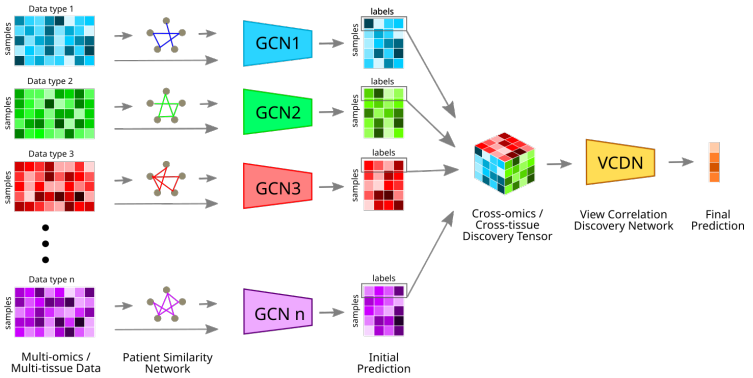
DEPARTMENT OF BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY
MADRAS
CHENNAI - 600036

# Cross-omic Deep Learning Networks for Identifying Disease Biomarkers




*A Project Report*

*Submitted by*

**ADITYA JEEVANNAVAR**

*In the partial fulfilment of requirements*

*For the award of the*

**BACHELOR OF SCIENCE AND MASTER OF SCIENCE**

July 2021

# CERTIFICATE

This is to undertake that the Project Report titled **CROSS-OMIC DEEP LEARNING NETWORKS FOR IDENTIFYING DISEASE BIOMARKERS**, submitted by me to the Indian Institute of Technology Madras, for the award of Bachelor of Science and Master of Science (Dual-Degree), is a bona fide record of the research work done by me under the supervision of Dr. Manikandan Narayanan. The contents of this Project Report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

**Place: Chennai 600 036**

**Date: 8th June 2021**

**Aditya Jeevannavar**
Dual Degree Student

**Dr. Manikandan Narayanan**
Project Guide

# ACKNOWLEDGEMENTS

I thank Dr. Manikandan Narayanan for his guidance and patience, and for weekly discussions that compelled me to work regularly and always have the project in my thoughts. I also thank him for making available to me the resources, computational *and* analytical, that made this project possible.

I also thank BIRDS (Bioinformatics and Integrative Data Science) group for listening to and critiquing my project presentations. Critique of my work as well as presentations of others in the group have provided me many ideas that enabled the progress of this project.

The results presented here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga, and METABRIC: https://ega-archive.org/studies/EGAS00000000083.

# ABSTRACT

KEYWORDS: Multi-omics; Data integration; Breast cancer; Deep Learning

With the current high throughput nature of the omics technologies, researchers are able to collect several omics data set on the same samples to create multi-omics data sets. We have created a Graph Convolutional Network and View Correlation Discovery Network based framework to integrate such multi-omics as well as multi-tissue data sets. To infer inter-omics relations, feature engineering is used to generate pairwise interaction features across omics. The GCN-VCDN model utilizes multi-omics data and inter-sample relations to classify samples. Model-agnostic interpretablity measures, SHAP and LIME, provide feature importance scores to enable personalized medicine and predict biomarkers. We also created an omics imputation framework in order to incorporate data sets that have missing omics. Additionally, the model can also integrate multi-tissue data to infer inter-tissue relations and perform sample classification.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# GLOSSARY

The following are some of the commonly used terms in this thesis:

| | |
|---|---|
| **Beta value** | A value that represents the ratio between the methylated array intensity and total array intensity. It falls between 0 (lower levels of methylation) and 1 (higher levels of methylation). |
| **Biomarker** | An objectively measurable feature, either molecular like gene expression level or clinical like body temperature, that is indicative of disease process. |
| **F1 Score** | The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst value at 0 |
| **Modality** | One view/modality of a multi-view/multi-modality data set, for example gene expression in a multi-omics data set |
| **Multi-omics data** | Matched data set with multiple -omes like genome and proteome such that data across multiple -omes exists for each sample |
| **Multi-modality integration** | Inference of intra-modality and inter-modality relations for downstream tasks like sample classification |
| **Multi-tissue data** | Matched gene expression data set for multiple tissues from the same patient |

# ABBREVIATIONS

| | |
|---|---|
| BRCA | Breast Invasive Carcinoma |
| GCN | Graph Convolutional Network |
| GEO | Gene Expression Omnibus |
| GSEA | Gene Set Enrichment Analysis |
| KNN | K Nearest Neighbours |
| LIME | Local Interpretable Model-agnistic Explanations |
| METABRIC | Molecular Taxonomy of Breast Cancer International Consortium |
| MSigDB | Molecular Signatures Database |
| RSEM | RNA-Seq by Expectation Maximization |
| SHAP | SHapley Additive exPlanations |
| SMOTE | Synthetic Minority Over-sampling TEchnique |
| TCGA | The Cancer Genome Atlas |
| VCDN | View Correlation Discovery Network |

# CHAPTER 1

# INTRODUCTION

## 1.1 What are multi-omics and multi-tissue data?

The -ome suffix in cellular and molecular biology forms nouns with the sense of "all constituents considered collectively". Genome, transcriptome, and proteome respectively consider all the genes, gene transcripts, and proteins of an organism collectively. Omics is the study of these collectives, as in, genomics is the study of the genome, transcriptomics the study of the transcriptome, and so on.

