes as input a variable number of gene labels that represent which molecular entities measured in an

experiment that showed a statistically significant difference in expression level between experimental conditions. ORA is performed by examining pathways associated with the set of differentially expressed genes. An overrepresented pathway is then identified by testing if a pathway has a higher proportion of associated genes in the list of those differentially expressed than would be randomly expected. The *a priori* relationships between genes and pathways are defined in ontologies. Confidence in overrepresentation can be calculated statistically with a hypergeometric distribution, binomial probability, chi-squared, or others (García-Campos et al., 2015).

For a hypergeometric distribution in overrepresentation analysis, the p-value for a pathway can be calculated with the following formula:

$$P(number\ of\ DE\ genes > K) = 1 - \sum_{i=0}^{K} \frac{\binom{J}{i}\binom{N-J}{M-i}}{\binom{N}{M}}$$

Where the p-value, $(P(number\ of\ DE\ genes > K))$, is the probability of having K or more DE genes in a specific pathway for the total number of DE genes. This is calculated where N is the number of genes, M is the number of DE genes, J is the number of genes associated with a specific pathway. Calculated is the probability of having more than K DE genes (Yang et al., 2014) due to chance alone, under the null hypothesis.

A contemporary popular tool that implements ORA is clusterProfiler mentioned previously (Yu et al., 2012).

While widely used, ORA represents only one type of pathway analysis, with the other major types being functional class scoring (FCS), implemented in Gene Set Enrichment Analysis (GSEA), and pathway-topology based methods, such as Pathway-Express. Limitations of ORA include the use of a statistical cutoff threshold, which will reduce the amount of data used in analysis (Pavlidis et al., 2004), and are arbitrary (Huang et al., 2008). Additionally, ORA assumes that pathways are independent to one another, which is contrary to observations of

pathway interactions in biology (Barabási & Oltvai, 2004; García-Campos et al., 2015). Despite the drawbacks of each major method of pathway analysis, all these methods see widespread implementation. Additionally, there is no perfect tool for every experimental condition (Nguyen et al., 2019), and in reality, oftentimes a combination of techniques is used to hone in on relevant biology (Alhamdoosh et al., 2016).

1.3 Neural networks

Neural networks are an implementation of artificial intelligence. They are a design paradigm within computer science denoted by its network structure that historically took inspiration from the structure and functioning of the human brain. Generally, they are designed to process vectorized data by performing a series of calculations in “layers” on input data typically in an attempt to classify the input into a category. This classification is made by having a period of training in which the “strengths” of the connections in the network are changed by altering parameters in the calculations. The following section will by no means be a thorough explanation of all neural net implementations, but due to the interdisciplinary nature of this work, we will provide an overview of the basic concepts of neural networks which will have relevance to bioinformatic applications explained later.

1.3.1 Artificial neurons

While there are many implementations of neural networks, the basis of almost any neural network is the artificial neuron. This neuron takes a set of information or inputs from adjacent neurons and produces an output. The inputs go through a calculation wherein each input has an individual weight which multiplies the input signal. This weight can be thought of as how strongly one neuron influences another, similar to the synaptic strength between biological

neurons. After this weighted signal is calculated, all inputs are summed into creating the activation signal. This activation can be many different nonlinear functions, such as the *tanh* signal or rectified linear unit (ReLu) function (Worden et al., 2023).

1.3.2 Architecture

Following the design of the artificial neuron, these artificial neurons can be now arranged into a network where layers of neurons connect into a subsequent layer of neurons in what is considered a “feedforward network”. This network structure also sees the inclusion of a bias element, which is a constant added to all neurons (which may equal zero) to impact further calculations of activation functions. Adding a bias to offset activation functions may be necessary for training a network to properly classify its input. The final layers of neurons in the network represent the output produced by the network. In many applications, this output is a numeric representation of the category of the input data (Worden et al., 2023). In the case of bioinformatic applications, classifying input data as healthy or diseased might take the form of classifying input as 0 or 1 respectively. An illustration of an example neural network can be seen in Figure 1.4.

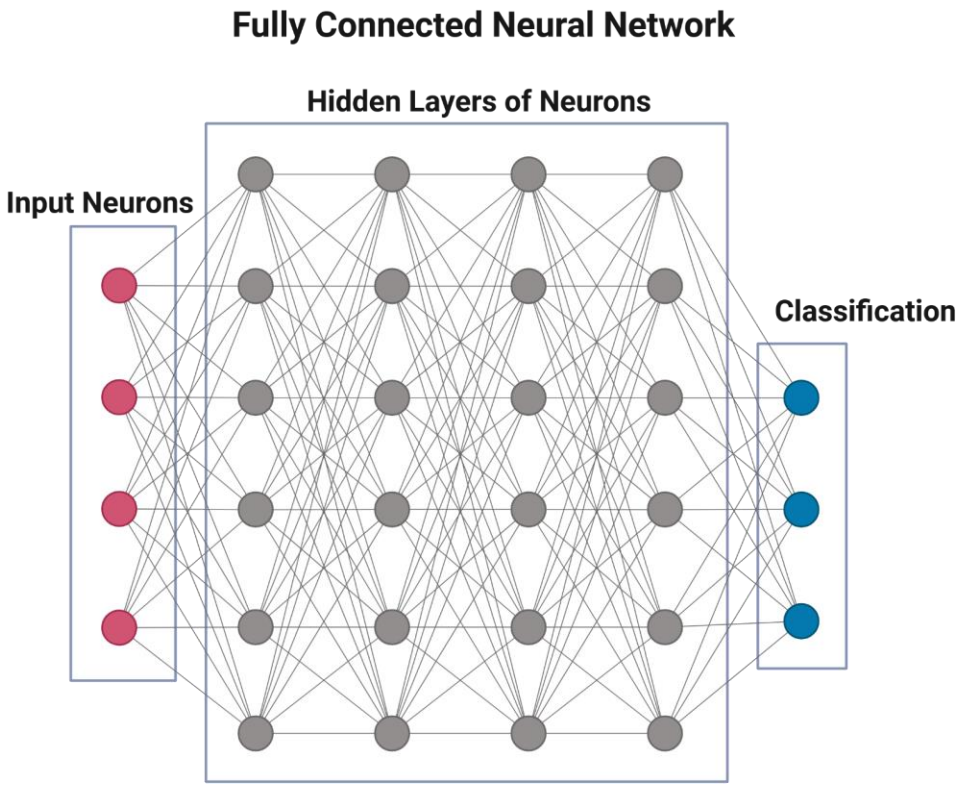

Figure 1.4 An illustration of an example fully connected neural network which shows layers of neurons connected to all neurons in the adjacent layers. The network is formed from an input layer, intermediary hidden layers and the output layer.

1.3.3 Training

In order for a neural network to classify correctly it goes through a training period. This is a period in which the weights of a neural network are adjusted to minimize a heuristic loss function. In a trivial example, once constructed and initial weights of inputs have been assigned, a neural network is trained in a “supervised” manner by providing input where the category is known, and seeing if the input would be correctly classified. If the error is within a range deemed tolerable by the implementation, the weights are considered correct. If this error is too great,the weights of the connections are adjusted using the technique of backpropagation. This process is

repeated over a set number of epochs. An epoch is complete once the entire training dataset has been evaluated for its accuracy (Worden et al., 2023).

Neural networks have been shown to have great predictive ability in situations where classes of data are defined in complex non-linear fashion. However, they are considered something of a “black box” in that it is not immediately clear how they are making their classifications (Selby et al., 2025).

1.3.4 Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) is a field interested in determining how important specific inputs or nodes are within a neural network towards classification (Ali et al., 2023). This has developed as the interest in the idea of “trustworthy AI” has grown (Markus et al., 2021). For reasons such as privacy, ethics, and misinformation, the ability for an AI system to show how inferences are being made is crucial (Bostrom & Yudkowsky, 2018). For instance, ensuring that an AI system is not using unethical prejudice in classification on the data of individuals requires the ability for the AI to provide information on how classification is being performed (Hardt et al., 2016; Mandel & Barnett, 2024). Early techniques of explainable AI included a connection weight metric (Garson, 1991), a salience metric (Ruck et al., 1990), and using partial derivatives for sensitivity analysis (Dimopoulos et al., 1995). More contemporary techniques such as SHAP or Deep SHAP values (Lundberg & Lee, 2017), Integrated Gradients (Sundararajan et al., 2017), or conductance (Dhamdhere et al., 2018) can perform such a role.

Implemented by Lundberg & Lee, Deep SHAP values take inspiration from techniques out of game theory and are a computationally less expensive approximation of Shapley scores. These scores can be produced for a neural network, where each node will be given a score of how important the node was in classifying each condition of the network (2017). This is an approximation of how less accurate the network would classify input if the node was removed (Hartman et al., 2023).

1.3.5 Biologically Informed Neural Networks

In their review, Selby et al. (2025) discuss a bioinformatic specific implementation of neural networks which emerged around 2018, the biologically informed neural network (BINN). They typically perform classification on omic data, use XAI techniques, and use a network architecture that leverages information from biological ontologies. Additionally, they have been reported to provide an alternate method for pathway analysis (Hartman et al., 2023, Cao et al., 2025). The promise of this technology can be seen in their ability to perform well with low sample sizes and leveraging explainable AI techniques to perform biomarker discovery and pathway analysis (Selby et al., 2025).

As compared to the fully-connected feed forward network from the previous section, BINNs create a network in the shape of the ontological terms associated with the genetic materials inputted into the network. This informs how many neurons (or nodes) to create based upon the input data, but also informs the connections between nodes. BINNs will prune any connections between nodes if these connections do not exist in the reference ontology. An illustration of a BINN architecture can be seen in Figure 1.5, which, in contrast to Figure 1.4, shows sparse connections which represent the relationship between ontological terms, similar to that seen previously defined in the Gene Ontology.

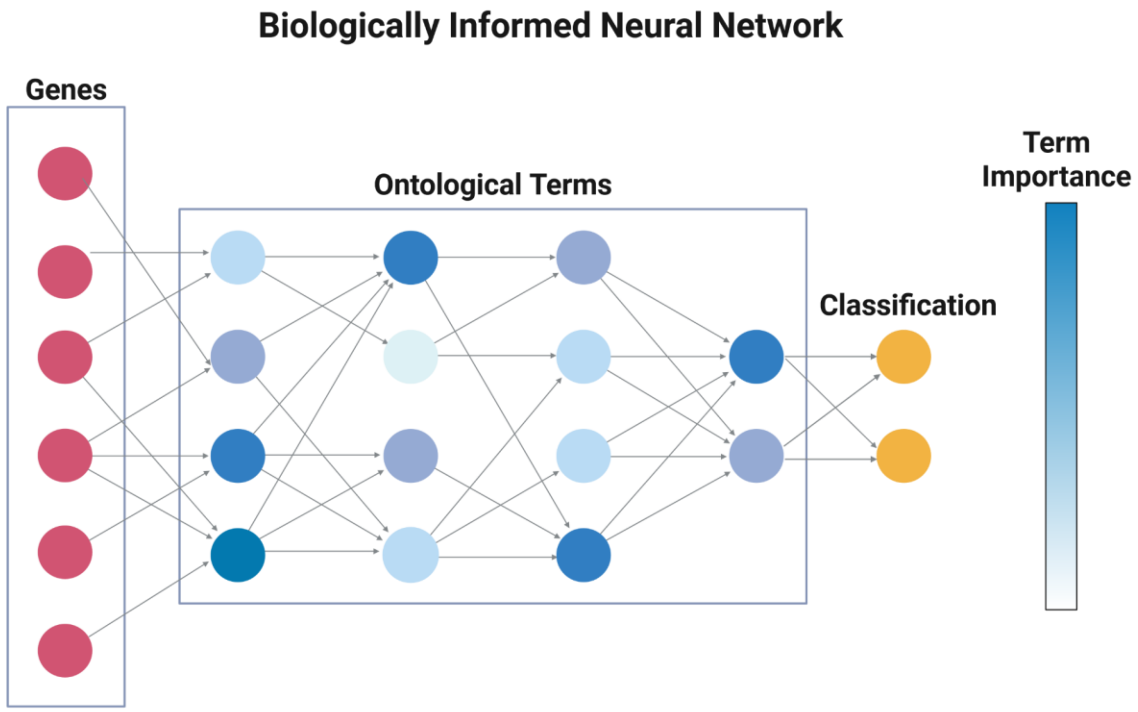

Figure 1.5 An illustration of a BINN. In contrast to the previous example, connections in the network are pruned so that the nodes represent an ontology. Once training has occurred, the importance of each term is calculated using an explainable AI metric.

The neural net representing biological information in it's architecture opens the network to greatly improved interpretability (Hao et al., 2018; Elmarakeby et al., 2021; Novakovsky et al., 2022). At the same time, despite a theoretical trade-off in machine learning between interpretability and predictive performance (Murdoch et al., 2019), these models have shown improved performance when compared with a densely connected counterpart with otherwise similar structures (Elmarakeby et al., 2021). Similarly, in their paper, Hartman et al. show that when using the Reactome ontology as reference for constructing connections between neurons, a BINN was able to classify cases and controls in septic acute kidney injury (AKI) and COVID-

19 based on proteomic input data with higher accuracy than current machine learning (ML) techniques, such as a support vector machine with radial basis function kernel, k-nearest neighbor, a random forest, and two boosted tree based methods (2023).

While integrating ontological information into network structure appears to improve classification, the use of XAI techniques holds promise to help inform biological research. **The intuition of applying these metrics is that if you are able to calculate the importance of a node (i.e. ontological term) or input (i.e. genetic material) towards classifying healthy or diseased samples, you may be calculating the importance of the genetic material associated with the ontological term to the disease itself**. This technique has been applied in discovering biomarkers across conditions such as cancer and COVID-19 (Hao et al., 2018; Hartman et al., 2023; Cao et al., 2025). These technologies have also been applied in basic science research, and have been used to explore genes of relevance in the biology of alternate splicing, using knockout experimental data (Cao et al., 2025).

1.4 Proposed Approach

Presented in this manuscript will be an approach of running and interpreting the explainable AI metrics of a BINN. We will first take publicly available bulk RNA-seq data and use a clustering algorithm described in Chapter 4 to separate this dataset into distinct modules of expression data as a quality control preprocessing step. For each module, a BINN will be trained to classify the experimental conditions of the expression data. At the same time, we will train “null BINNs” on the same data by scrambling the labels of the dataset before training. For each term in the network, we will compare the value of the explainable AI metric in the “null BINNs” to the “real BINN”. We consider the distribution of metrics for a term in the “null BINNs” as an estimate of the null hypothesis for each term in the network. Instead of using the explainable AI metric itself as a measure of relevance, we use a measure of the dissimilarity of

the “real BINN” to the “null BINNs”, and show this value recapitulates the traditional, hypergeometric approach with a higher correlation across a wide spread of sample sizes. The scrambling approach and results will be presented in Chapter 5. An illustration of our proposed approach can be seen in Figure 1.6.

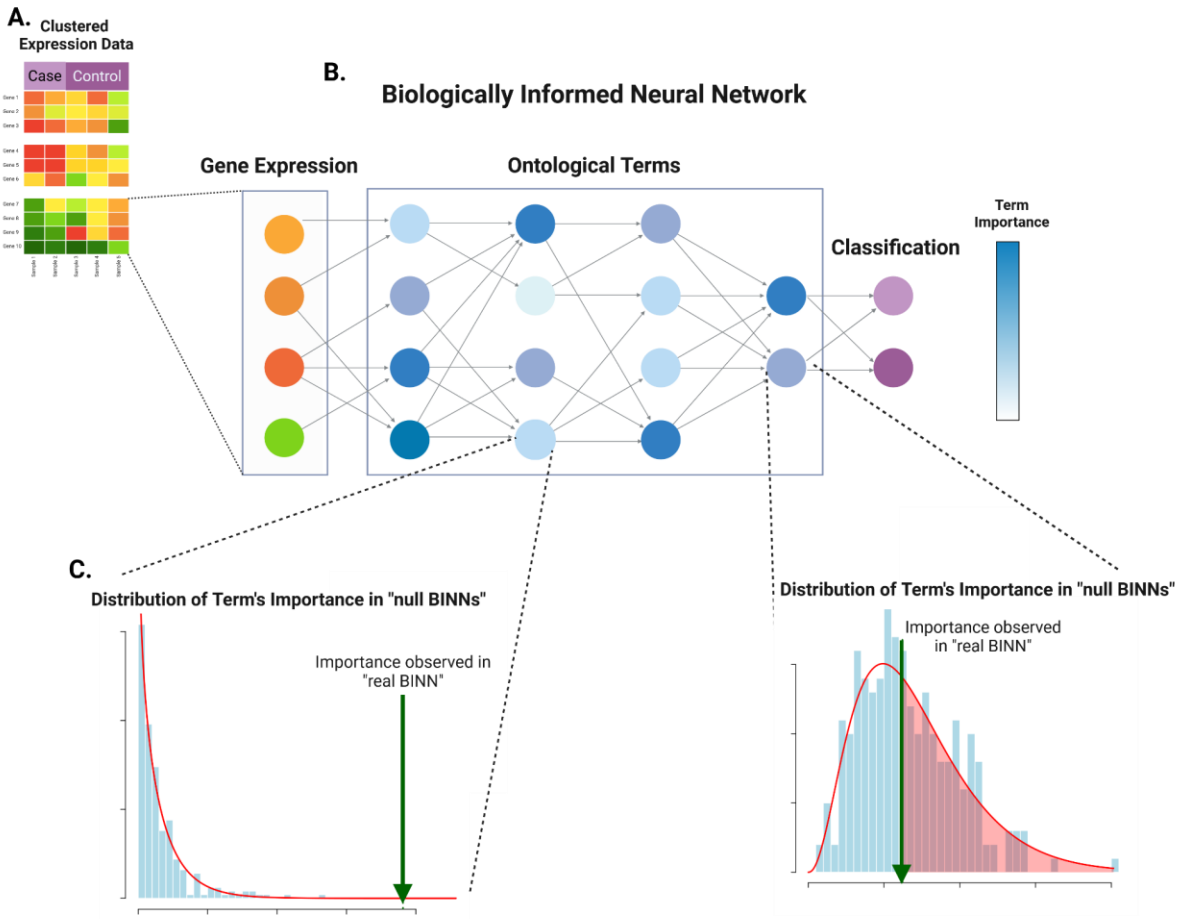

Figure 1.6 Our proposed approach to examining BINNs. Panel A shows an omics dataset broken into modules of genes. Described in Chapter 4, a BINN will be trained on each module to classify the experimental phenotype, shown in Panel B. Panel C shows a methodology explained in Chapter 5 in which the importance of a node observed in the BINN will be compared to the importance calculated for that term in a BINN trained on the same module, but with the phenotypic labels scrambled (the “null BINNs”).

Chapter 2. Rationale, Hypothesis, and Objectives

2.1 Rationale

BINNs are emerging as a novel pathway analysis technique. They may perform better than current methods in low sample size environments, which are common within translational medicine. There are limited assessments of XAI metrics for use in pathway analysis, and they have been largely explored for only biomarker discovery. BINNs also have many features which are desirable in pathway analysis methods (multi-omic integration, low sample size tolerance, not constrained to binary experimental designs).

Performing pathway analysis is difficult in a translational setting due to the lack of sample size, particular in rare diseases. Application of pathway analysis offers a technique for hypothesis generation by discovering aspects of biology disrupted by disease. The relevant information reported by these tools may be used to discover druggable targets, design future wet-lab studies, or better understand disease pathogenicity.

However, best practices in applying BINNs for pathway analysis have not been developed. Evaluating how BINNs perform in comparison to traditional techniques may lay the groundwork for further exploration of these unique features.

2.2 Hypothesis

BINNs may be able to reproduce the results of traditional approaches to pathway analysis, and we may be able to improve their ability to reproduce these results by applying techniques common in bioinformatics. Application of clustering of features to input data, and the addition of a permutation testing framework to BINNs may improve the ability for BINNs to recapitulate information from traditional approaches. Recapitulation of a binary approach may allow us to extend the BINN approach to complex designs and multi-omics.

2.3 Research objectives

Aim 1: Explore the results of BINNs' explainable AI metrics to a hypergeometric ORA approach when applied to the same dataset.

Aim 2: Explore the change in ORA recapitulation and neural network accuracy when there is a change in sample size

Aim 3: Explore the impact of clustering input features the input data to BINNs on ORA recapitulation and neural network training accuracy

Aim 4: Explore the role of a permutation test framework on the ability for BINNs to recapitulate the ORA approach

Chapter 3: Comparing Deep SHAP Importances to Overrepresentation Analysis p-values

3.1 Motivation

While largely explored for their strength in classification, BINNs offer promise as a relatively novel and alternative approach to perform pathway analysis (Selby et al., 2025). Examining dysregulated biology through this type of method has been an essential avenue through which omics experiments can be translated into actionable insight (García-Campos et al., 2015). There are unique features of BINNs which present potential benefits for translational medicine research when compared with other existing techniques.

Firstly, BINNs are able to perform multi-omic analyses (van Hilten et al., 2024). A multi-omic approach, where data is integrated from multiple different high throughput sequencing data sources (proteomics, transcriptomics, metabolomics, etc.) has shown strong advantages in uncovering disease pathogenicity and drug discovery (Chen et al., 2023). Many current techniques for performing multi-omic analyses combine techniques applied separately to each data source, while BINNs offer a method to perform a single analysis which takes multiple omic sources as input.

Secondly, BINNs have been noted to excel with high-dimensionality, low sample data, (referred to as “wide data”, where there are few samples and many features) which is common within translational research. It has been documented a large portion of omics studies will only have samples “in the tens” (Kirpich et al., 2018). This may be due to the fact that there is often a scarcity of samples in the research of rare diseases, wherein the “precious samples” of these diseases create a bottle-neck of sample size (Barrett et al., 2023). By leveraging the machine learning techniques alongside the external biological information, it is hoped that BINNs may mitigate the impact of small sample sizes seen in the current best practice pathway analysis techniques.

Additionally, BINNs also can classify a variable number of treatment levels, which some statistical tests used in traditional pathway analysis approaches, such as the hypergeometric (or Fisher's exact test) are unable to do (Nayak & Hazra, 2011). Additionally, in our experience, a multiplicity of treatment levels create a situation where statistical approaches become untenable. For instance, our collaborators have used experimental designs containing 20 treatment levels, encompassing biologic sex, tissue site, timepoint and disease exposure. This creates a situation in which to apply a statistical analysis of variance (ANOVA), you must perform 190 post-hoc pairwise analyses and the appropriate multiple testing correction for examining over 30,000 ontological terms, totalling over 5,700,000 tests. With the sample sizes seen in our practice, it is nearly impossible to find a statistical signal strong enough to counteract multiple testing corrections for 5,700,000 inferences in this case.

Despite their strength as a classifier, there has not emerged a consensus way of interpreting the explainable AI metrics generated by BINNs for pathway analysis. The lack of guidance in applying BINNs has led to ad-hoc interpretations of these metrics, which may be inconsistent or unreliable (Chen et al., 2024).

Furthermore, it appears that BINNs are primarily used for the discovery of biomarkers (i.e. genetic materials of relevance) rather than leveraging their architecture to determine relevant ontological terms (Hartman et al., 2023; Cao et al., 2025). We seek to interpret the ontological terms themselves.

Some formalized approaches of performing pathway analysis with BINNs have emerged, as Hartman et al. compared the terms in their BINN with the highest Deep SHAP importance to a traditional analysis pathway analysis using Metascape on the same dataset (2023). In many cases for pathway analysis, it is common for researchers to compare the overlap in results between two methodologies, which lends credence to examining the overlap in BINNs metrics and traditional approaches. Other ML techniques used to determine significant pathways, such

as the implementation by Cao et al., compared results with literature to interpret their BINN focused pathway analysis results (2025).

There may be first-principles issues with comparing the output of one method to another, however it has been proposed that this may be a pragmatic way to validate novel pathway analysis tools (García-Campos et al., 2015). Pathway analysis tools may be built on similar approaches, which may lead to a circular validation, and large sample sizes may be required to avoid this. However, different approaches, such as creating “gold-standard” data to evaluate pathway analysis tools, have proved to be difficult (Tarca et al., 2008; García-Campos et al., 2015). In this light, a comparison of methods has been posited as a reasonable technique.

3.2 Materials and Methods

3.2.1 Introduction and Workflow

Considering the relatively narrow scope of previous evaluations of BINNs as pathway analysis tools, and in keeping with suggested methods described previously, we propose a way to compare the metrics produced by BINNs to traditional bioinformatic approaches. We will compare all terms evaluated in a BINN to a traditional, hypergeometric ORA approach on publicly available RNA-seq data.

3.2.2 RNA-seq Expression Data

To produce results for both methods, RNA-seq expression data from human cancer cell-lines from the Cancer Cell Line Encyclopedia (CCLE) was downloaded through the EMBL-EBI Expression Atlas web portal (Barretina et al., 2012; Ghandi et al., 2019; EBI Functional Genomics Team, n.d.).

The published dataset was previously processed with iRAP 1.0.1. From this, the expression level data of 55,843 genes were normalized to a FPKM value for 169 unique cancer

types from 1,019 human cancer cell lines. Undetected transcripts were marked as not available (NA) in the dataset. Transcripts were labelled with Ensembl human transcript gene IDs using Ensembl release 95.

To evaluate BINNs, we used the two most abundant cancer types from the CCLE dataset, which was two different phenotypes of lung cancer: lung carcinoma, which had 68 samples, and small cell lung adenocarcinoma, which had 49 samples. From these samples, we formed four datasets: a dataset comprising all samples from each cancer, a dataset of 24 samples from each cancer, a dataset of 12 samples from each cancer, and finally a dataset of 6 samples from each cancer.

This dataset was chosen primarily as 117 samples is a large number of samples within the field of translational omics. Large sample counts benefit the statistical power of methods used in overrepresentation analyses (Gan, 2025). Additionally, the sample size was comparable in sample size in the work which produced the *binn* library. Hartman et al. used 150 samples of normalized proteomics expression data to classify two severity levels of Sepsis-Associated Acute Kidney Injury using a BINN, and produced high classification accuracy.

3.2.3 Biologically informed neural networks

To create and train a BINN, we used a Python library implementation of a BINN, *binn,* version 0.1.1 produced by Hartman et al. This was chosen as it offered several features which aided analysis. Firstly, it was developed using packages which did not require specific architecture such as libraries designed for use with graphics processing units. Additionally, it allowed for use of custom ontologies and mapping of genetic entities to ontological terms to form network structure. All implementation was performed using this package in a Python version 3.11.8 environment.

The *binn* library uses PyTorch version 2.7.1 (Paszke et al., 2019) to produce a sparse feed forward neural network. This tool was validated using proteomic data, but its input layer

may be configured to take any expression level data from a single-omic source. The sparse network is set prior to training and connections are formed between terms which exist in a .txt file designating connections in an ontology. This forms layers of linear nodes which are pruned to represent the connections in the ontology. Between each of these layers is a layer of nodes which performs batch normalization, then a layer which performs dropout to prevent overfitting. A representation of the neural network architecture implemented by default in the library can be seen in Figure 3.1. A more detailed visualization of the hidden layers can be seen in Figure 3.2.

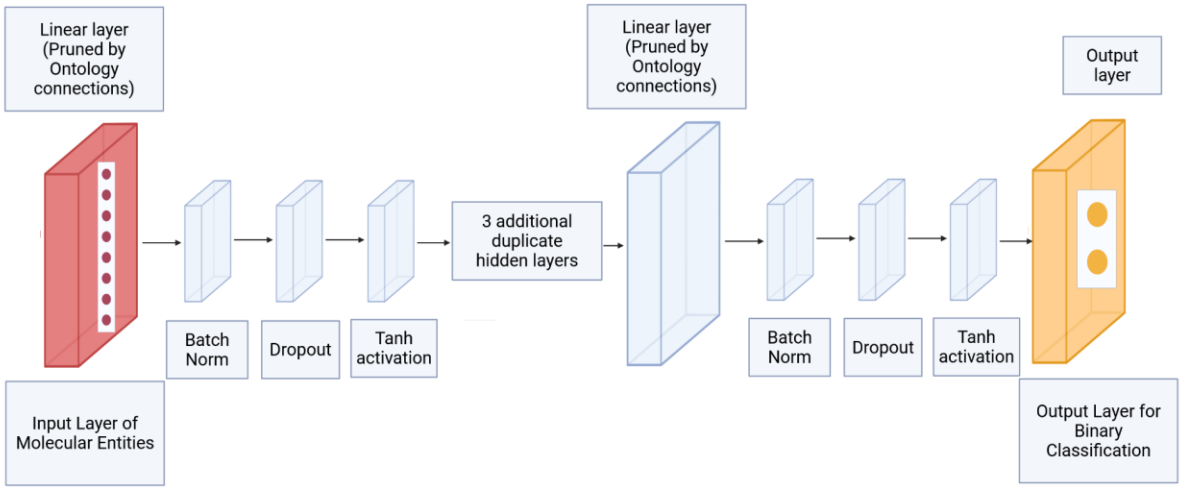

Figure 3.1 The default network design implemented in the *binn* library produced by Hartman et al. (2023).

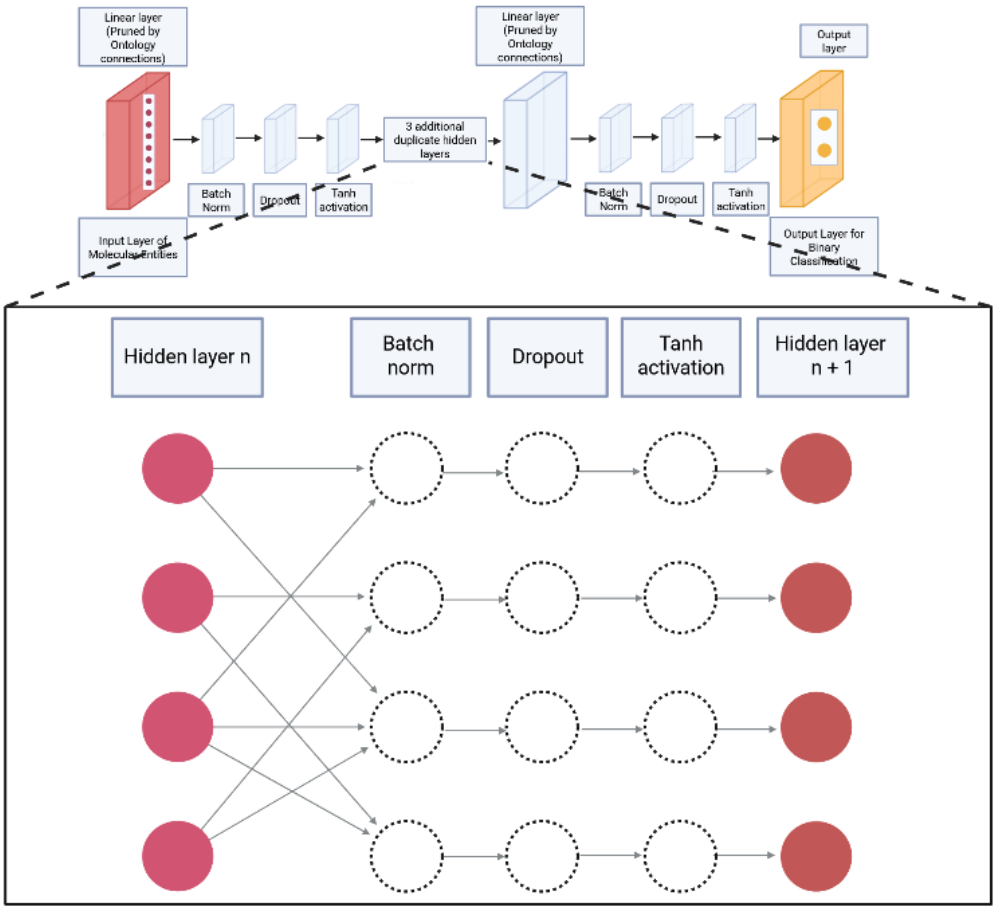

Figure 3.2 The hidden layers in the default implementation of the *binn* library. Each layer contains a linear layer which is pruned to represent the connections in the ontology. This layer then has batch normalization, dropout and a tanh activation function after the linear layer.

This implementation of a BINN used the explainable-AI Deep SHAP importance metric to assign a numeric value to each node in the network. This value is a computationally inexpensive approximation of the importance of a node in classification. It roughly corresponds to how much worse the network would classify samples if it were to be removed. In the implementation of BINNs we employed, Hartman et al. perform a normalization step, in which a

node's importance is divided by the logarithm of the number of connections which precede the node (2023). These values are generated after the network is trained with the package's data-explainer library for each node in the network (all of which correspond to a term in the ontology). Later in this chapter, we will compare this metric to the p-values produced in a traditional hypergeometric test.

Our implementation of this library uses the default settings as supplied by the package. This uses a 80/20 split of training and validation data, with 0.2 dropout and 4 layers of nodes in the network, and is trained over 100 epochs. The library is implemented to construct a network with an activation function of ReLu, tanh, sigmoid, and others, but tanh was used in our implementation. The network was trained to classify the two lung cancers. This model used a cross-entropy loss function. As input, the expression level data was supplied in a tab-separated value (.tsv) file, along with a .tsv containing the classification of each sample (zero or one, corresponding to the two lung cancers). The *binn* library imputes NA values in the input as zero. Metrics of training (training loss, validation loss, training accuracy and validation accuracy) were reported by the *binn* library to console and recorded using the built in contextlib library from Python.

In order to map the inputs of the RNA-seq dataset to the nodes in the network, the relationships between gene identifiers and ontological terms were supplied in a .txt file. This file contains a mapping of gene identifiers to ontological terms supplied as a .tsv. The mapping file contained two columns corresponding to an ENSEMBL human transcript gene identifier, and the next column was the term in the Gene Ontology accessed from the R library GO.db version 3.18.0 (Carlson, 2023), which was used to construct this file.

To test the ability of BINNs to perform at lower sample sizes, we trained four BINNs, one on the full dataset, and three subsets. The subsets were a dataset made out of 6, 12 and 24 samples from each of the two lung cancer chosen at random without replacement.

3.2.4 Differential Expression Analysis

To compare Deep SHAP importances from a BINN approach with the results of an overrepresentation approach, we ran the CCLE expression data through a differential expression analysis pipeline. This was developed in R version 4.3.1 and RStudio server version 2024.03.999-dev+999 (Ocean Storm) for Linux (R Core Team, 2023).

First, a list of differentially expressed genes was determined. This was performed with a two-sided t-test applied to the expression level of each transcript using the t.test() function implemented in the stats package version 4.3.1 (R Core Team, 2023). The t-test was performed to determine if a difference in expression between the two phenotypes of cancer existed, with each cancer type coded as 0 or 1 in R. A transcript would be skipped if either group did not contain any observed values, or if the observed values were effectively constant as determined by the t.test() library. The p-value of the t-test for each transcript was collected.

Multiple testing correction was performed by adjusting the t-test p-values using p.adjust() with a Benjamani-Hochberg correction, from the R stats package version 4.3.1 (R Core Team, 2023). An adjusted p-value of <0.05 was considered statistically significant for differential expression. As an overview of the most statistically significant terms, we produce a table of the top 25 terms as measured by smallest adjusted p-value (See Table 3.3).

3.2.5 Parsing the explanations produced by BINNs

The Deep SHAP values are stored in an “Explanations” object from the package. This was converted to a comma-separated value (.csv) and loaded into R for parsing. Each explanation was stored as a connection, with each connection between a term and another term having an importance. Using an R script, a term’s importance was calculated by summing all normalized importances for connections leaving a node. As each node represents an ontological term, from this we have a dataframe of normalized Deep SHAP values for each term

related to the input genetic elements to the BINN. As an overview of the most important terms, we produce a table of the top 25 terms as measured by highest importance (See Table 3.4).

3.2.6 Hypergeometric test

Pathway analysis was performed with an overrepresentation analysis using the R package GOfuncR version 1.28.0 (Grote, 2025). This was carried out with the go_enrich() function, which took as input the name of the organism, and a dataframe which contains a list of gene identifiers as a column, and a second column containing a 1 or a 0, indicating if the gene's differential expression was statistically significant, or not. In our workflow, as specified by Grote in the GOfuncR package (2025), we supplied the tool with all genes from the CCLE dataset for which the previous t-test was possible, with the gene identifier converted from the ENSEMBL identifier from CCLE to a gene symbol using the R library biomaRt version 2.58.2 (Durinck et al., 2009).

To calculate an ontological term's differential expression the go_enrich() function uses a hypergeometric test. This produced a list of ontological terms with two p-values for overrepresentation or underrepresentation of the term. The resulting hypergeometric p-values, which represent the statistical significance of the ontological terms, were adjusted with p.adjust() using Benjamani-Hochberg, which resulted in a list of adjusted p-values. All terms from the Gene Ontology were used, which was accessed from the R package GO.db version 3.18.0. To consider overall differential expression, we take the minimum p-value of the overrepresentation p-value and underrepresentation p-value. We finally examine the top 25 terms by the lowest p-value.

3.2.7 Comparison of Deep SHAP importances and Hypergeometric p-values and other evaluations

To evaluate the similarity between BINNs and the overrepresentation method for performing pathway analysis, we compared the ontological term's adjusted p-values (from the ORA) to the Deep SHAP importances for each term. All analyses were performed in R. We assessed the Pearson correlation between the -log(adjusted p-value) and the Deep SHAP importance value for each term.

Additionally, a receiver operating characteristic (ROC) curve analysis was performed to visualize the relationship between sensitivity and specificity to true positive rate and false positive rate of Deep SHAP importances in classifying ontological terms. In this test, "positives" were considered terms with a p-value less than <0.05 as determined by the hypergeometric approach. Briefly, the sensitivity and specificity analysis involved considering all possible thresholds of importances as significant, and examining how many terms under the threshold were significant according to the hypergeometric. The ROC curve plots the true positive rate and false positive rate over all thresholds.

Additionally, a Spearman correlation was calculated to determine the similarity in how each term was ranked for each metric with cor() in the R stats library (R Core Team, 2023). This calculated how close to a monotonic relationship terms had when ranked by ascending Deep SHAP importance and ascending p-value. This test will also report if a positive or negative monotonic relationship exists between metrics. To evaluate the performance of the neural network in classification, we visually examined the training loss and validation loss of the neural network. We also examined the validation accuracy, to see if the network was trained accurately.

3.3 Results

3.3.1 Differential expression analysis

In performing differential expression analysis, 15,380 transcripts of a total of 55,843 were skipped when performing the t-test due to all values being missing or essentially constant data. In the remaining transcripts, 15,926 transcripts had statistically significant differential expression before multiple testing correction, and 12,687 transcripts had statistically significant differential expression after multiple testing correction was applied.

3.3.2 Pathway analysis

Performing a hypergeometric test with the GOfuncR package as described previously on the transcriptomic data returned 22,045 GO terms and their statistical relevance in terms of overrepresentation or underrepresentation. After performing multiple testing correction, 282 GO terms were statistically significantly overrepresented, and 57 were statistically significantly underrepresented. The minimum p-value of these two terms was saved as a “Global” p-value.

In Figure 3.3, a heatmap of the Z-scores of our transcriptomic data shows large homogeneity in our data, however despite this, the majority of the samples of the two treatment levels appear clustered and we are still able to find relevant biological processes and terms through pathway analysis.

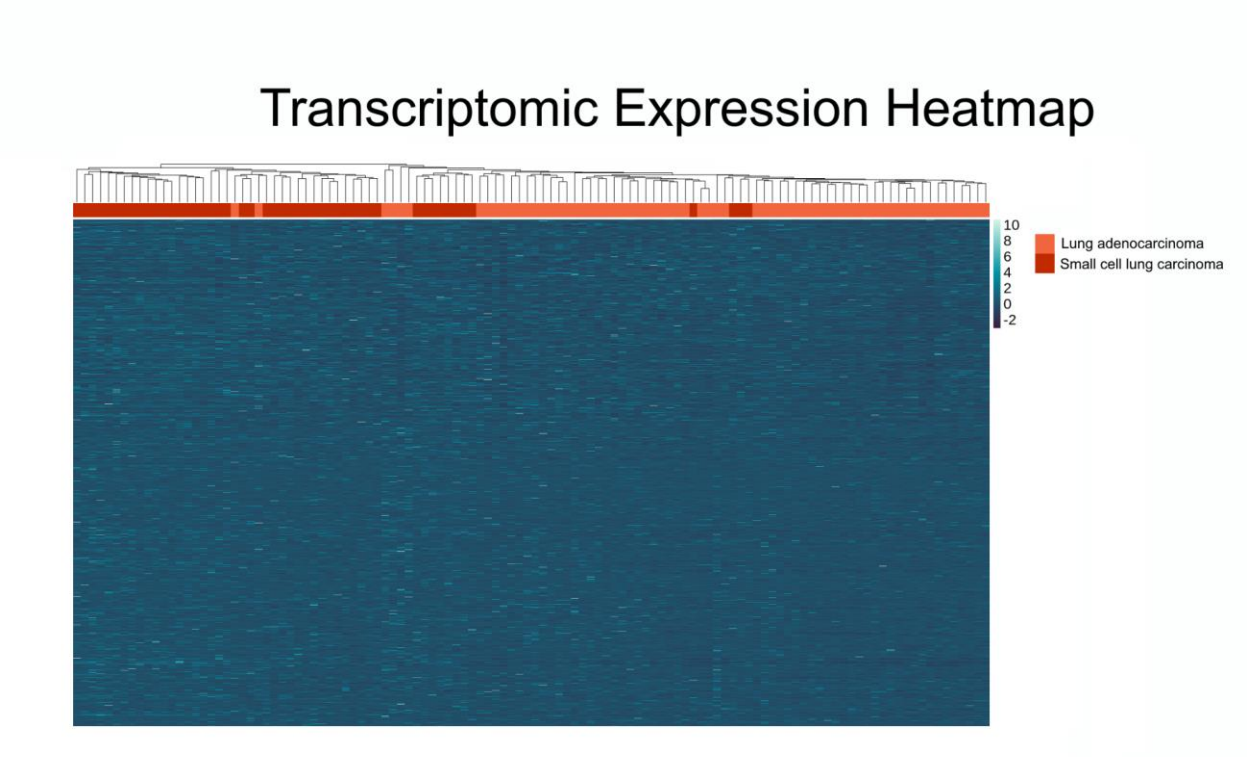

Figure 3.3 A heatmap of the Z-scores of the expression level of all transcripts from the dataset for the dataset formed from both lung cancer phenotypes.

3.3.3 Training BINNs

We observed the training and validation loss curves to evaluate how the BINN model performed as a classifier. Plotting the loss curves in Figure 3.4 over the 100 training epochs for the dataset of all samples shows a stable loss curve which reduces as training continues. This indicates a neural net model in which some overfitting occurred, since the training set was always lower than the training dataset. Additionally, the final network had a validation accuracy

of 91.67%, meaning the network was capable of correctly classifying which cancer each sample represented to a high degree of accuracy.

The BINNs trained on subsets of the data showed varying results, and are visualized in Figure 3.5, 3.6 and 3.7. The dataset formed out of 6 and 12 samples from each cancer was able to achieve 100% validation accuracy. The dataset formed out of 24 samples achieved 90% accuracy. The 24 sample dataset showed a less stable, yet decreasing loss curve, however the validation loss curve rises as training continues, suggesting examining more epochs, or selecting an earlier epoch would be more optimal. For the datasets formed from 6 and 12 samples, training loss decreases and plateaus early, however validation loss does not lower until later epoch, perhaps more epochs should have been considered to determine if this value remained stable.

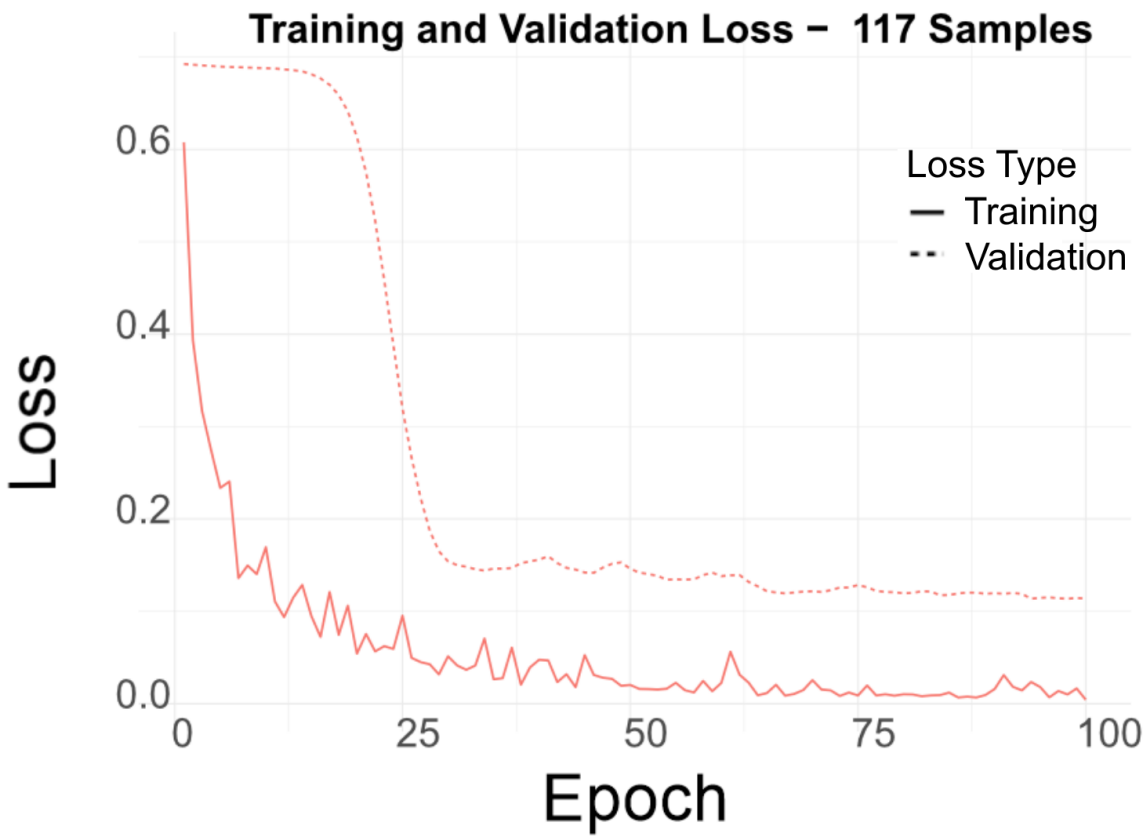

Figure 3.4 The training and validation loss curves for BINN trained on all samples from the dataset comprised the two lung cancer phenotypes.

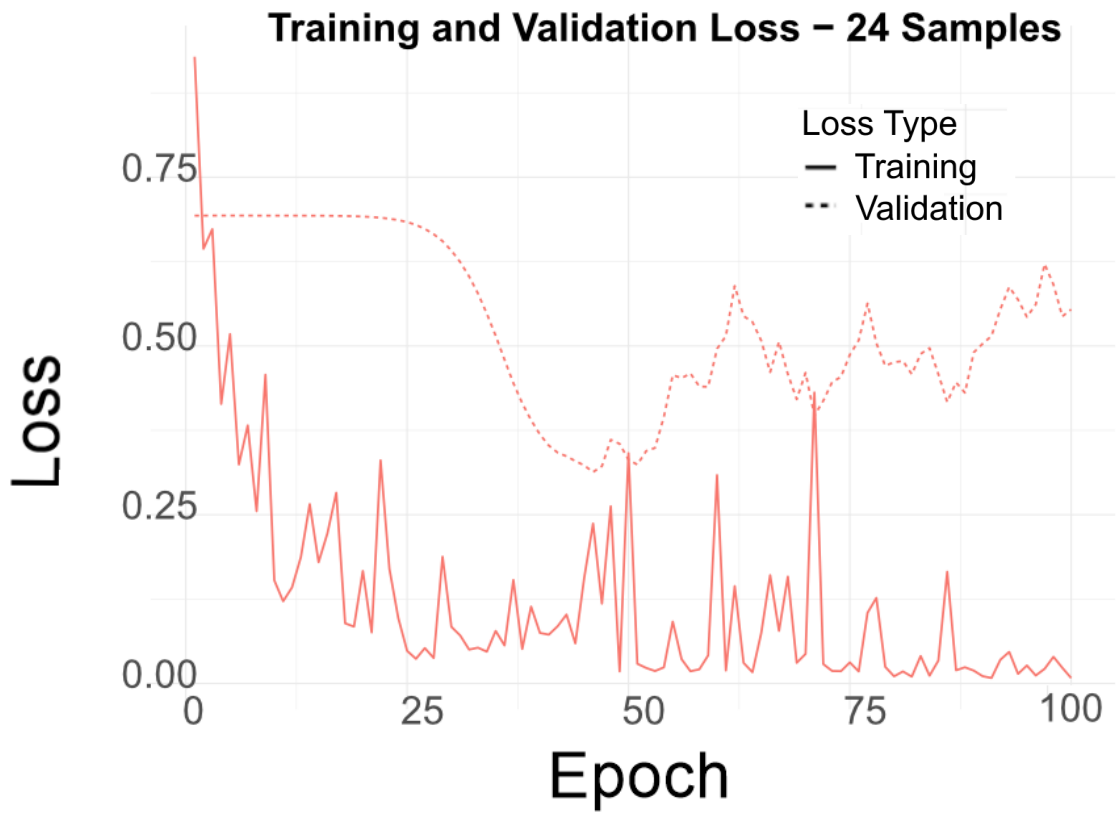

Figure 3.5 The training and validation loss curves for BINN trained on 24 samples from each of the two lung cancer phenotypes.

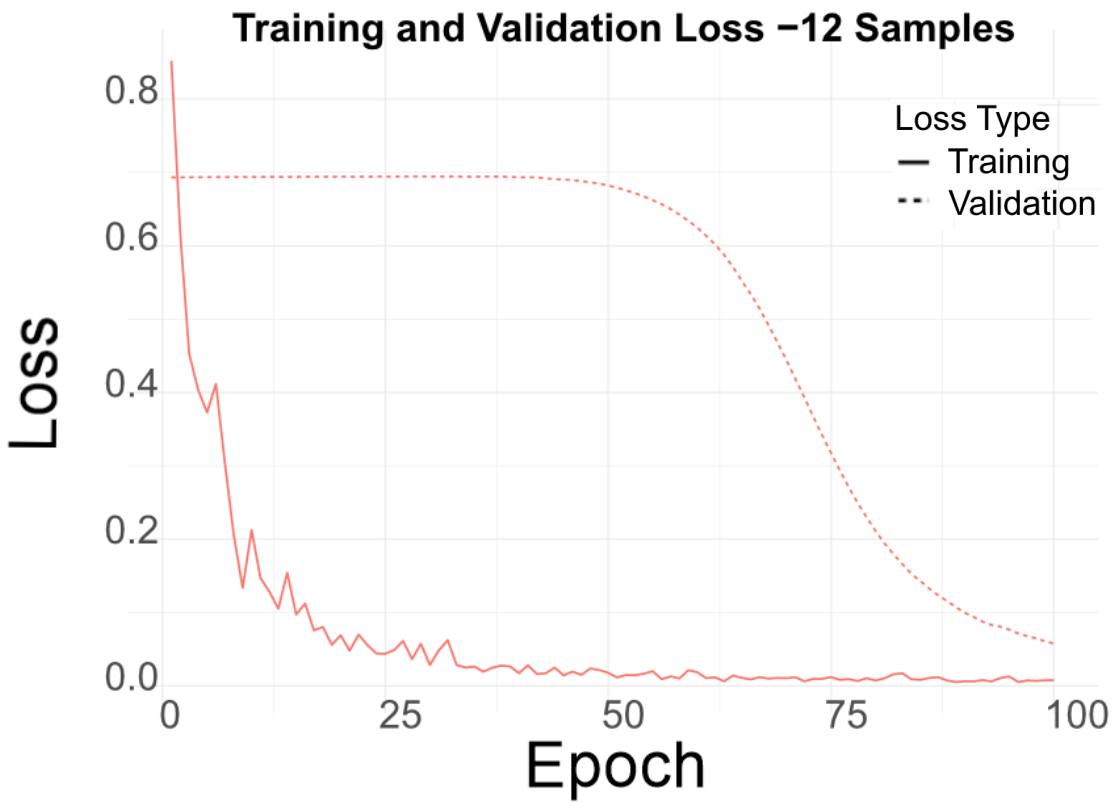

Figure 3.6 The training and validation loss curves for BINN trained on 12 samples from each of the two lung cancer phenotypes.

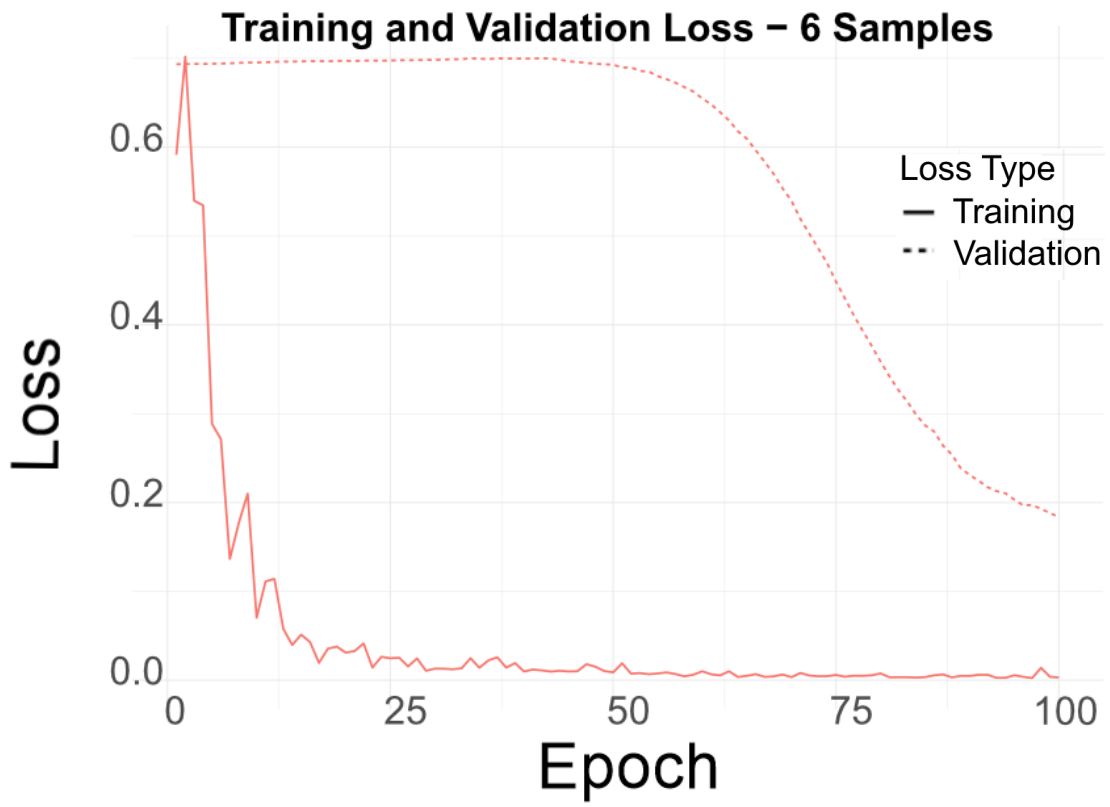

Figure 3.7 The training and validation loss curves for BINN trained on 6 samples from each of the two lung cancer phenotypes.

To examine the relationship between p-values and Deep SHAP importances, we plot the -log("Global" p-value) and importance for all terms for which both were created, which can be seen in Figure 3.8. This reveals a low correlation, with an adjusted R^2 value of 1.01×10^{-3} as calculated by a linear model in R. Additionally, of the 3,121 terms reported, of the highest 10% importance terms, only 1 was a term deemed statistically significant by the hypergeometric approach.

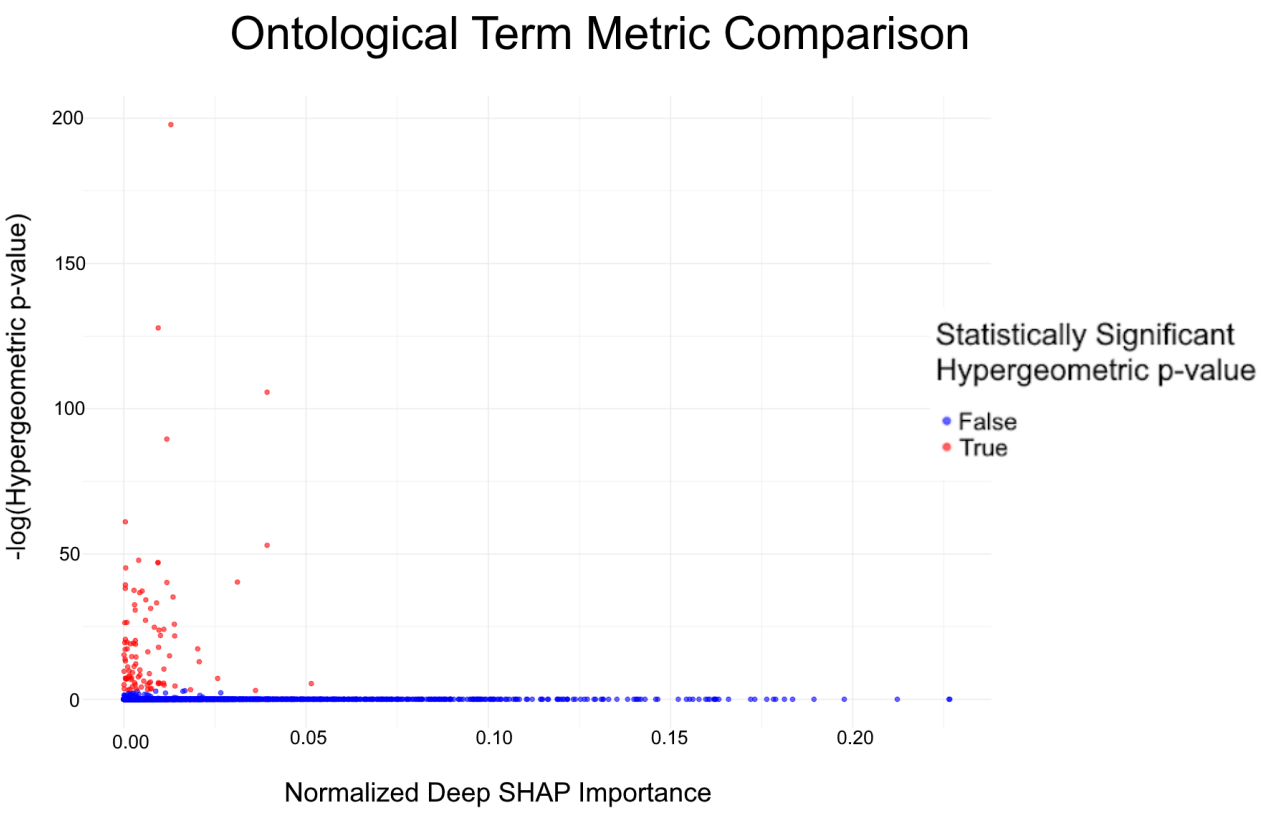

Figure 3.8 A scatter plot of all terms returned by the BINN, with the normalized Deep SHAP importance on the x-axis, and the negative logarithm of the hypergeometric p-value on the y-axis.

To further quantify the relationship between these metrics, we applied a Spearman correlation to determine if there existed a positive or negative monotonic relationship between importance and statistical significance. If importances can recapitulate the information from the hypergeometric approach, we would anticipate an inverse relationship (i.e. negative Spearman relation). When performing a Spearman correlation on the overall dataset we observe a value of 0.121294, indicating a weak positive monotonic relationship between importances and p-values. The other datasets had a Spearman test applied with a statistically significant p-value as before, and showed a similar, weakly positive relationship. The results are summarized in Table 3.1.

Table 3.1 Spearman Correlations between the Deep SHAP importance and hypergeometric p-value, along with the p-value for the Spearman test, for each dataset

Sample	Spearman correlation	P-value of Spearman test
All samples	0.121294	1.496e-11
24	0.1396363	6.928e-15
12	0.1235283	6.118e-12
6	0.1162719	9.864e-11

We finally produce a ROC curve to evaluate the importance metrics. We used this to visualize and calculate if importances were able to be used as a metric to classify which terms were statistically significant from the hypergeometric approach. We observe an ROC curve with an area under curve (AUC) of 0.625, indicating this performs better than random classification, however it is not considered a very strong classifier. For the subsets of the full dataset, we observe an AUC of 0.662, 0.636 and 0.613 from the subset of 24 samples, 12 samples and 6 samples respectively. The ROC curves can be seen in Figure 3.9, 3.10, 3.11, and 3.12.

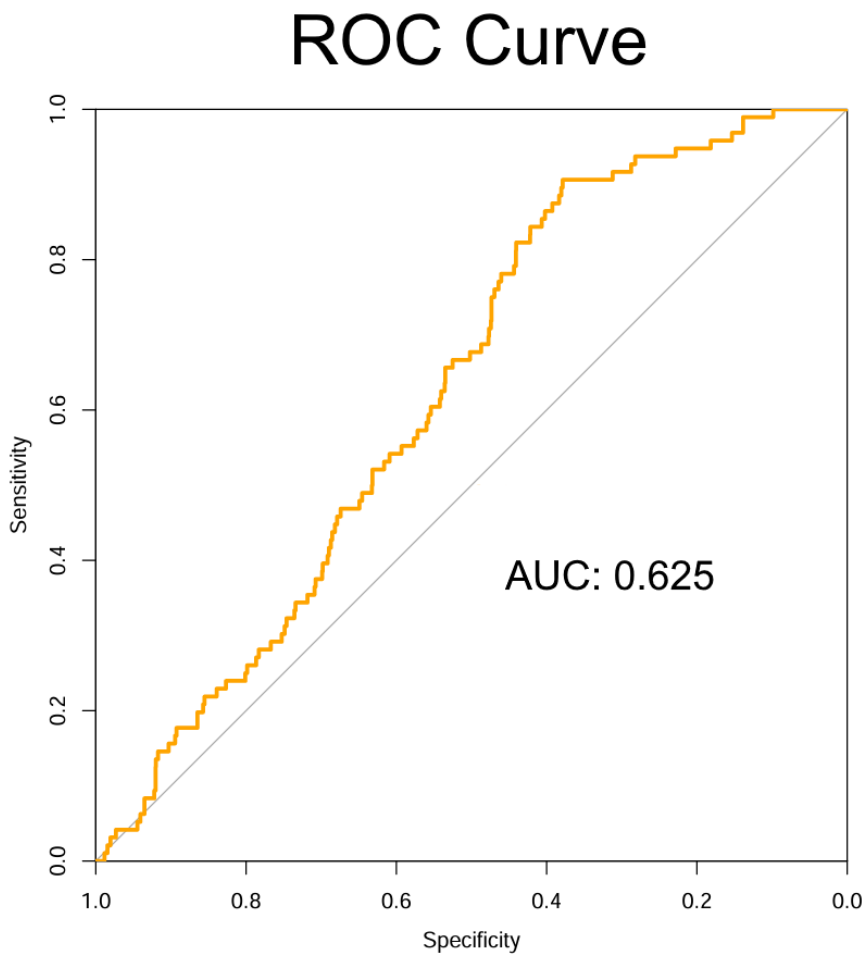

Figure 3.9 The ROC curve which evaluates Deep SHAP importances being used for classifying terms with statistically significant p-values from the hypergeometric ORA. Values based on the BINN trained on the dataset containing all samples of the lung cancer dataset. AUC of 0.625.

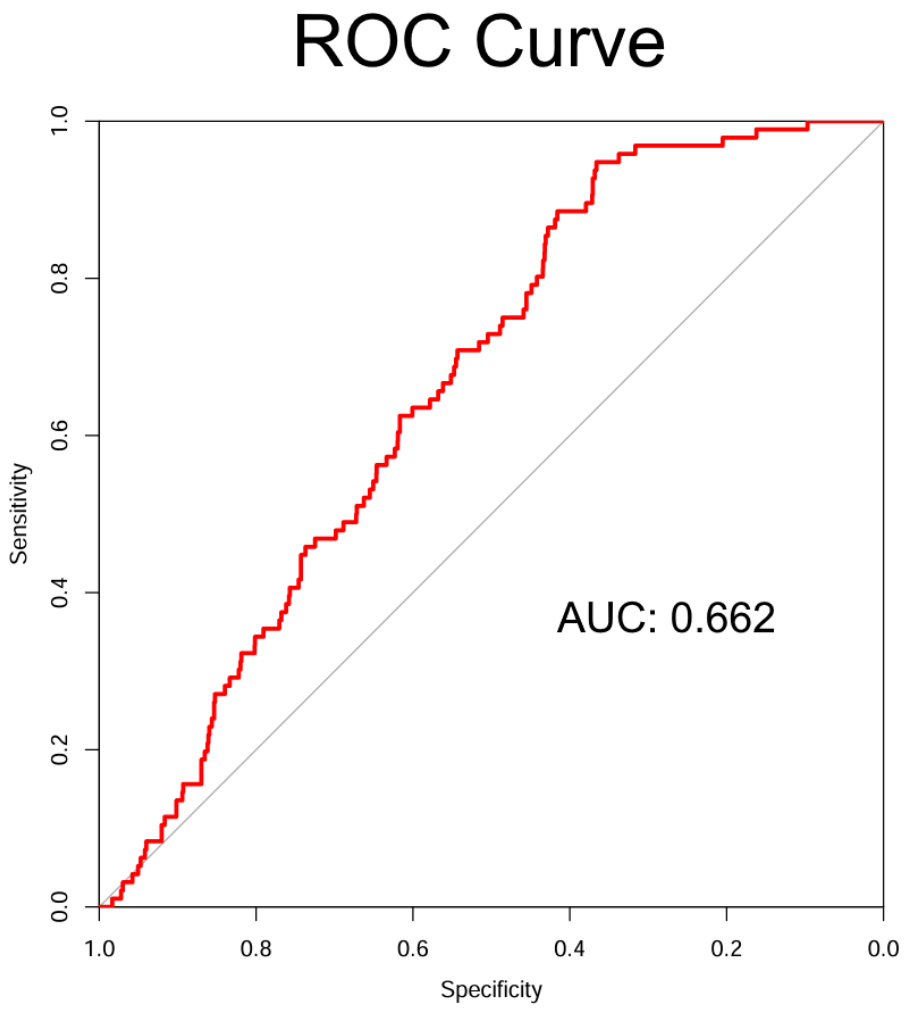

Figure 3.10 The ROC curve which evaluates Deep SHAP importances being used for classifying terms with statistically significant p-values from the hypergeometric ORA. Values based on the BINN trained on the dataset containing 24 samples of each phenotype from the lung cancer dataset. AUC of 0.662.

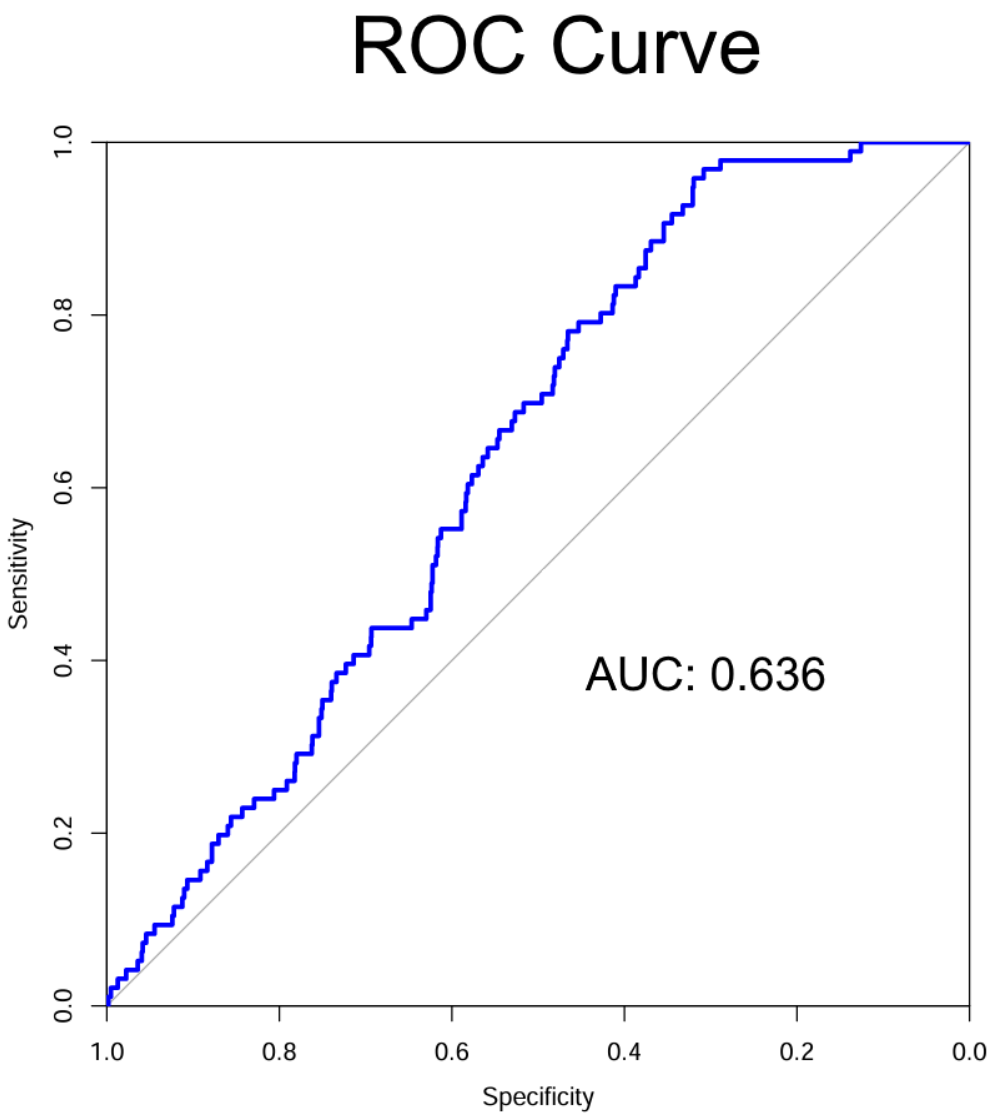

Figure 3.11 The ROC curve which evaluates Deep SHAP importances being used for classifying terms with statistically significant p-values from the hypergeometric ORA. Values based on the BINN trained on the dataset containing 12 samples of each phenotype from the lung cancer dataset. AUC of 0.636.

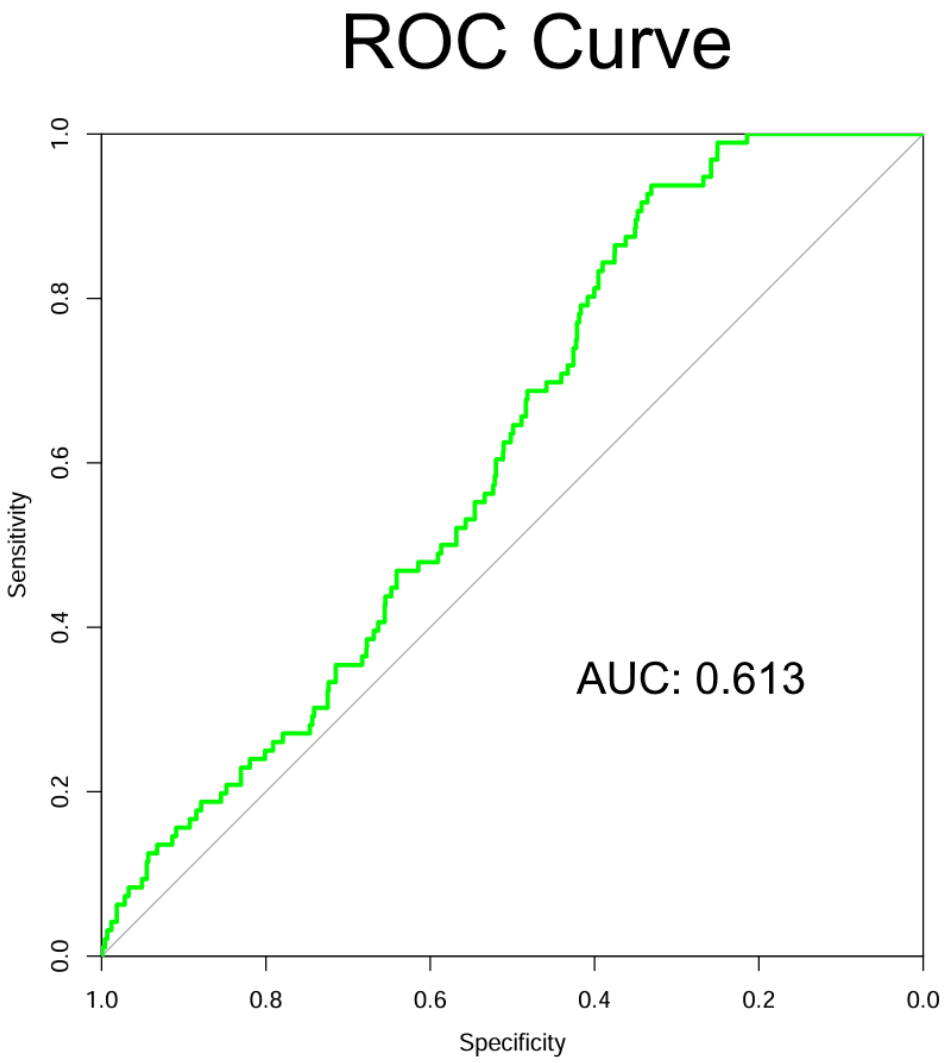

Figure 3.12 The ROC curve which evaluates Deep SHAP importances being used for classifying terms with statistically significant p-values from the hypergeometric ORA. Values based on the BINN trained on the dataset containing 6 samples of each phenotype from the lung cancer dataset. AUC of 0.613.

We finally examined the top 25 terms by Deep SHAP importance produced by the BINN trained on all datasets and summarized them in Table 3.2. We also examine the most statistically significant terms from the hypergeometric approach in Table 3.3. While 25 terms is not a gold standard of evidence, we note there are no terms which exist in each table.

Table 3.2 The GO term ID, GO term name and Deep SHAP importance value for the 25 terms with the highest Deep SHAP importance from the dataset containing all samples.

GO Term ID	GO term name	Deep SHAP importance value
GO:0014806	smooth muscle hyperplasia	0.226607312
	negative regulation of mesenchymal to epithelial transition involved in metanephros	
GO:0003340	morphogenesis	0.2263766979
GO:0015228	coenzyme A transmembrane transporter activity	0.21220672
GO:0003322	pancreatic A cell development	0.1976579208
GO:0004484	mRNA guanylyltransferase activity	0.189344988
GO:0014737	positive regulation of muscle atrophy	0.1834787808
GO:0002130	wobble position ribose methylation	0.1812334138
GO:0009234	menaquinone biosynthetic process	0.1788427674
	right ventricular compact myocardium	
GO:0003226	morphogenesis	0.1782336617
GO:0015607	ABC-type fatty-acyl-CoA transporter activity	0.1763756678
GO:0001681	sialate O-acetylesterase activity	0.1731070791

GO:0018117	protein adenylylation	0.1719731692
	N-acetylglucosaminylphosphatidylinositol	
GO:0000225	deacetylase activity	0.1658899827
GO:0010370	perinucleolar chromocenter	0.1632307631
GO:0018283	iron incorporation into metallo-sulfur cluster	0.1625238154
GO:0003342	proepicardium development	0.1622625039
GO:0004063	aryldialkylphosphatase activity	0.1619848318
GO:0008493	tetracycline transmembrane transporter activity	0.1619362442
GO:0007501	mesodermal cell fate specification	0.1607630019
GO:0006501	C-terminal protein lipidation	0.1602280216
GO:0004441	inositol-1,4-bisphosphate 1-phosphatase activity	0.1596337576
GO:0003097	renal water transport	0.1577941018
GO:0014015	positive regulation of gliogenesis	0.1561256749
GO:0018063	cytochrome c-heme linkage	0.1552465159
GO:0007093	mitotic cell cycle checkpoint signaling	0.154350135

Table 3.3 The GO term ID, GO term name and Hypergeometric p-value for the 25 terms with the lowest p-values from the dataset containing all samples.

GO Term ID	GO Term Name	Hypergeometric p-value
GO:0005622	intracellular anatomical structure	1.23E-86
GO:0005634	nucleus	3.04E-56
GO:0019219	regulation of nucleobase-containing compound metabolic process	1.25E-46
GO:0005654	nucleoplasm	1.28E-39
GO:0003677	DNA binding	2.98E-27
GO:0006139	nucleobase-containing compound metabolic process	9.54E-24
GO:0006996	organelle organization	1.67E-21
GO:0005737	cytoplasm	3.63E-21
GO:0016070	RNA metabolic process	3.90E-21
GO:0003676	nucleic acid binding	2.33E-20
GO:0009593	detection of chemical stimulus	3.03E-18
GO:0004888	transmembrane signaling receptor activity	3.52E-18
GO:0001067	transcription regulatory region nucleic acid binding	7.89E-18
GO:0005515	protein binding	2.63E-17
GO:0005576	extracellular region	5.24E-17
GO:0007606	sensory perception of chemical stimulus	6.63E-17
GO:0004930	G protein-coupled receptor activity	1.16E-16

GO:0004984	olfactory receptor activity	5.14E-16
GO:0006974	DNA damage response	1.37E-15
GO:0003700	DNA-binding transcription factor activity	3.93E-15
GO:0015630	microtubule cytoskeleton	7.54E-15
GO:0016043	cellular component organization	2.64E-14
GO:0005829	cytosol	4.79E-14
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	1.56E-12
GO:0007049	cell cycle	3.23E-12

3.4 Discussion

While we were able to train a BINN with high classification accuracy in most cases, we do not observe a very strong ROC or Spearman correlation between Deep SHAP importances and the p-values from a hypergeometric approach. In Chapters 4 and 5, we will perform similar analyses and contrast them with the results seen here.

Our neural network implementation using the library developed by Hartman et al. showed healthy metrics, as a reducing and stable loss curve was observed. However, the training curve for 24 samples began to rise which may imply there was some degree of overfitting in this model (Fleuret, 2024). The smaller sample sizes indicated potentially less accurate models.

Of note is the high classification accuracy of samples did not result in the ability to reproduce the results of traditional approaches to pathway analysis; however, we were able to achieve high classification accuracy at lower sample sizes as well. Except in the case of 24

samples, the AUC of the smaller sample sizes were lower than that of the full dataset, however they were all relatively similar. This may indicate that BINNs can produce consistent results with low sample sizes. If true, this stability will be useful for translational research. If BINNs provide similar results when looking at a study with six samples per treatment group when compared to 49 and 68, that would be a great benefit to research to rare diseases and other cases where samples are precious. Additionally, observing the scatter plot of log transformed p-values and importances revealed a low correlation between the different metrics.

While we only examined four datasets, the patterns in classification accuracy seem unclear. The dataset which consisted of 24 samples of each lung cancer obtained the highest AUC, yet appeared to have the least stable loss curve. This may potentially lead one to believe that this high AUC is erroneous. Potential future directions may be in investigating the importances of a BINN trained on the same dataset at different epochs, however comparison with the hypergeometric does not show a strong pattern between the quality of neural network training and the ability to reproduce traditional approaches. For the remainder of the discussion, we will pragmatically choose the same hyperparameters used here and explore the role of other factors in assessing the usefulness of BINNs.

Chapter 4: Clustering applied to expression data in biologically informed neural networks

4.1 Motivation

Clustering is a technique which sees use in omics, where a dataset is broken into separate groups of features based on a shared criterion (Karim et al., 2020). It is common to use mathematical clustering of features as part of a process to reduce the dimensionality of data and produce modules of genes. There are reasons to consider clusters of correlated groups of genes as clusters. It has been observed that genes express in sets, and that genes that co-express tend to contribute to similar biological processes (Raina et al., 2022).Treating genes as a cluster may better represent biology, and this has led to clustering being used in place of statistical differential analysis for pathway analysis in some instances. Additionally, in the context of research, not all genes will be affected by the condition of interest. As such, there is rationality to consider which cluster of genes is most relevant to the experiment.

Limiting input features (in the context of omics, features are measured genetic entities) to neural networks is generally recommended, with conventional recommendations existing of a 10-to-1 ratio of samples per input feature (Alwosheel et al., 2018). Yet a balance between samples and features is unrealistic within omics datasets. However, clustering may offer a way to segregate a dataset into smaller, more relevant modules, which may in turn be more appropriate for input into a neural network.

High dimensionality in other bioinformatic tools has been handled with clustering techniques, with improvement in classification observed. Single cell RNA (scRNA) sequencing analysis uses clustering as part of the standard Seurat workflow by clustering data by predicted cell type of origin (Butler et al., 2018). Additionally, in bioinformatic workflows that incorporate machine learning, clustering has also been applied. Frameworks like omics-CNN use hierarchical clustering before applying a neural network model to omics data (Zompola et al., 2023).

Weighted gene co-expression network analysis (WGCNA) is a technique which implements clustering and sees widespread use within bioinformatics. It is a correlation-based clustering algorithm developed in 2008 which has been cited over 24,000 times at the time of writing (Langfelder & Horvath, 2008; Vahabi & Michailidis, 2022). It sees many uses, has been used as an alternate approach to finding DE genes (Zeng et al., 2021), and has also been used to reduce the amount of input data into a machine-learning method, which improved the classification of the algorithm (Gakii et al., 2023).

As our initial investigation revealed, Deep SHAP values alone may not immediately recapitulate the traditional hypergeometric approach; we aim to investigate clustering as a means to boost relevant biological insight. To this end, we examined if training BINNs on all clusters of the same dataset will show that certain clusters can recapitulate the hypergeometric approach. As before, we also wish to see the role of sample sizes in a BINNs ability to classify and recapitulate the information we see in a traditional approach.

4.2 Materials and Methods

4.2.1 RNA-expression data

As before, we used the lung cancer samples from the human cancer cell-line dataset from the CCLE. We additionally used the subsets of this dataset described before.

4.2.2 WGCNA

To test the effects of clustering on BINN input data, WGCNA was implemented using the WGCNA R library version 1.73. A workflow was generated in R and applied to all datasets to generate modules of genes. When filtering outliers in the preprocessing, we supplied the goodSampleGenes() function with a variance cutoff, which was higher for lower sample sizes in order to have successful clustering. This function will remove genes with variance lower than

the supplied cutoff. Cutoffs are listed in Table 4.1. When creating clusters with the blockwiseModules() function as described in documentation, we used a max block size of 1,000, a minimum module size of 30, and an unsigned topological overlap matrix (TOM). Similarity between genes as determined by the TOM is used to determine coexpression. The WGCNA algorithm is applied in a series of blocks, the maximum number of input features considered for clustering at a time. A dendrogram representing the TOM can be seen in Figure 4.1 for nine of the blocks analyzed.

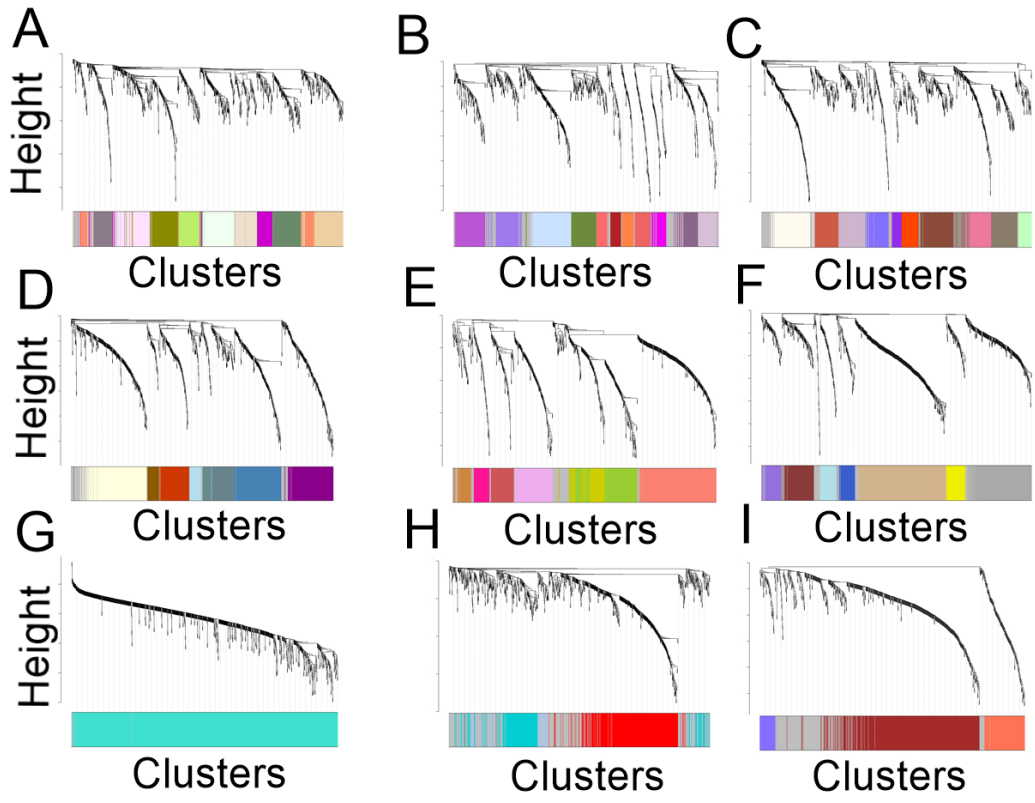

Figure 4.1 The dendrogram based on the TOM for different blocks from the analysis of the full dataset. Different clusters in the same block are coloured differently.

Table 4.1 The variance cutoff used in the goodSampleGenes() function in WGCNA for the various sample sizes

Sample size	Variance cutoff
All samples	1e-10 * (maximum absolute value in expression level)
24	1e-10 * (maximum absolute value in expression level)
12	1e-5 * (maximum absolute value in expression level)
6	1e-4 * (maximum absolute value in expression level)

4.2.3 BINNs applied to clusters

Each module generated by WGCNA was used as the input to train a BINN as implemented previously. A “Cumulative Module” would be formed for each dataset, in which the Deep SHAP importances produced by each module would be normalized to 1 by dividing all importances by the highest importance, then combined into one module. If repeat terms existed, the entry with the highest importance would be kept. We also examine the top 25 terms from the cumulative module formed from the dataset containing all samples.

4.2.4 Comparison

Similar to the exploration in chapter 3, each BINN was examined for its classification metrics, ROC curve, and spearman correlation. All visualizations and calculations were performed in R.

4.3 Results

4.3.1 WGCNA clustering

The final number of clusters produced from each sample size, along with the range of sizes can be seen in the table below. This also shows the number of genes considered valid for WGCNA and the module sizes after applying preprocessing.

Table 4.2 A summary of the running of WGCNA on the different datasets, including the number of samples left after filtering low variance genes, the number of modules.

	6 samples	12 samples	24 samples	All samples
Good sample genes	4686	23,206	24,335	24,228
Number of modules	36	101	93	170
Range of module sizes	32-354	31-996	35-926	30-1323
Average module size	117.429	213.576	208.478	120.917

4.3.2 BINN training

The subsequent BINNs trained on each cluster were examined to see the patterns in validation loss curves. These can be seen in Figures 4.2, 4.3, 4.4, and 4.5. This shows a divergence from the previous implementation, in which many of the clusters show poor training, as they have less stable and increasing loss curves. There are, however, clusters for which a stable, decreasing, validation loss curve can be seen.

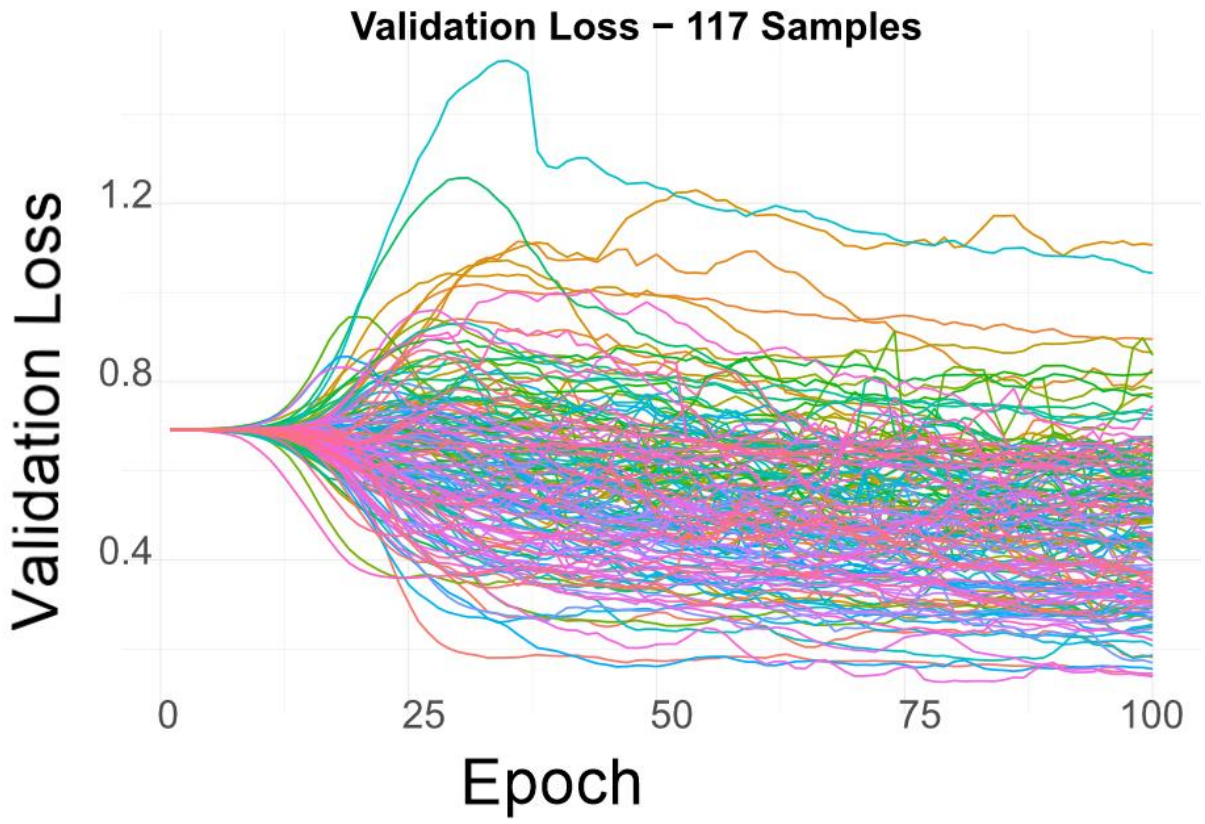

Figure 4.2 The validation loss curves for BINNs trained on all clusters from the dataset which contained all samples from the two lung cancer phenotypes.

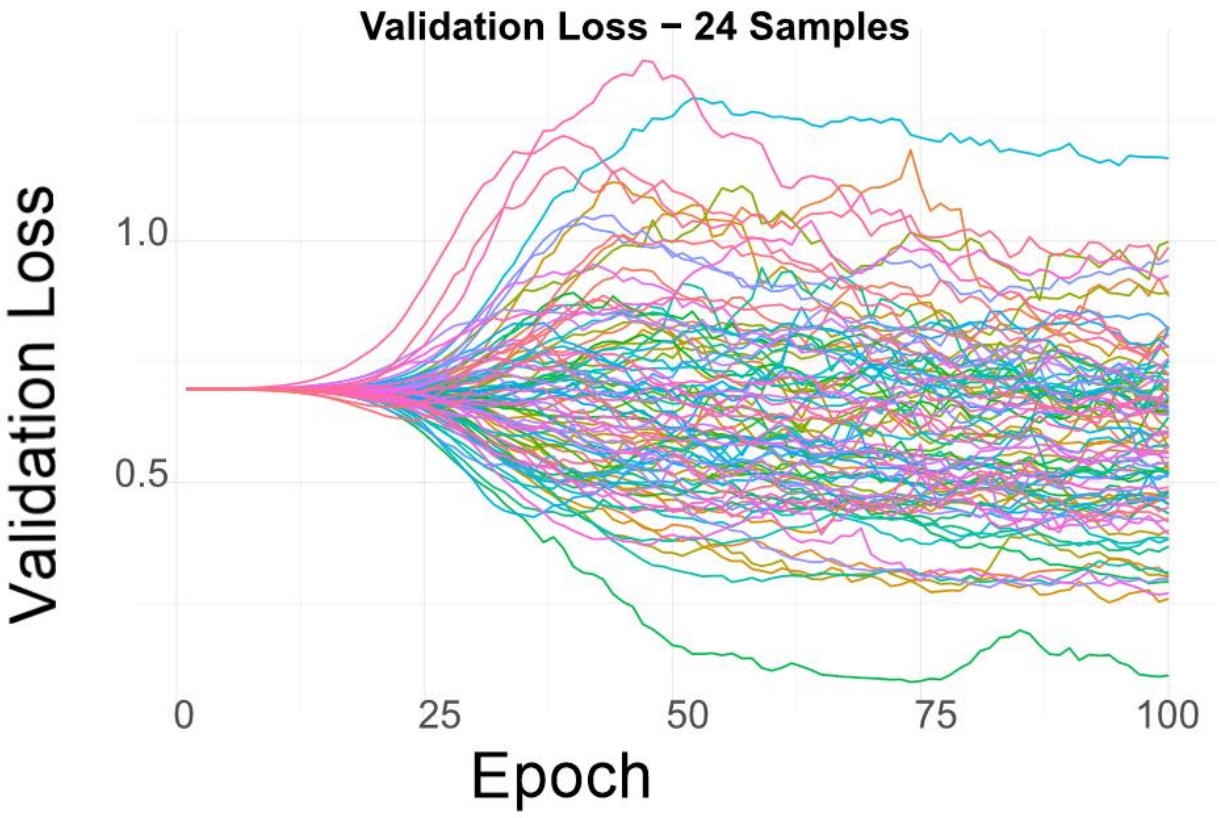

Figure 4.3 The validation loss curves for BINNs trained on all clusters from the dataset which contained 24 samples from each lung cancer phenotypes.

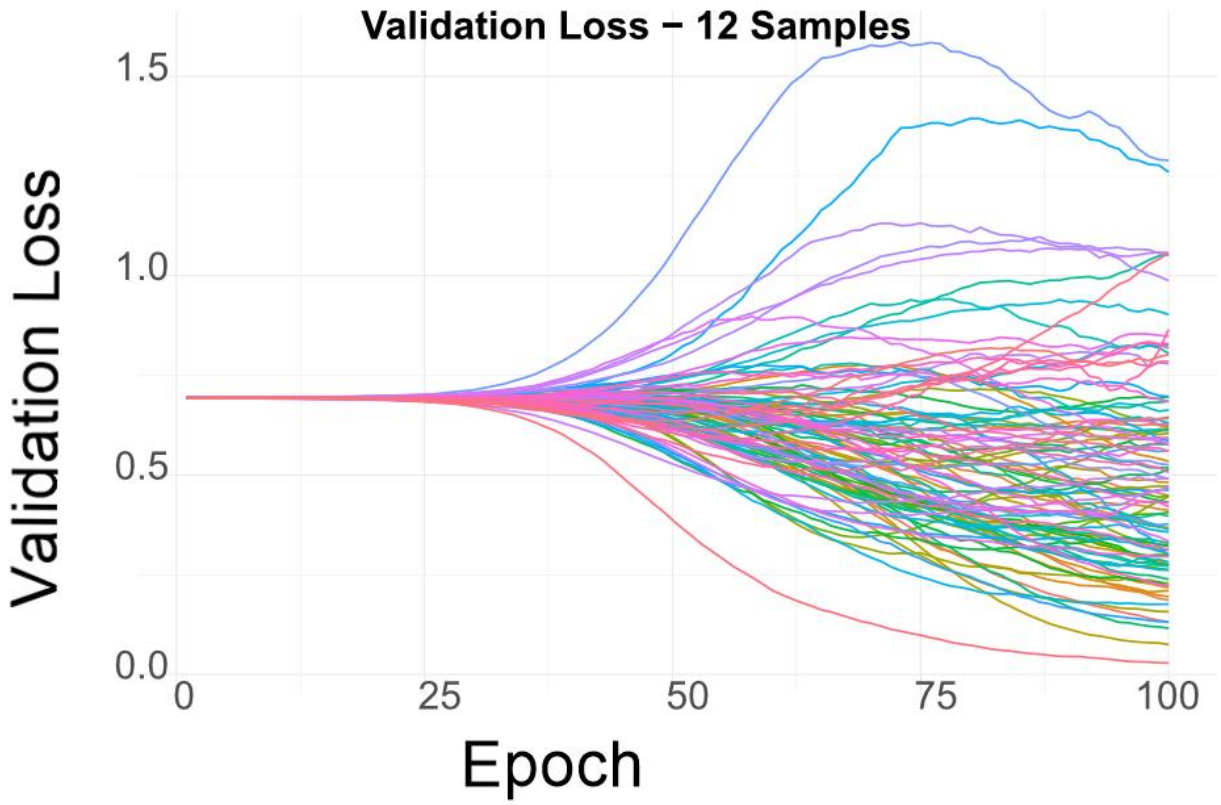

Figure 4.4 The validation loss curves for BINNs trained on all 99 clusters from the dataset which contained 12 samples from each lung cancer phenotype.

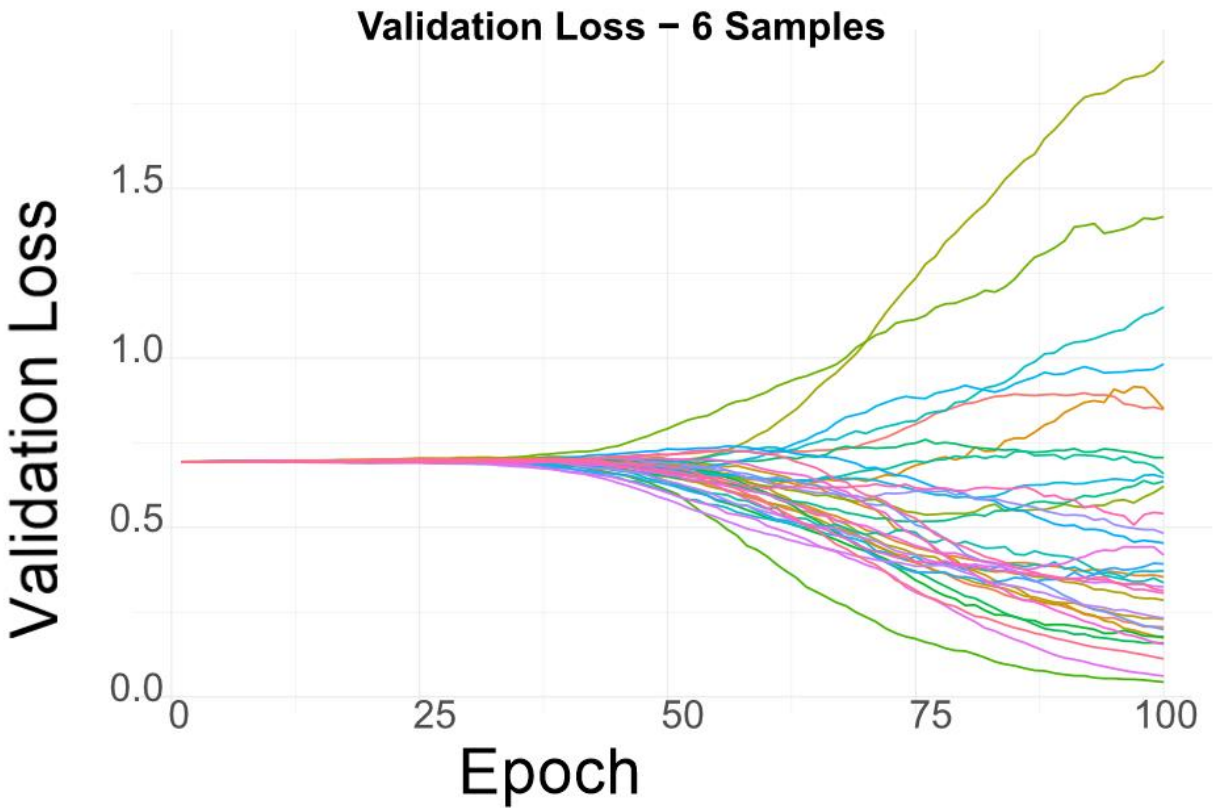

Figure 4.5 The validation loss curves for BINNs trained on all clusters from the dataset which contained 6 samples from each lung cancer phenotype.

Alongside this, a wide spread of validation accuracies was observed, which can be seen in Table 4.3. Also examined was the relationship between the validation accuracy (the percentage of samples in validation dataset classified correctly) and the AUC of all modules across all datasets in a scatter plot in Figure 4.6, which showed a positive relationship, however this also showed a low adjusted R^2 value of 0.01704.

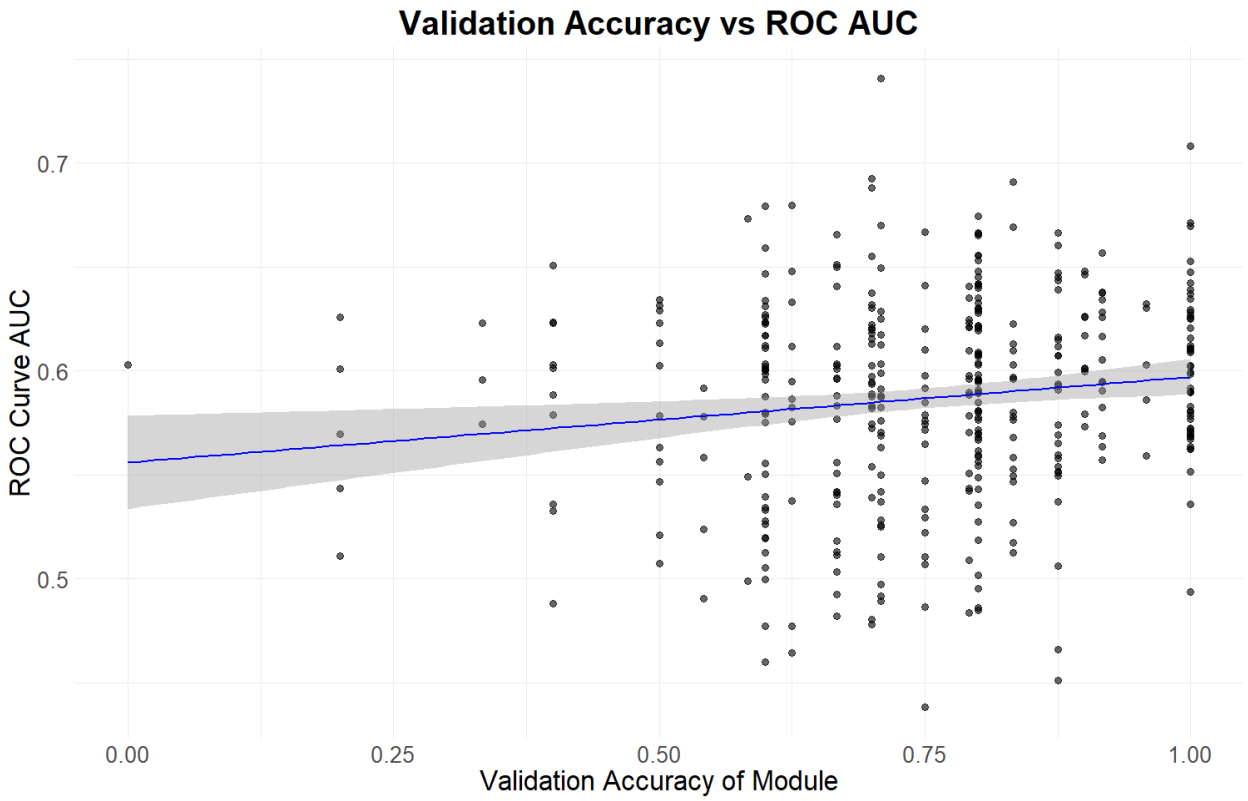

Figure 4.6 A scatter plot showing the validation accuracy on the X-axis and AUC of that cluster's ROC curve for all clusters from all datasets.

Table 4.3 A summary of the machine-learning training metrics from trained BINNs on all clusters produced by WGCNA.

Number of samples	6 samples	12 samples	24 samples	All samples
Minimum validation accuracy	0	0.20	0.4	0.3333
Maximum validation accuracy	1	1.00	1	0.9583
Median validation accuracy	1	0.80	0.7	0.7917
Mean validation accuracy	0.803932	0.74	0.726087	0.79736

To examine the effect of clustering on the relationship between p-values and the importances, we applied the Spearman correlation again to quantify the relationship between importance and statistical significance. The results for the cumulative dataset formed from all clusters can be seen in Table 4.4. We additionally produce ROC curves to calculate if importances in the cumulative dataset functioned better than the non-clustered data to classify terms deemed statistically significant from the hypergeometric approach. The ROC curves for the cumulative modules can be seen in Figure 4.7.

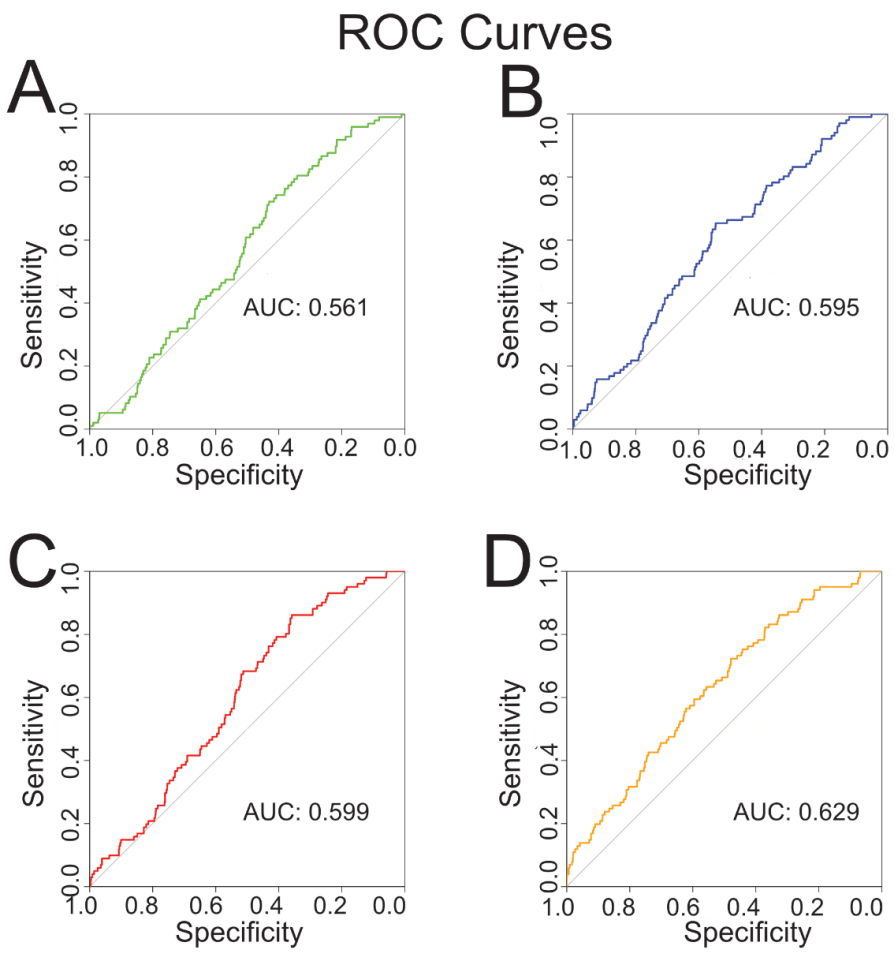

Figure 4.7 The ROC curves showing Deep SHAP importances ability to predict statistically significant hypergeometric p-values for the “Cumulative Module” formed from each module in a dataset. Panel A is formed from the dataset containing 6 samples, with an AUC of 0.561. Panel B is formed from the dataset containing 12 samples, with an AUC of 0.595. Panel C is formed from the dataset containing 24 samples, with an AUC of 0.599. Panel D is formed from the dataset containing all samples, with an AUC of 0.629.

Table 4.4 A summary of calculating the AUC for a ROC curve and the Spearman correlation for each dataset formed from accumulating all clusters from WGNCA.

Dataset	AUC of ROC	Spearman Correlation	Spearman P-value
All samples	0.5775958	-0.1760057	2.2e-16
24 samples	0.5952061	-0.1428115	7.797e-15
12 samples	0.6143549	-0.1512095	2.2e-16
6 samples	0.5427351	-0.06054823	0.003362

The ROC curves of the individual modules which produced the highest and lowest AUC for each of the four datasets are listed in Table 4.5. Visualizations of these ROC curves can be seen in Figures 4.8, 4.9, 4.10 and 4.11.

Table 4.5 Summary of the ROC curves graphing Deep SHAP importances classifying statistically significant terms from the hypergeometric approach. Results show the AUC for the ROC curves for the highest and lowest three performing clusters generated by WGCNA.

Dataset	Top module 1	Top module 2	Top module 3	Bottom module 1	Bottom module 2	Bottom module 3
All samples	0.741	0.69	0.68	0.438	0.451	0.465
24 samples	0.693	0.69	0.669	0.478	0.48	0.486
12 samples	0.679	0.67	0.671	0.46	0.477	0.485
6 samples	0.708	0.65	0.65	0.503	0.513	0.518

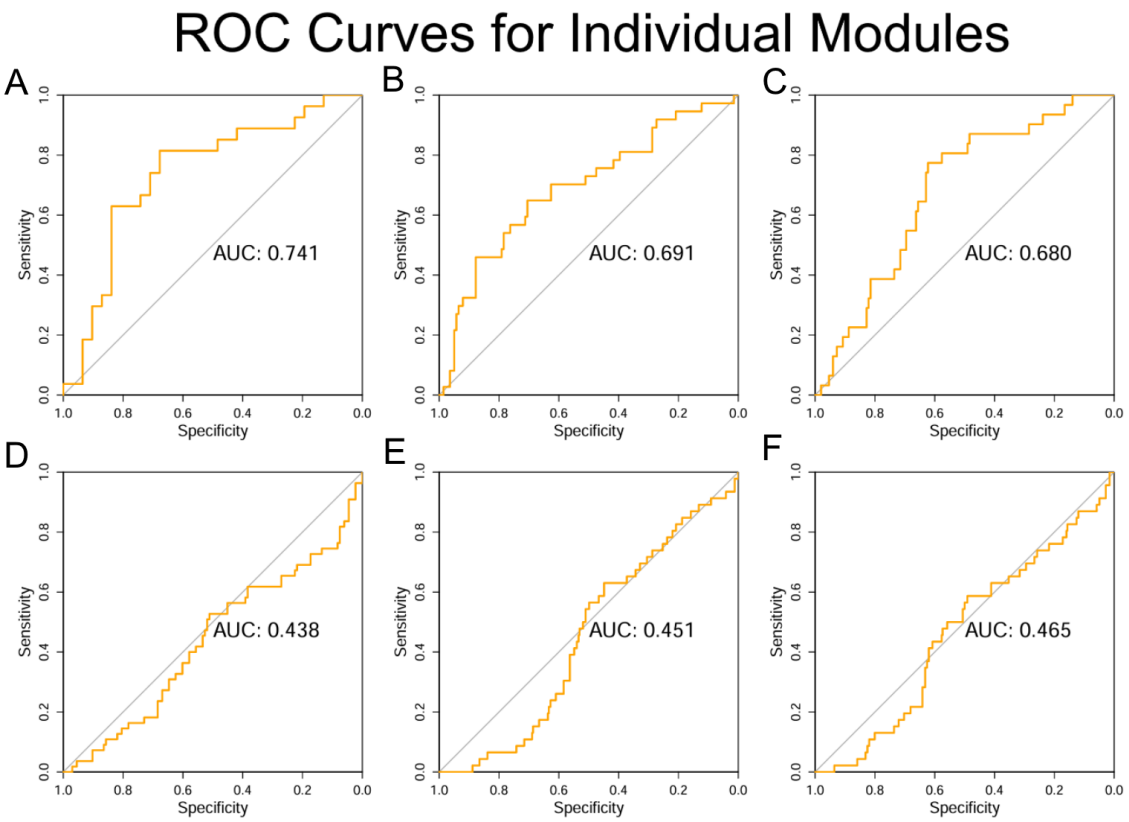

Figure 4.8 The ROC curves showing Deep SHAP importances ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the full dataset containing all samples. Panels A, B, C are the ROC curves generated from the modules with the 3 highest AUC values, and D,E, F are generated from the modules with the 3 lowest AUC values. The highest AUC observed for a module is 0.741, and the lowest is 0.438.

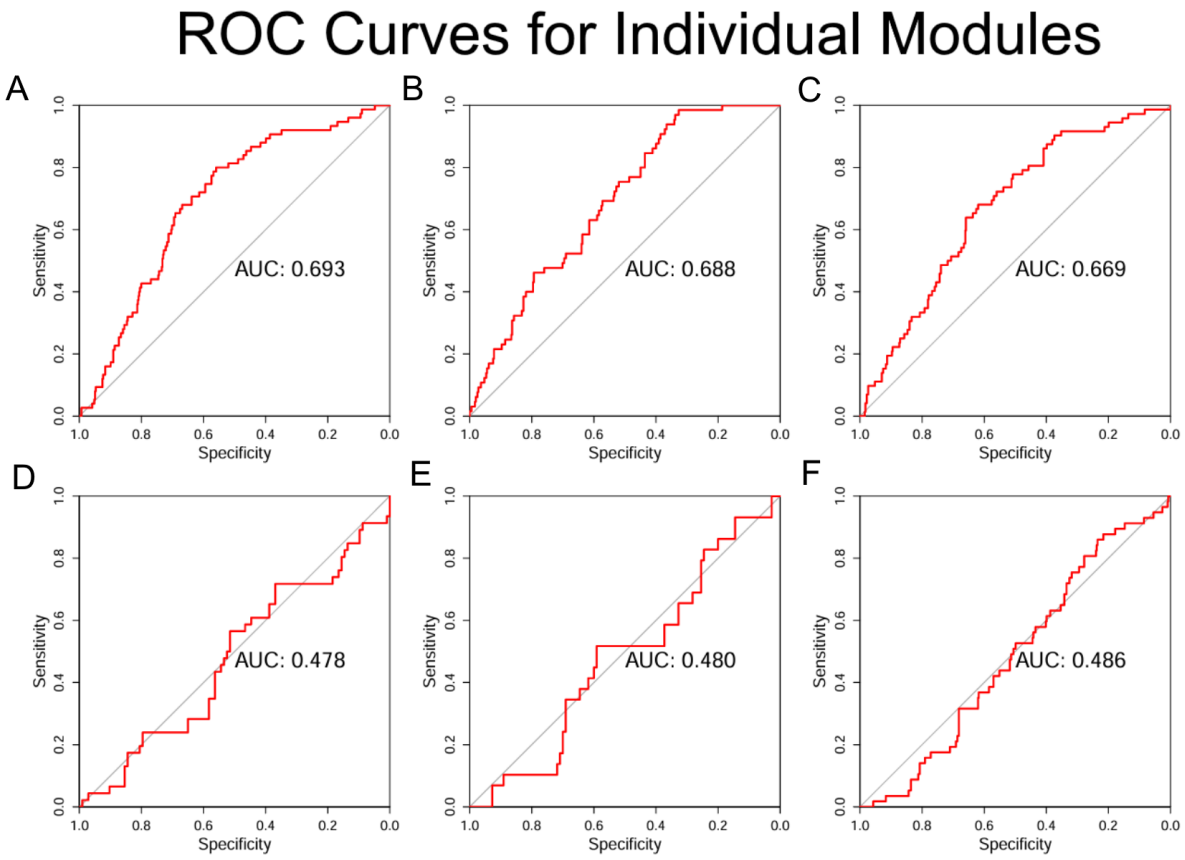

Figure 4.9 The ROC curves showing Deep SHAP importances ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 24 samples from each lung cancer. Panels A, B, C are the ROC curves generated from the modules with the 3 highest AUC values, and D,E, F are generated from the modules with the 3 lowest AUC values. The highest AUC observed for a module is 0.693, and the lowest is 0.478.

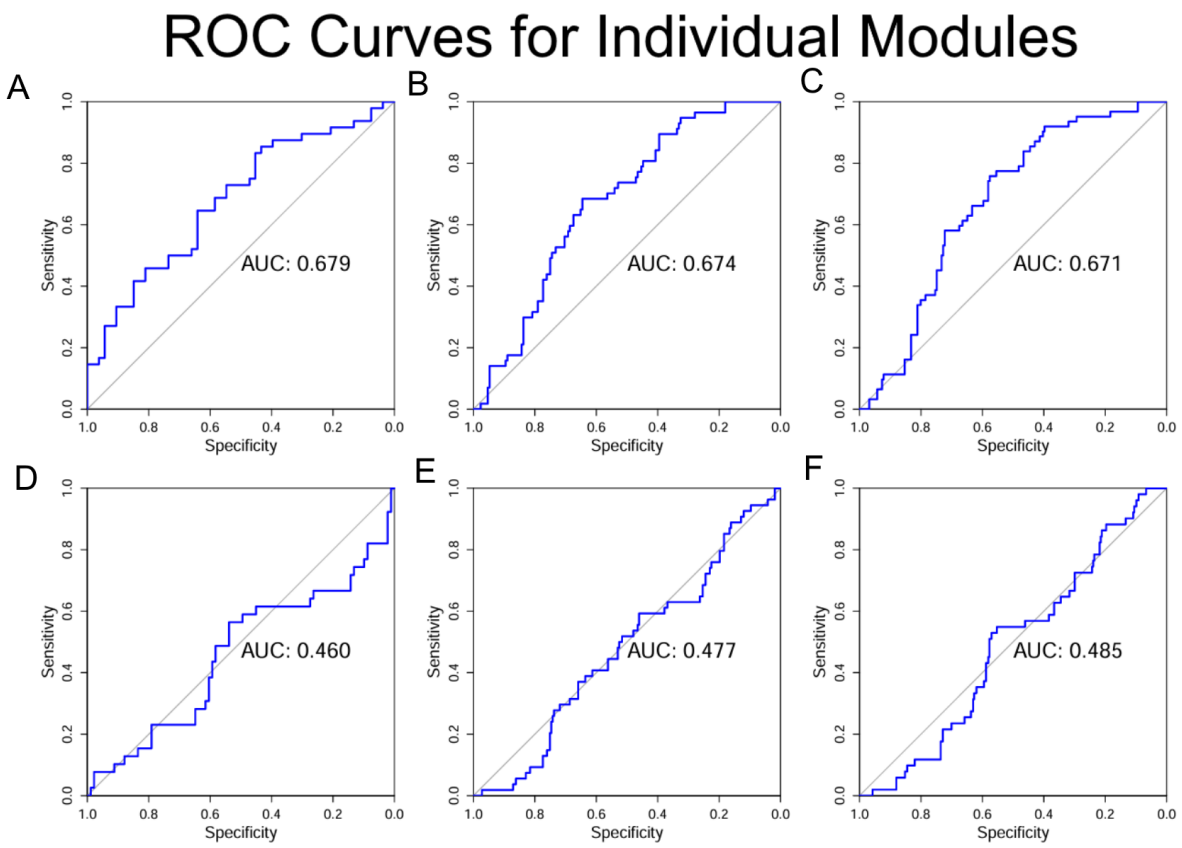

Figure 4.10 The ROC curves showing Deep SHAP importances ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 12 samples from each lung cancer. Panels A, B, C are the ROC curves generated from the modules with the 3 highest AUC values, and D,E, F are generated from the modules with the 3 lowest AUC values. The highest AUC observed for a module is 0.679, and the lowest is 0.460.

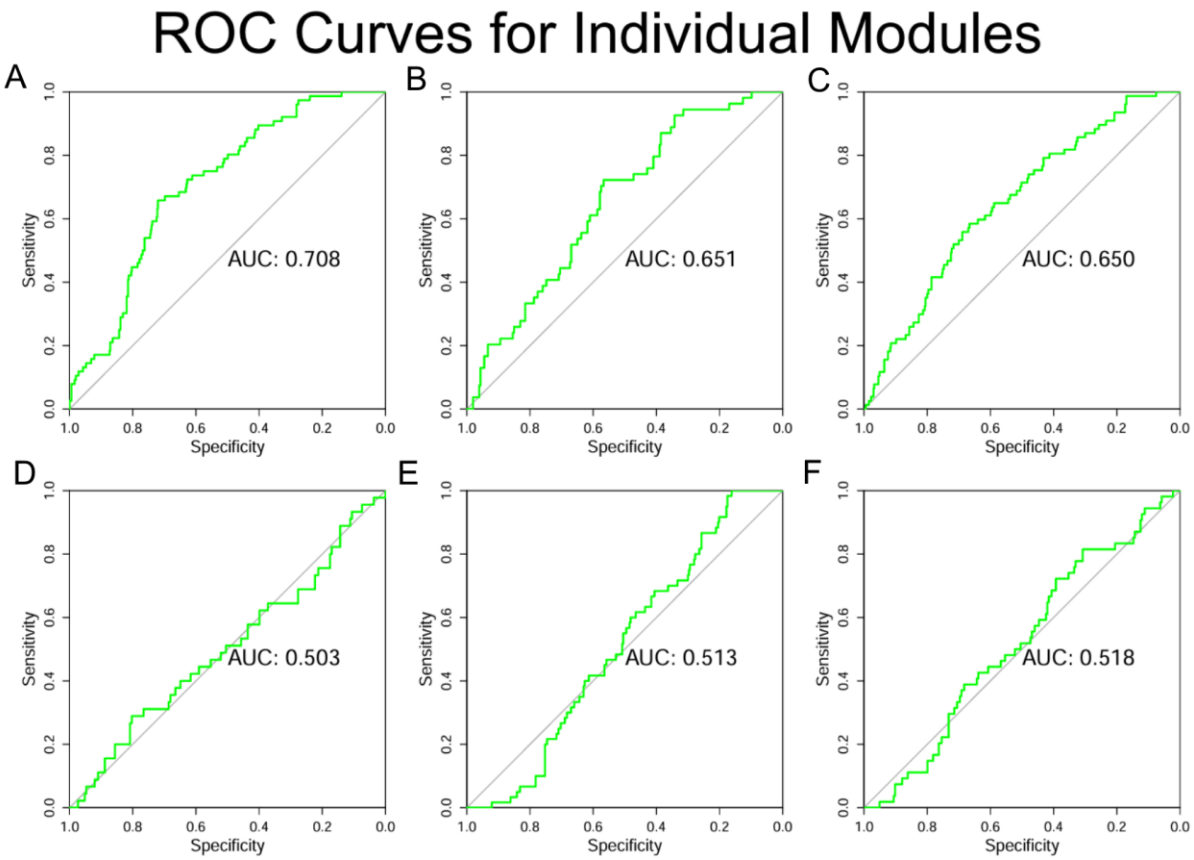

Figure 4.11 The ROC curves showing Deep SHAP importances ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 6 samples from each lung cancer. Panels A, B, C are the ROC curves generated from the modules with the 3 highest AUC values, and D,E, F are generated from the modules with the 3 lowest AUC values. The highest AUC observed for a module is 0.708, and the lowest is 0.503.

The 25 terms with the highest importance from the “Cumulative Module” for the dataset containing all samples can be in Table 4.6.

Table 4.6 The GO term ID, GO term name and normalized Deep SHAP importance for the 25 terms with the highest Deep SHAP importance from the “Cumulative Module” for the dataset containing all samples.

GO Term ID	GO term name	Normalized Deep SHAP Importance
GO:0003678	DNA helicase activity	1
GO:0005215	transporter activity	1
GO:0005635	nuclear envelope	1
GO:0005886	plasma membrane	1
GO:0006869	lipid transport	1
GO:0006886	intracellular protein transport	1
GO:0015990	electron transport coupled proton transport	1
GO:0016604	nuclear body	1
GO:0002159	desmosome assembly	0.9608744718
GO:0015297	antiporter activity	0.9446328695
GO:0007271	synaptic transmission, cholinergic	0.9134745725
GO:0007283	spermatogenesis	0.8923656998
GO:0003724	RNA helicase activity	0.8753178114
GO:0008544	epidermis development	0.865534216
GO:0007498	mesoderm development	0.8556243957

GO:0001764	neuron migration	0.854141292
GO:0007596	blood coagulation	0.8530508305
GO:0000209	protein polyubiquitination	0.8246577241
GO:0007131	reciprocal meiotic recombination	0.8151083994
GO:0006612	protein targeting to membrane	0.8135210814
GO:0005353	fructose transmembrane transporter activity	0.8100017647
GO:0007088	regulation of mitotic nuclear division	0.8045326472
GO:0000785	chromatin	0.8038703785
GO:0005874	microtubule	0.8009995632
GO:0007595	lactation	0.8008018231

4.4 Discussion

Training BINNs on the individual clusters determined by WGCNA show some modules have a much higher AUC for their ROC curve. This supports the idea that different groups of the transcriptome have different relevance to conditions. Conversely, we would expect some modules to have a very low similarity to the hypergeometric approach, as not all of the transcriptome is relevant to the condition observed. Using the WGCNA algorithm did provide a mechanistic way to remove genes of low variance, and limit the scope of input features.

However, comparing modules to one another may not be very helpful in better understanding BINNs. It may be the case that the variation in AUC and validation accuracy can be better explained through the size of the module, another metric, or even random chance.

The wide range in validation loss, as well as the weak relationship between validation accuracy and AUC does not give us a stronger view of how the health of neural net training relates to biological relevance. Further complicating this comparison is the wide range in module and sample sizes, which may be confounding variables with both AUC and validation accuracy. A comparison of modules of the same size using different hyperparameters may show more insight into this relationship for future works.

When examining the cumulative module for each dataset, we do not see a great change in how the ROC curve performs. We do observe an increase of AUC with sample size, however it does not show great improvement over using the Deep SHAP importances of the non-clustered dataset as explored in chapter 3, and in some dataset shows a drop in AUC. This is showing that ORA recapitulation, as measured by the ROC curves, shows an increase when applying clustering. However, this increase is only seen for the dataset composed of all samples, and not the subsets of this data. Additionally, we see no overlap in the top 25 terms for the cumulative module formed from the dataset containing all samples.

Chapter 5: Point Estimates Applied to Deep SHAP Importances

5.1 Motivation

Previous results may indicate that BINNs partially recapitulate traditional approaches depending on the input data, even at lower sample sizes. However, especially given the complexity of factors that can impact Deep SHAP values such as neural network training, we propose other factors which may increase the interpretability of Deep SHAP values.

Importances are fundamentally different from p-values. P-values correspond to a statement of statistical confidence, while importances do not. It is not clear how much confidence one can place in a statement of importance; however, Hartman has explored comparing the variance of the importance of a term when training multiple identical BINNs on the same dataset. This showed that training multiple BINNs using the same input data and network structure resulted in varying importances. There exploration concluded that low variance of a term's importance should improve one's confidence in that term (2025). This addresses one of the pitfalls described by Chen et al. in their review on explainable AI for bioinformatic applications. In this review they suggest determining which terms are more robust than others by the stability of their importance across training results and training parameters (2024). However, this does not propose a standard or analytic statement on the relationship between the stability of a term and relevance to interpretation.

We propose an alternate approach to examining the importances. Described later in the methods section will be an approach of applying a non-parametric permutation test to accept or reject a null hypothesis on the value of a Deep SHAP importance within a BINN. The null hypothesis for this test is as follows: if there is no difference between phenotypes, the importance of a term should not be different if the phenotypic labels are scrambled relative to the input data. That is to say, if we scramble the phenotypic labels of the dataset and there exists no difference in importance, we fail to reject the null hypothesis. Since in this case, we

know there are real differences in phenotypic conditions, and scrambling labels did not affect the importance of the term, we have less confidence in the importance of this term representing observed reality. An illustration of this general work flow is shown in Figure 5.1.

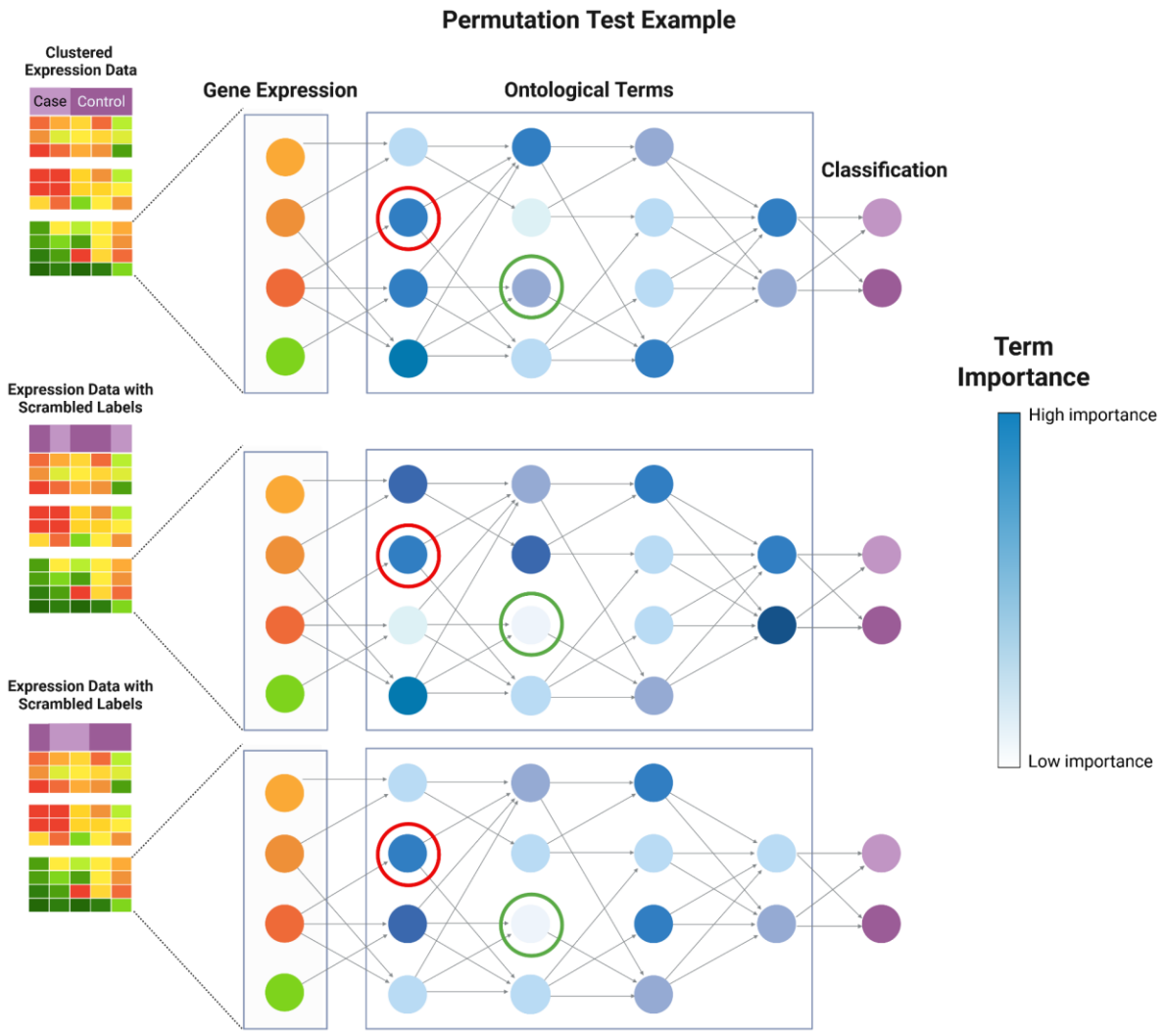

Figure 5.1 An illustration of the permutation test applied to BINNs. In the image three BINNs are trained. The top BINN is trained on the omic data collected, but the other 2 are trained on the same data with the labels of case and control scrambled with replacement. If we examine the term in the network circled in red, the BINN trained on the data with unscrambled labels and the BINNs trained on data with scrambled labels have a similar importance. In contrast, the term circled in green shows a much lower importance in the BINNs trained on data with scrambled

labels compared to the BINN trained on data with unscrambled labels. While the importance of the node in the green circle is lower than the importance of the node in the red circle, we assert the importance of the node in the green circle more accurately represents reality.

Permutation tests are performed by comparing statistics calculated on experimental data to statistics calculated from the same data. This is done under the assumption that making input data intentionally random should produce a distribution of importances under the null hypothesis that there is no difference between the two groups.

This is of value in bioinformatics, as this approach does not require statistical assumptions regarding the distribution of errors, such as is done in t-tests and analysis of variance. These assumptions can be difficult to establish in complex data, such as data from omics studies (Gel et al., 2015). This is further complicated by the issue of widespread small sample sizes.

Although considered computationally more expensive than some other statistical tests, permutation tests see wide use within bioinformatic workflows, including genome-wide association studies, GSEA, and other pathway analysis tools. Additionally, this methodology also sees use within machine learning and is applied to RandomForest models (Breiman, 2001). A similar approach is also seen with bootstrapping, a statistical resampling method which also sees use in bioinformatics (Khoshgoftaar et al., 2015).

Possibly the most widely adopted pathway analysis method which compares values calculated based on correct and incorrectly labeled data is GSEA (Subramanian et al., 2005). The key feature of GSEA is that it calculates enrichment scores by using the same dataset and comparing the observed data to statistics based on scrambled versions of itself (i.e. copies of itself where the phenotypic labels are permuted). This ensures that it is being compared to data with the same internal variance structure.

We seek to apply a similar concept to analyzing importances. The scrambling approach to neural network input has been proposed by Mandel and Barnett, where it was discussed in networks trained for multiple applications, including genetic analyses (2023). However, this was applied to explainable AI metrics on a one-layer neural network, and not a sparse multi-layer network as we are working with.

5.2 Materials and Methods

5.2.1 Permutation test

We propose comparing the importance of a term from a BINN that was classified with scrambled labels to the importance of that term when a BINN that was trained on correctly labeled data. The importances seen in the scrambled data should approximate the null distribution of importances specific to a given ontological term. We posit, if there is little difference between the importance of a term in classifying scrambled or unscrambled information, this should decrease confidence in the importance of this term. Whether or not these disparities exist is also of interest and could help the interpretation of BINNs.

5.2.2 Dataset

As before we use the human lung cancer data from the CCLE with 6, 12 and 24 samples from each lung cancer, as well as the full lung cancer dataset (49 and 68 samples). This used the modules generated from WGCNA from the previous implementation.

5.2.3 Permuting

The scrambling technique was implemented with a combination of Python, R, and Snakemake (Mölder et al., 2021). Snakemake is a pipelining technology common in bioinformatics, and was used to automate running BINNs on the scrambled data.

Scrambling was performed in R. To ensure the scrambled data would have the same number of each cancer as the real data, the *sample()* function in base R was used which allowed us to perform substitution with replacement (R Core Team, 2023). For each cluster, 250 scrambles would be generated. The maximum possible number of scrambles for a dataset with two equally sized treatment levels can be determined with the combinatorics “n choose r” calculation where n is the number of samples per group, and r is the total number of samples. Following this, the maximum number of scrambles available to a study which examined two phenotypes with 5 samples in each phenotype (totalling 10 samples), would be 251.

The Snakemake pipeline would run the Python scripts to train a BINN for each scramble provided. BINNs were implemented with the same hyperparameters as used previously. Snakemake allows for parallelization, significantly decreasing the walltime to run each analysis.

To accelerate the analysis, institutional high performance resources were used. We used the supercomputing resource Grex at the University of Manitoba. We used the resources and libraries available from Digital Research Alliance of Canada to create a Python virtual environment which allowed us to use Snakemake to run portions of our pipeline.

An additional step was taken to generate a metric from the scrambled BINNs. First, using the R library MASS c 7.3-60 (Ripley & Venables, 2009) for each term, a gamma distribution would be fitted to the importances observed on the scrambled BINNs. Before fitting, all importances from a BINN would be normalized by dividing by the maximum importance observed. Fitting a gamma distribution is possible due to the non-negativity of the importances generated by the BINN. This may not be the correct modeling of this distribution and will require further investigation for future works.

Next, a point-estimate is generated from this distribution. This point estimate for each term is defined as the area under the gamma distribution from the importance observed from the real BINN to positive infinity. We posit that this may be treated similar to a p-value, as very low point estimates will represent a small area under the curve, as will be the case if the

importance of the real BINN is much higher than those observed in the scrambled BINNs. A visualization of the point estimate can be seen in Figure 5.2

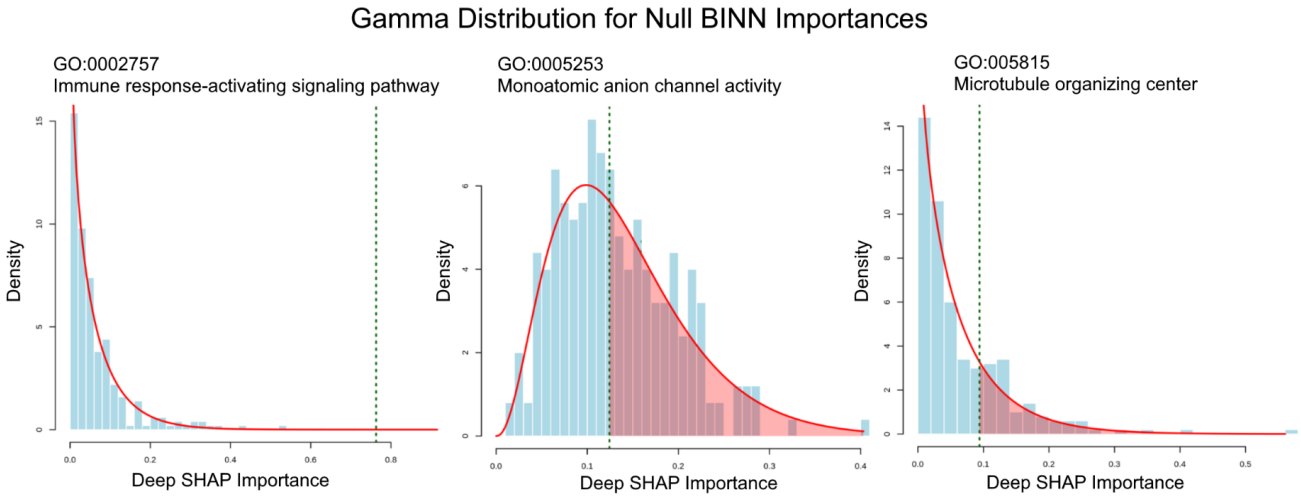

Figure 5.2 The gamma distribution and point estimate calculation for three ontological terms. In these examples, the histogram represents the distribution of the importances observed in the "null BINNs". The line in red is the gamma distribution applied to this histogram. The point estimate is calculated by taking the area under this curve from the observed importance in the "real BINN" (marked with a green dotted line) to positive infinity. This area can be seen coloured in red in the figure.

These point estimates were assessed to determine if they were able to predict statistical significance from our hypergeometric analysis using a ROC curve comparing point estimates to the importances themselves.

Once all point estimates are generated for a dataset, a cumulative dataset of terms is formed by amalgamating all terms from all clusters, with repeat terms being removed except for the row with the minimum point estimate.

5.2.4 Evaluating point estimates

To evaluate the relevance of the point-estimate metric to translational medicine, a similar analysis from before was performed. As was originally conducted in Chapter 3, a ROC curve for point estimates to predict statistically significant terms from the hypergeometric approach was generated, along with Spearman correlation. These were done on all individual clusters, as well as the cumulative dataset. For further validation, a confusion matrix was generated for each dataset on the cumulative terms. This was done to assess the agreement of point estimates and p-values. We examined the number of ontological terms with a point estimate <0.05 that also have a hypergeometric p-value of <0.05. We also consider the point estimates <0.0001, <0.001 and <0.01. We also examine the top 25 terms to show the range of point estimates and most relevant terms.

5.3 Results

Fitting of gamma curves to each term in the datasets was possible except in the case of a small percentage of terms for each dataset. These cases were caused by the function from the R library MASS being unable to fit due to the importance of 0 in all BINNs trained. Table 5.1 indicates the percentage of terms without a gamma distribution calculated.

Table 5.1 Summary of fitting gamma distributions to all terms with the MASS library, and how many terms were skipped due to values of 0.

Dataset	Number of Terms	Terms without a Gamma Distribution	Percentage
All samples	63,087	1,945	3.08%
24 samples	50,080	1,272	2.54%
12 samples	53,084	1,199	2.26%
6 samples	20,355	448	2.20%

Using a ROC curve to predict statistically significant terms from the hypergeometric test showed relatively similar results to using the importance values across all datasets. Figure 5.3, 5.4, 5.5 and 5.6 show the ROC curves for the clusters with the three highest and three lowest AUC scoring from each dataset.

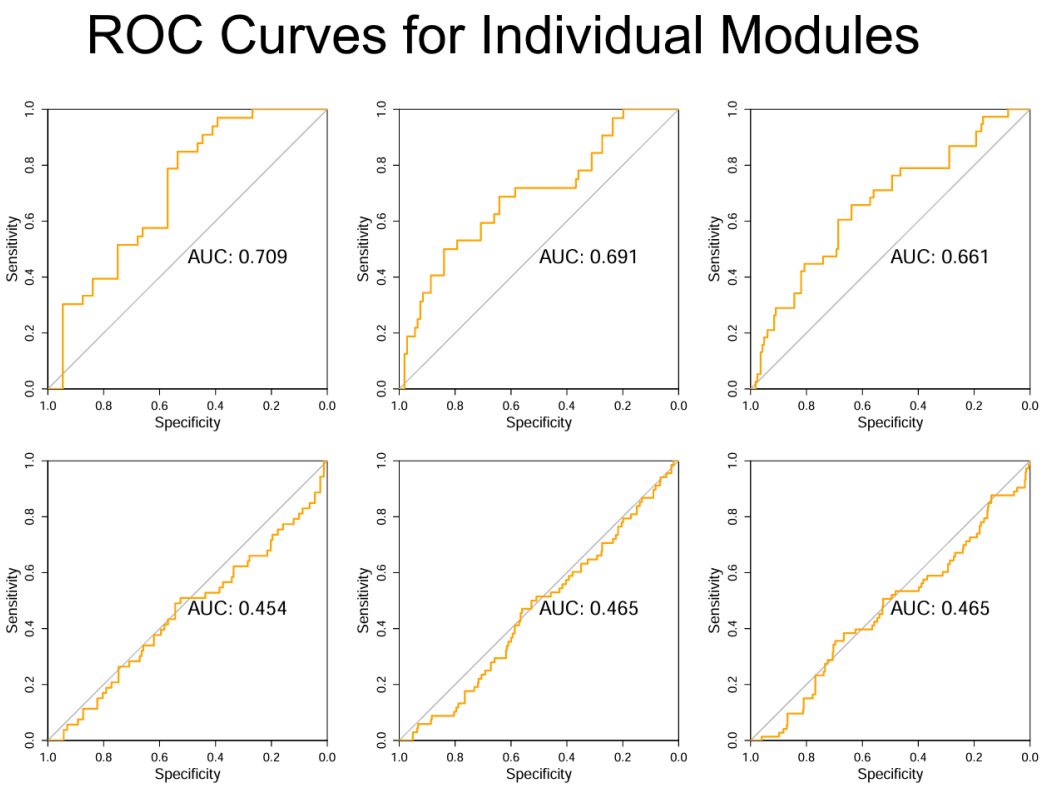

Figure 5.3 The ROC curves showing our point estimate's ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 6 samples from each lung cancer. The highest AUC observed for a module is 0.709, and the lowest is 0.454.

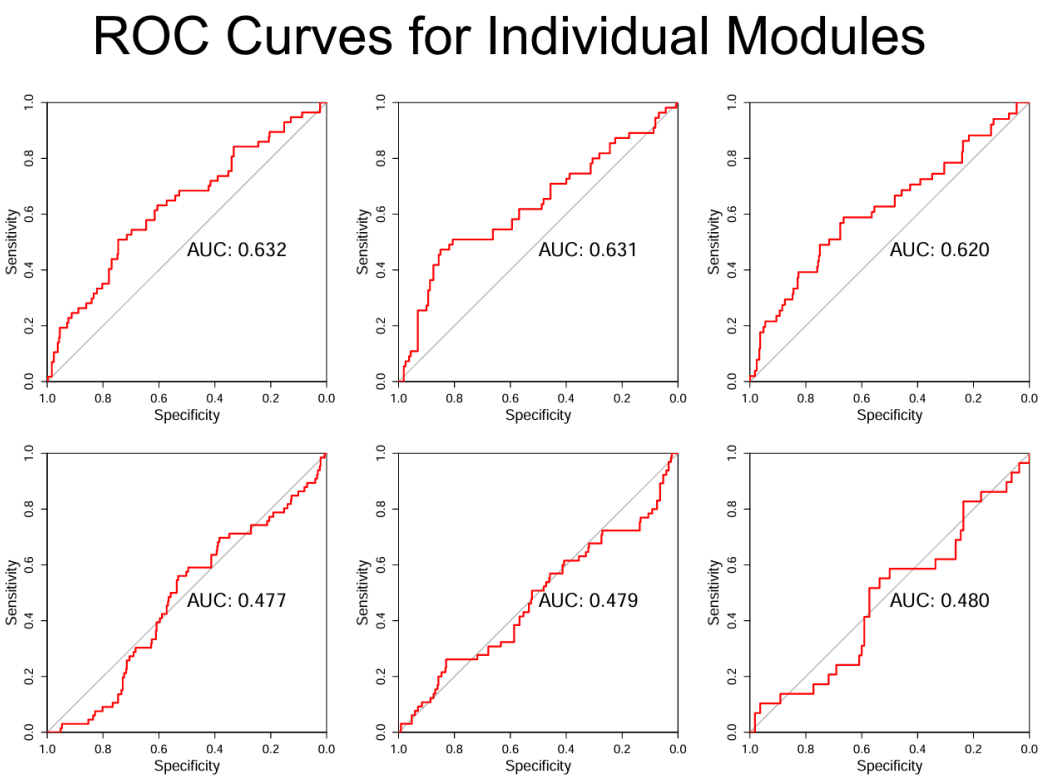

Figure 5.4 The ROC curves showing our point estimate's ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 24 samples from each lung cancer. The highest AUC observed for a module is 0.632, and the lowest is 0.477.

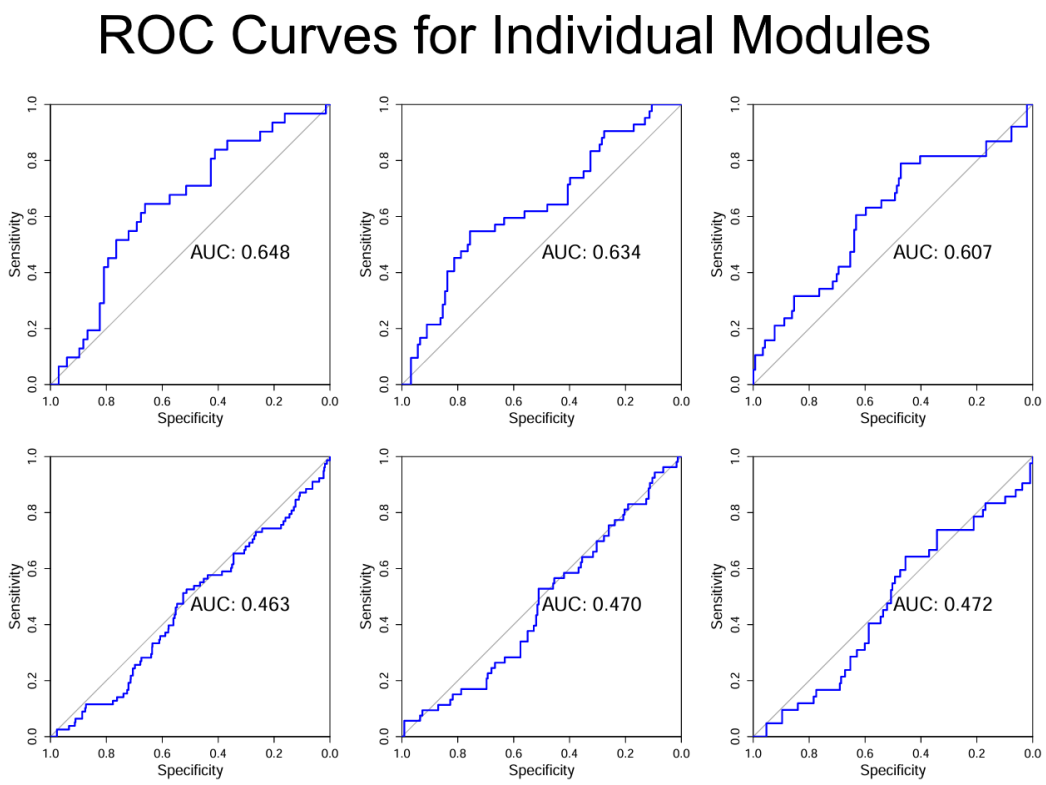

Figure 5.5 The ROC curves showing our point estimate's ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 12 samples from each lung cancer. The highest AUC observed for a module is 0.648, and the lowest is 0.463.

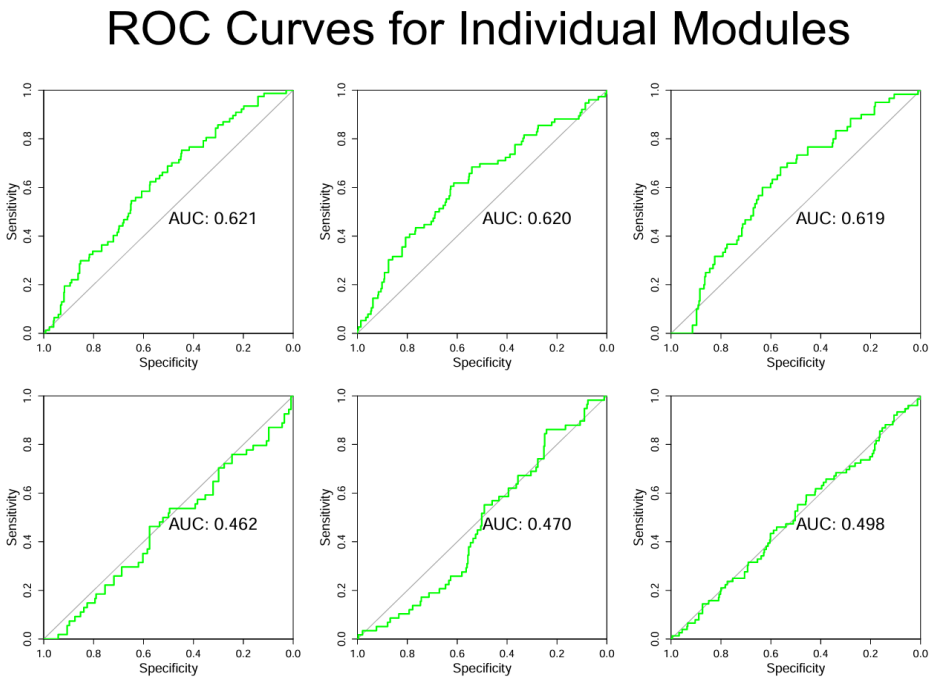

Figure 5.6 The ROC curves showing our point estimate's ability to predict statistically significant hypergeometric p-values for the highest and lowest AUC modules from the dataset formed from 6 samples from each lung cancer. The highest AUC observed for a module is 0.621, and the lowest is 0.462.

Cumulative clusters from the point estimates show a marked improvement from the cumulative cluster using the importances. The cumulative ROC curve from each dataset is shown in Figure 5.7.

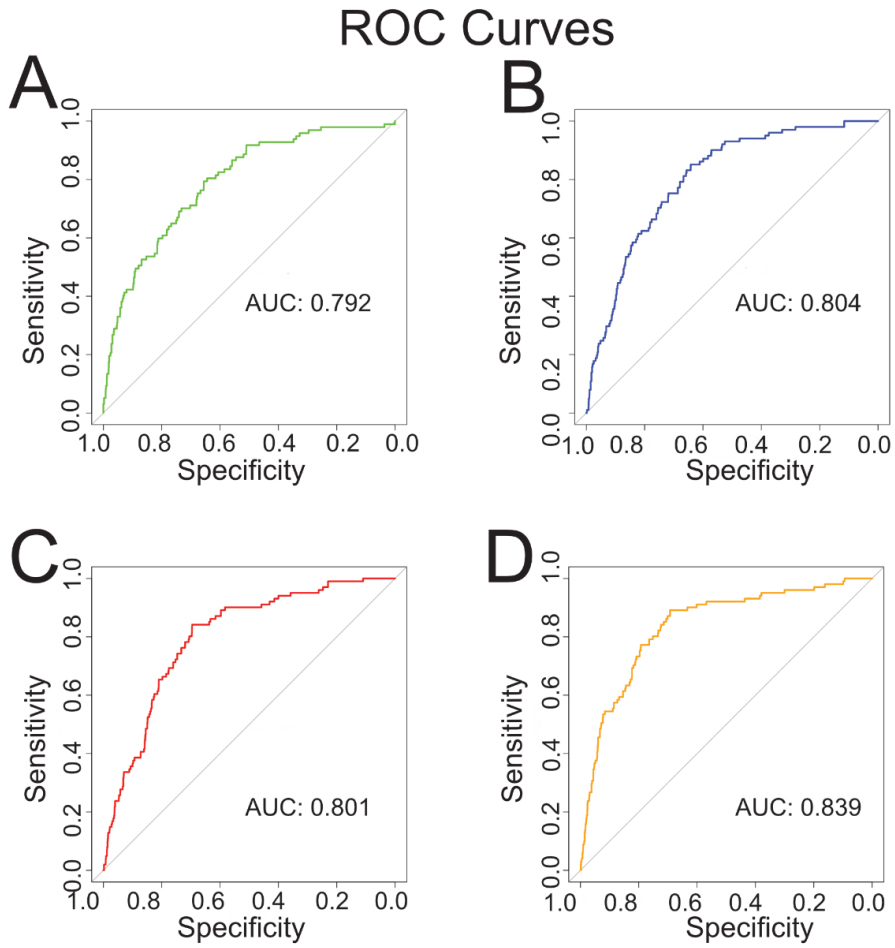

Figure 5.7 The ROC curves evaluating the point-estimates for the “Cumulative Modules” for each dataset. Panel A is for the 6 sample dataset, B is for the 12 sample dataset, C is for the 24 sample dataset, and D is for the dataset of all samples.

The similarity of the point estimate metric to the p-values was measured using the Spearman correlation and ROC as before. This was performed for all cumulative datasets. Results of the AUC and Spearman test can be seen in Table 5.2.

Table 5.2 A summary of the AUC value of the ROC curve when using the point estimate to classify statistically significant hypergeometric terms, along with the Spearman correlation of the point estimate and hypergeometric p-value. All datasets are formed from accumulating the terms from each module from WGCNA as described

Dataset	AUC of ROC	Spearman correlation	Spearman p-value	Number of terms
All samples	0.8251529	0.3895322	< 2.2e-16	13,967
12 samples	0.8113638	0.3453106	< 2.2e-16	14,546
6 samples	0.8081749	0.3408234	< 2.2e-16	6,027
24 samples	0.7937538	0.2883561	< 2.2e-16	13,224

To visualize the difference between the Spearman correlation calculated for the point estimate and the Spearman correlation calculated for the Deep SHAP importance previously, the rankings of each term by both metrics compared with the ranking of the p-value are plotted in Figure 5.8. The values came from the cumulative module from the dataset containing all samples.

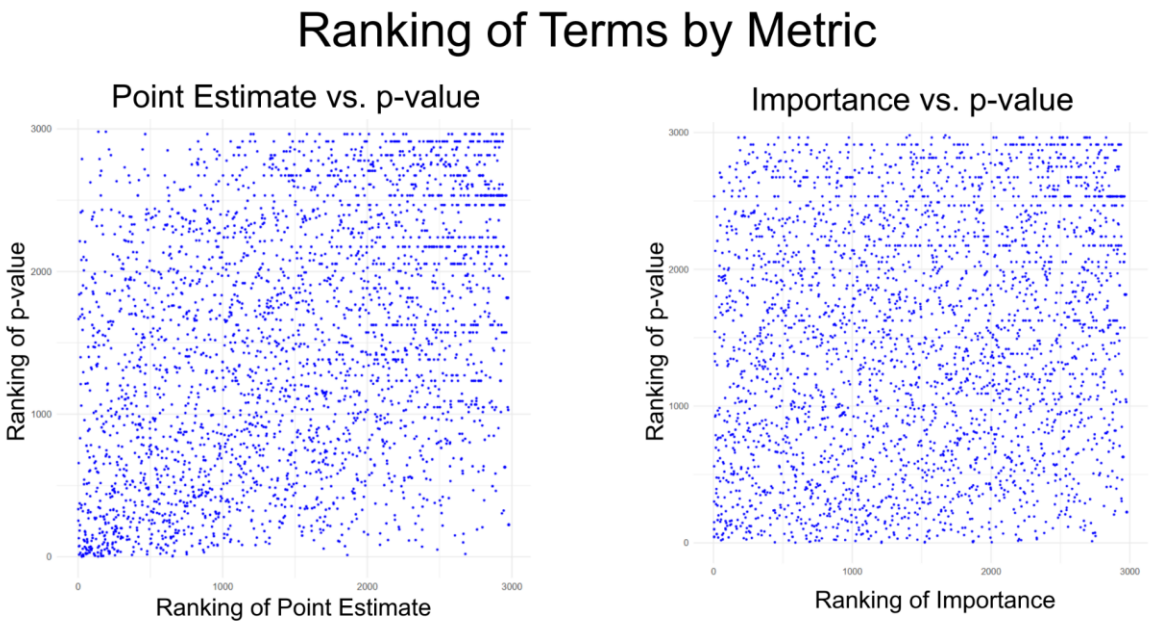

Figure 5.8 Each chart graphs the ranking of each term by the different metrics. On the y-axis is the ranking of the term by p-value. On the left, the x-axis shows the ranking of the term by point estimate, on the right the x-axis shows the ranking of the term by Deep SHAP importance.

Each individual module had an ROC curve and Spearman correlation calculated. The modules with the highest AUC are shown in Table 5.3.

Table 5.3 A summary of the metrics for individual clusters with the highest AUC from each dataset. AUC value of the ROC curve when using the point estimate to classify statistically significant hypergeometric terms, along with the Spearman correlation of the point estimate and hypergeometric p-value.

Module	Original Dataset	AUC of ROC	Spearman correlation	Spearman p-value	Number of terms
Module 156	All samples	0.7099567	0.3977902	0.0001133	119
Module 46	24 samples	0.6357045	0.06675606	0.145	592
Module 100	12 samples	0.6537002	-0.2145399	0.03297	127
Module 7	6 samples	0.6193477	-0.1678412	5.4e-05	718

Comparing the distribution of statistically significant hypergeometric values to the point estimates below the range of thresholds defined in methods can be seen in the confusion matrices in Figure 5.9.

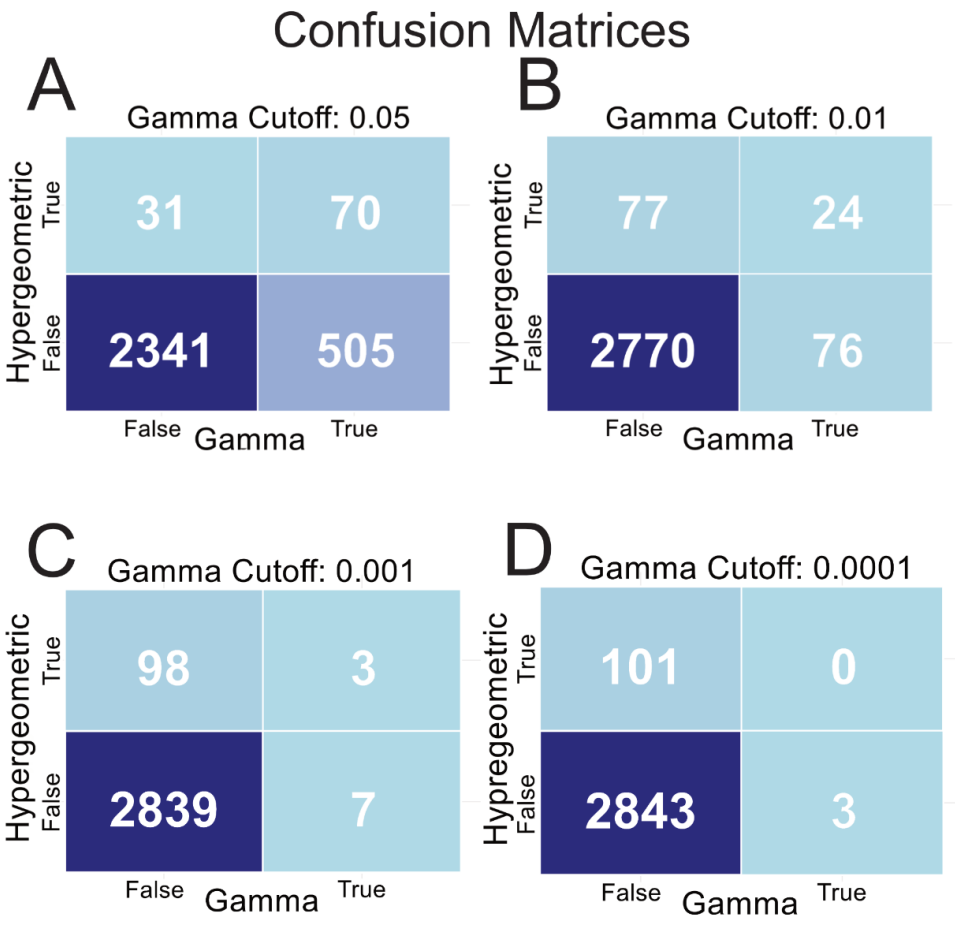

Figure 5.9 Confusion matrices showing the agreement between significant hypergeometric results and “significant” point estimates across different cutoffs of point estimates. Panel A uses a significance cutoff for point estimates of 0.05. Panel B shows a cutoff of 0.01, Panel C shows a cutoff of 0.001. Panel D shows a cutoff of 0.0001. The point estimates in this visualization are based off of the “Cumulative Module” from the dataset containing all samples of lung cancer data.

The 25 terms with the lowest point estimate from the “Cumulative Module” for the dataset with all samples can be seen in Table 5.4.

Table 5.4: The GO term ID, GO term name and point estimate for the 25 terms with the lowest point estimate

GO Term	GO term name	Gamma point estimate
GO:0002757	immune response-activating signaling pathway	1.90E-05
GO:0006939	smooth muscle contraction	3.56E-05
GO:0002218	activation of innate immune response	3.67E-05
GO:0006937	regulation of muscle contraction	0.000129907773
GO:0006275	regulation of DNA replication	0.0001822044596
GO:0006260	DNA replication	0.0001822045111
GO:0001067	transcription regulatory region nucleic acid binding	0.0001949934117
GO:0015804	neutral amino acid transport	0.0002167608631
GO:0000228	nuclear chromosome	0.0003464591895
GO:0016746	acyltransferase activity	0.0004985453116
GO:0016772	transferase activity, transferring phosphorus-containing groups	0.001064103773
GO:0007616	long-term memory	0.001150357011
GO:0016829	lyase activity	0.001231276158
GO:0016835	carbon-oxygen lyase activity	0.001231276543

GO:0006824	cobalt ion transport	0.001320908622
GO:0009893	positive regulation of metabolic process	0.001676528264
GO:0002097	tRNA wobble base modification	0.00174124445
GO:0015807	L-amino acid transport	0.001979813772
GO:0015825	L-serine transport	0.001979813855
GO:0009966	regulation of signal transduction	0.002074743635
GO:0005801	cis-Golgi network	0.002488917809
GO:0016817	hydrolase activity, acting on acid anhydrides	0.002524755905
GO:0016787	hydrolase activity	0.002524756368
GO:0003824	catalytic activity	0.002524756722
GO:0005737	cytoplasm	0.00252973961

5.4 Discussion

Examining ROC curves indicates that the point estimate metric is a better classifier of significant hypergeometric terms compared to Deep SHAP importances. However, our confusion matrices illustrate a problem in our experiment, that false positives always outweigh true positives, providing a poor likelihood ratio. We see a small effect of sample size on improving the AUC of the ROC curves, however, even at our lowest sample size, we achieve what is considered a strong AUC.

Additionally, the ranking of terms by metric shows that there is much more clustering around the low p-value terms and the low point estimate terms, when compared to the low p-value terms and the high Deep SHAP importance terms. Examining the top 25 terms, there are 2 GO terms which exist in the top 25 hypergeometric terms and the top 25 point estimate terms,

transcription regulatory region nucleic acid binding (GO:0001067), and cytoplasm (GO:0005737).

This shows a strong improvement from using only the importances, lending some credence to modeling a null distribution of importances. This may offer a computationally expensive way to provide greater signal from small sample sizes. Which may be an effective trade-off especially in the case of conditions where tissue samples are rare and precious, and researchers have access to high performance computational clusters.

Further investigation will have to be made into whether or not a gamma curve is an appropriate distribution to model the null importances. Also of interest is the effects of a larger number of permutations, with recommendations existing to use permutations in the thousands (Marozzi, 2004; GSEA and MSigDB Team, 2019).

Additionally, using the hypergeometric test as a point of truth to validate against is an imperfect compromise. Comparing point estimates using other techniques, such as comparing against other techniques, or a comparison using a curated (Alavi-Majd et al., 2014; Geistlinger et al., 2020) or simulated dataset (Maciejewski, 2013) may provide more conclusive evidence.

We were unable to find an immediate relationship between validation accuracy and the AUC. There is an unclear relationship with validation accuracy and biological relevance (as measured with a ROC curve). It stands to reason that a high sample size, low feature dataset may be classified accurately, however it is not clear if the accurate classification gives us a sense of how meaningful the explainable AI metrics are. Other factors may play a role and should be examined to help inform researchers how to best use BINNs to gain biological insight.

Metrics out of WGCNA may help add predictive power to our analysis. This tool examines the internal correlation structure of our data, and returns information relating to the coexpression of the specific data we are working with. Leveraging this information may help narrow our analysis in the future. For instance, a module with a high average coexpression may

be more appropriate to train a BINN with, and a module with low coexpression should potentially be skipped.

Chapter 6. Significance, limitations, and future work

Our analysis showed that using the Deep SHAP importance of an ontological term as implemented by Hartman et al. did not strongly reproduce a hypergeometric approach. Pre-clustering our dataset showed a small improvement in recapitulating the hypergeometric approach. Finally, using a point-estimate for each ontological term's Deep SHAP importance generated by a permutation test showed strong improvement of the ability to reproduce the hypergeometric approach. The ability to reproduce the hypergeometric approach appears stable at low sample sizes, which are common within translational research.

Due to the realities facing translational omic research, methods which allow researchers to glean actionable information for disease, within their limitations, may have great value. Application of machine learning algorithms offers a novel approach to perform analyses, and in the case of BINNs, these approaches are able to leverage the vast amounts of ontological information which has gone into developing traditional techniques. In our research, the permutation method explored showed improved ability to recapitulate a traditional approach using small sample sizes with the explainable AI techniques employed by BINNs. While we did not explore complex experimental design with more than two phenotypic conditions, a permutation test would still be possible to apply to more complex designs. This may form the basis for further exploration into BINNs for complex designs.

Pathway analysis is implemented in various tools, and an exploration of how this method compares to other approaches may offer more insight, considering the number of false positives still present in our technique. We did not examine how using a hypergeometric overrepresentation analysis may have impacted our results, and using a different "point of truth" may change our conclusions. Additionally, using curated datasets, where the important pathways should be known *a priori*, may allow us to evaluate this method without a comparison to other techniques, avoiding a potential circular validation.

There is also an evident high dimensionality of characteristics which impact this approach. The clustering method, hyperparameters of BINN training, and the calculation of our point estimate, all offer avenues for further exploration. These were not widely explored in our analysis, and these may offer a way to have greater confidence in the explainable AI metrics of BINNs. Such information may help inform best practices for applying BINNs, and help researchers determine the relevance of output.

Additionally, the other features of BINNs: integrating multi-omic data sources, and classifying multiple treatment levels, were not explored. These uses may show even greater utility for BINNs.

References

Alavi-Majd, H., Khodakarim, S., Zayeri, F., Rezaei-Tavirani, M., Tabatabaei, S. M., & Heydarpour-Meymeh, M. (2014). Assessment of gene set analysis methods based on microarray data. *Gene*, *534*(2), 383–389. https://doi.org/10.1016/j.gene.2013.08.063

Alhamdoosh, M., Ng, M., Wilson, N. J., Sheridan, J. M., Huynh, H., Wilson, M. J., & Ritchie, M. E. (2016). Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, *33*(3), 414–424. https://doi.org/10.1093/bioinformatics/btw623

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*, 167–182. https://doi.org/10.1016/j.jocm.2018.07.002

Anonymous. (2019, November). GSEA user guide. https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

Anonymous. *About the GO*. Gene Ontology Resource. (2025, November 5). https://geneontology.org/docs/introduction-to-go

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. https://doi.org/10.1038/nrg1272

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of Anticancer Drug Sensitivity. *Nature*, *483*(7391), 603–607. https://doi.org/10.1038/nature11003

Barrett, J. S., Betourne, A., Walls, R. L., Lasater, K., Russell, S., Borens, A., Rohatagi, S., & Roddy, W. (2023). The future of Rare Disease Drug Development: The rare disease cures accelerator data analytics platform (RDCA-DAP). *Journal of Pharmacokinetics and Pharmacodynamics*, *50*(6), 507–519. https://doi.org/10.1007/s10928-023-09859-7

Bernabé, C. H., Queralt-Rosinach, N., Silva Souza, V. E., Bonino da Silva Santos, L. O., Mons, B., Jacobsen, A., & Roos, M. (2023). The use of foundational ontologies in biomedical research. *Journal of Biomedical Semantics*, *14*(1). https://doi.org/10.1186/s13326-023-00300-z

BioRender. (n.d.). *Scientific Image and Illustration Software*. https://www.biorender.com/

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In Artificial intelligence safety and security (pp. 57–69). Chapman and Hall/CRC.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096

Cao, Y., Yin, C., Zhou, X., & Zhao, Y. (2025). Fr-Binn: Biologically informed neural networks for enhanced biomarker discovery and pathway analysis. *International Journal of Molecular Sciences*, *26*(14), 6670. https://doi.org/10.3390/ijms26146670

Carlson, M. (2023). GO.db: A set of annotation maps describing the entire Gene OntologyVersion (R package version 3.18.0.).

Chen, C., Wang, J., Pan, D., Wang, X., Xu, Y., Yan, J., Wang, L., Yang, X., Yang, M., & Liu, G. (2023). Applications of multi-omics analysis in human diseases. *MedComm*, *4*(4). https://doi.org/10.1002/mco2.315

Chen, V., Yang, M., Cui, W., Kim, J. S., Talwalkar, A., & Ma, J. (2024). Applying interpretable machine learning in Computational Biology—Pitfalls, recommendations and opportunities for new developments. *Nature Methods*, *21*(8), 1454–1461. https://doi.org/10.1038/s41592-024-02359-7

Chua, A. E., Pfeifer, L. D., Sekera, E. R., Hummon, A. B., & Desaire, H. (2023). Workflow for evaluating normalization tools for OMICS data using supervised and unsupervised machine learning. Journal of the American Society for Mass Spectrometry, 34(12), 2775–2784. https://doi.org/10.1021/jasms.3c00295

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). David: Database for annotation, visualization, and Integrated Discovery. *Genome Biology*, *4*(9). https://doi.org/10.1186/gb-2003-4-9-r60

Dhamdhere, K., Sundararajan, M., & Yan, Q. (2018). How important is a neuron?. arXiv preprint arXiv:1805.12233.

Dimopoulos, Y., Bourret, P., & Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters, 2(6), 1–4.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomart. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

EBI Functional Genomics Team. (n.d.). *Expression atlas*. EBI. https://www.ebi.ac.uk/gxa/home

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., Arnoff, T. E., Park, J., Hahn, W. C., & Van Allen, E. M. (2021). Biologically informed deep neural network for Prostate Cancer Discovery. *Nature*, *598*(7880), 348–352. https://doi.org/10.1038/s41586-021-03922-4

Fleuret, F. (2024). The little book of deep learning. University of Geneva.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., & Shlomi, T. (2011). Predicting selective drug targets in cancer through Metabolic Networks. *Molecular Systems Biology*, *7*(1). https://doi.org/10.1038/msb.2011.35

Gakii, C., Mukami, V., & Too, B. (2023). Feature selection for classification using WGCNA and spread sub-sample for an imbalanced rheumatoid arthritis RNASEQ Data. *Informatics in Medicine Unlocked*, *43*, 101402. https://doi.org/10.1016/j.imu.2023.101402

Gan, W. (2025). Impact of sample size and its estimation in medical research. *AJPM Focus*, 100451. https://doi.org/10.1016/j.focus.2025.100451

García-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway analysis: State of the art. *Frontiers in Physiology*, *6*. https://doi.org/10.3389/fphys.2015.00383

Garson, D. G. (1991). Interpreting neural network connection weights.

Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., Zimmer, R., & Waldron, L. (2020). Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*, *22*(1), 545–556. https://doi.org/10.1093/bib/bbz158

Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., & Malinverni, R. (2015). Regioner: An R/bioconductor package for the Association analysis of genomic regions based on permutation tests. *Bioinformatics*, *32*(2), 289–291. https://doi.org/10.1093/bioinformatics/btv562

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, *569*(7757), 503–508. https://doi.org/10.1038/s41586-019-1186-3

Grote, S. (2025). GOfuncR: Gene ontology enrichment using FUNCVersion (R package version 1.28.0). Retrieved from https://bioconductor.org/packages/GOfuncR.

Gruber, T. R. (1995). Toward principles for the design of Ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5–6), 907–928. https://doi.org/10.1006/ijhc.1995.1081

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hao, J., Kim, Y., Kim, T.-K., & Kang, M. (2018). PASNet: Pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2500-z

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

Hartman, E. (2025, February). *Binn/docs/robustness.ipynb at main · infectionmedicineproteomics/binn*. GitHub. https://github.com/InfectionMedicineProteomics/BINN/blob/main/docs/robustness.ipynb

Hartman, E., Scott, A. M., Karlsson, C., Mohanty, T., Vaara, S. T., Linder, A., Malmström, L., & Malmström, J. (2023). Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-41146-4

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., & O'Donovan, C. (2014). The Goa Database: Gene Ontology Annotation Updates for 2015. Nucleic Acids Research, 43(D1). https://doi.org/10.1093/nar/gku1113

ISO/IEC 21838-1:2021. ISO. (2021, August 6). https://www.iso.org/standard/71954.html

Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D. M., Duncan, W. D., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., … Peters, B. (2021). Obo Foundry in 2021: Operationalizing Open data principles to evaluate ontologies. *Database*, *2021*. https://doi.org/10.1093/database/baab069

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2020). Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, *22*(1), 393–415. https://doi.org/10.1093/bib/bbz170

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2). https://doi.org/10.1371/journal.pcbi.1002375

Khoshgoftaar, T. M., Fazelpour, A., Dittman, D. J., & Napolitano, A. (2015). Alterations to the bootstrapping process within Random Forest: A case study on imbalanced bioinformatics data. *2015 IEEE International Conference on Information Reuse and Integration*, 342–348. https://doi.org/10.1109/iri.2015.59

Kirpich, A., Ainsworth, E. A., Wedow, J. M., Newman, J. R., Michailidis, G., & McIntyre, L. M. (2018). Variable selection in OMICS DATA: A practical evaluation of small sample sizes. *PLOS ONE*, *13*(6). https://doi.org/10.1371/journal.pone.0197910

Kulski, J. K. (2016). Next-generation sequencing — an overview of the history, tools, and “Omic” applications. *Next Generation Sequencing - Advances, Applications and Challenges*. https://doi.org/10.5772/61964

Lambrix, P. (2014). Semantic web, ontologies, and Linked Data. *Comprehensive Biomedical Physics*, 67–76. https://doi.org/10.1016/b978-0-444-53632-7.01127-8

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1). https://doi.org/10.1186/1471-2105-9-559

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Maciejewski, H. (2013). Gene set analysis methods: Statistical models and methodological differences. *Briefings in Bioinformatics*, *15*(4), 504–518. https://doi.org/10.1093/bib/bbt002

Mandel, F., & Barnett, I. (2024). Permutation-based hypothesis testing for Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(13), 14306–14314. https://doi.org/10.1609/aaai.v38i13.29343

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for Health Care: A Comprehensive Survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Marozzi, M. (2004). Some remarks about the number of permutations one should consider to perform a permutation test. Statistica, 64(1), 193-201.

Mazandu, G. K., & Mulder, N. J. (2012). A topology-based metric for measuring term similarity in the gene ontology. *Advances in Bioinformatics*, *2012*, 1–17. https://doi.org/10.1155/2012/975783

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable Data Analysis with snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.1

Nayak, B., & Hazra, A. (2011). How to choose the right statistical test? *Indian Journal of Ophthalmology*, *59*(2), 85. https://doi.org/10.4103/0301-4738.77005

Nguyen, T.-M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: A Comprehensive Review and assessment. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1790-4

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2022). Obtaining genetics insights from deep learning via Explainable Artificial Intelligence. *Nature Reviews Genetics*, *24*(2), 125–137. https://doi.org/10.1038/s41576-022-00532-2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

Pavlidis, P., Qin, J., Arango, V., Mann, J. J., & Sibille, E. (2004). Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*, *29*(6), 1213–1222. https://doi.org/10.1023/b:nere.0000023608.29741.45

R Core Team. (2023). R: A Language and Environment for Statistical Computing. computer software, R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Raina, P., Guinea, R., Chatsirisupachai, K., Lopes, I., Farooq, Z., Guinea, C., Solyom, C.-A., & de Magalhães, J. P. (2022). GeneFriends: Gene co-expression databases and tools for humans and model organisms. *Nucleic Acids Research*, *51*(D1). https://doi.org/10.1093/nar/gkac1031

Ripley, B., & Venables, B. (2009). Mass: Support functions and datasets for Venables and Ripley's mass. *CRAN: Contributed Packages*. https://doi.org/10.32614/cran.package.mass

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). <tt>edger</tt>: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rogers, K. (n.d.). *OMICS*. Encyclopædia Britannica. https://www.britannica.com/science/omics

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A Review. *Computational and Structural Biotechnology Journal*, *23*, 1154–1168. https://doi.org/10.1016/j.csbj.2024.02.018

Ruck, D. W., Rogers, S. K., & Kabrisky, M. (1990). Feature selection using a multilayer perceptron. Journal of neural network computing, 2(2), 40–48.

Schultz, C. M., DiGeronimo, R. J., & Yoder, B. A. (2007). Congenital diaphragmatic hernia: A simplified postnatal predictor of outcome. Journal of Pediatric Surgery, 42(3), 510–516. https://doi.org/10.1016/j.jpedsurg.2006.10.043

Selby, D. A., Sprang, M., Ewald, J., & Vollmer, S. J. (2025). Beyond the black box with biologically informed Neural Networks. *Nature Reviews Genetics*, *26*(6), 371–372. https://doi.org/10.1038/s41576-025-00826-1

Silva, M. C., Eugénio, P., Faria, D., & Pesquita, C. (2022). Ontologies and knowledge graphs in oncology research. *Cancers*, *14*(8), 1906. https://doi.org/10.3390/cancers14081906

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The Obo Foundry: COORDINATED EVOLUTION OF ONTOLOGIES to support Biomedical Data Integration. *Nature Biotechnology*, *25*(11), 1251–1255. https://doi.org/10.1038/nbt1346

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. International conference on machine learning,

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J., Kim, C. J., Kusanovic, J. P., & Romero, R. (2008). A novel signaling pathway impact analysis. *Bioinformatics*, *25*(1), 75–82. https://doi.org/10.1093/bioinformatics/btn577

Tõnisson, L., & Preden, J. (2024). Ontology-based data representation prototype for indoor air quality, building energy performance, and Health Data Computation. *Sustainability*, *16*(13), 5677. https://doi.org/10.3390/su16135677

Vahabi, N., & Michailidis, G. (2022). Unsupervised multi-omics data integration methods: A comprehensive review. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.854752

Vallejo-Cremades, M., Merino, J., Carmona, R., Córdoba, L., Salvador, B., Martínez, L., Tovar, J. A., Llamas, M. Á., Muñoz-Chápuli, R., & Fresno, M. (2024). Toll-like receptors ligand immunomodulators for the treatment congenital diaphragmatic hernia. Orphanet Journal of Rare Diseases, 19(1). https://doi.org/10.1186/s13023-024-03384-7

van Hilten, A., van Rooij, J., Heijmans, B. T., ’t Hoen, P. A., Meurs, J. van, Jansen, R., Franke, L., Boomsma, D. I., Pool, R., van Dongen, J., Hottenga, J. J., van Greevenbroek, M. M., Stehouwer, C. D., van der Kallen, C. J., Schalkwijk, C. G., Wijmenga, C., Zhernakova, S., Tigchelaar, E. F., Slagboom, P. E., … Roshchupkin, G. V. (2024). Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data. *Npj Systems Biology and Applications*, *10*(1). https://doi.org/10.1038/s41540-024-00405-w

Vitorino, R. (2024). Transforming clinical research: The power of high-throughput omics integration. *Proteomes*, *12*(3), 25. https://doi.org/10.3390/proteomes12030025

W3C OWL Working Group (Ed.). (2012, December 11). *OWL 2 Web Ontology Language Document Overview (second edition)*. W3C. https://www.w3.org/TR/2012/REC-owl2-overview-20121211/#Documentation_Roadmap

Wehling, M. (2021). Introduction and definitions. *Principles of Translational Science in Medicine*, 3–7. https://doi.org/10.1016/b978-0-12-820493-1.00020-9

Wijesooriya, K., Jadaan, S. A., Perera, K. L., Kaur, T., & Ziemann, M. (2022). Urgent need for consistent standards in functional enrichment analysis. *PLOS Computational Biology*, *18*(3). https://doi.org/10.1371/journal.pcbi.1009935

Worden, K., Tsialiamanis, G., Cross, E. J., & Rogers, T. J. (2023). Artificial Neural Networks. *Computational Methods in Engineering & the Sciences*, 85–119. https://doi.org/10.1007/978-3-031-36644-4_2

World Bank Group. (n.d.). *Mortality rate, infant (per 1,000 live births)*. World Bank Open Data. https://data.worldbank.org/indicator/SP.DYN.IMRT.IN

World Health Organization. (n.d.). *Maternal health*. https://www.who.int/health-topics/maternal-health

Xie, C., Jauhari, S., & Mora, A. (2021). Popularity and performance of bioinformatics software: The case of gene set analysis. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04124-5

Yang, R., Rodriguez-Fernandez, M., St. John, P. C., & Doyle, F. J. (2014). Systems biology. *Modelling Methodology for Physiology and Medicine*, 159–187. https://doi.org/10.1016/b978-0-12-411557-6.00008-2

Yinyin Yuan, & Chang-Tsun Li. (2008). Probabilistic framework for gene expression clustering validation based on gene ontology and graph theory. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 625–628. https://doi.org/10.1109/icassp.2008.4517687

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). Clusterprofiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zeng, F., Shi, M., Xiao, H., & Chi, X. (2021). WGCNA-based identification of hub genes and key pathways involved in nonalcoholic fatty liver disease. *BioMed Research International*, *2021*(1). https://doi.org/10.1155/2021/5633211

Zompola, A., Korfiati, A., Theofilatos, K., & Mavroudi, S. (2023). OMICS-CNN: A comprehensive pipeline for predictive analytics in quantitative omics using one-dimensional convolutional neural networks. *Heliyon*, *9*(11). https://doi.org/10.1016/j.heliyon.2023.e21165