With the current high throughput nature of the "omics" technologies, researchers are able to collect several omics data sets on the same experimental samples. This has led to the advent "multi-omics". The heterogeneity of multi-omics can be seen in the non-existence of a simple one-to-one relation between the features of all the different omics data sets. While the base of the information is the genome, multi-omics is still a study of heterogeneous data sets. In other words, while the DNA holds the information for all molecular and cellular processes, there exist numerous regulatory mechanisms that introduce variables that cannot simply deciphered by the genome alone.

The analysis of multi-tissue data aims to identify and elucidate the cross-talk between the different tissues just as the analysis of multi-omics data aims to identify and elucidate the cross-talk between the different omics layers. For example, Seldin and Lusis (2019) use weighted gene correlated network analysis (WGCNA) to look at pathway-based interactions between liver and adipose tissue. Their framework, QENIE, ranks tissue-tissue interactions by global patterns of correlation.

The analysis of multi-omics and multi-tissue data are abundantly similar in that they both analyse the data across different phenotypes such as disease states or categories, for example Alzheimer's affected or normal control. The only difference is that multi-omics analysis involves inferring relations among different omics data on the same tissue and multi-tissue analysis involves inferring relations among different tissue data

on the same omics.

## 1.2 Why do we need to integrate multi-modality data?

For over a decade, genomics has been used for finding biomarkers to help in diagnosis and prognosis of disease or disorder and for finding causative variants and pathways to help in the cure. Now, with the availability of multi-omics data sets, there is a better opportunity for performing these functions. The information in the other omics can be used to fill the gaps that remain when only genomics is used. Also, the use of multi-omics can reduce the noise encountered in the analysis of single omics data. Thus, there is the need for a reliable multi-omics integration method that combines information across different omics types to predict better biomarkers and causative pathways. It has largely been accepted that such an integrative analysis is necessary for a comprehensive understanding of a biological system. (Gligorijević and Przulj, 2015)

For a detailed literature review of multi-modality integration tools, frameworks, and reviews, refer my mid-year project report. (Jeevannavar, 2020)

## 1.3 What about studies that collect only a single omics?

Collecting multi-omics data is expensive. There are many cohorts and research groups across the world that collect biological data but not all of them are equally well funded. Some groups are able to collect 4-5 different types of omics data on the same samples, while some are able to collect only 2-3. While the lesser funded studies comparatively generate a limited set of data, this data is still valuable and must not go to waste. Even with the missing omics data, it should be integrate-able in the same multi-omics integration models. This can be done by imputing the missing omics data based on the features in the other collected omics types.

Another case for the imputation of missing omics types and inclusion of such data sets is that integration models/methods are unable learn all the features from a limited set of samples, and therefore using the imputed data can increase the performance of the model. While the imputed data is not novel data that has been collected directly

from the samples, it can nevertheless provide valuable information to the model. We have thus implemented an imputation framework that can be used before the data enters the integration pipeline.

Most studies on data imputation, whether omics/multi-omics imputation or other general data imputation, deal with missing values that are missing at random or missing completely at random. We, in this study, are looking to deal with data that is missing not at random in the specific case that we know why the data is missing. Here, we have studies where one or more omics types were not measured due to constraints like budget and thus the data is missing certain values. We use the relations between different omics and make a predictive model that can impute values across studies.

## 1.4 What is this study about?

We have created a Graph Convolutional Network and View Correlation Discovery Network based framework to integrate multi-omics as well as multi-tissue data. To infer inter-modality relations, feature engineering is used to generate pairwise interaction features across modalities. The GCN-VCDN model utilizes multi-modality data and inter-sample relations to classify samples. Model-agnostic interpretablity measures, SHAP and LIME, provide feature importance scores to enable personalized medicine and predict biomarkers.

Many existing studies have performed multi-modal integration before (as described in my mid-year project report Jeevannavar (2020)), but they have mostly focused on overlaying different omics onto a single layer like the genome or the proteome. This project, on the other hand, treats each omics as equally dependent on the others and instead focuses on the samples' similarity with one another for classification. This enables us to extend the work to multi-tissue data integration as well. With omics imputation also in the pipeline, missing omics or tissue data can be dealt with easily.

Towards this, we provide a GCN + VCDN model that achieves an average accuracy of 84% at cancer subtype classification of the TCGA BRCA data set. Inclusion of the novel pairwise interaction features provides an improvement of 4% on the accuracy. The model performs subtype classification with 79% accuracy with only mRNA ex-

pression data, while it improves to 82% accuracy with gene expression *and* other omics imputed from the gene expression. The model also classifies an independent validation set (METABRIC) with 74% accuracy.

# CHAPTER 2

# DATA

In this chapter, we have two descriptions of the multiple data sets used. The first two, TCGA (Network, 2012; Weinstein *et al.*, 2013) and METABRIC (Curtis *et al.*, 2012), are breast cancer associated multi-omics data sets. The third, HBTRC, is a brain disorder associated multi-tissue data set.

## 2.1 The Cancer Genome Atlas (TCGA)

### 2.1.1 Data Availability

The TCGA Breast Invasive Carcinoma (BRCA) multi-omics data set used in this project is freely accessible through the Broad GDAC (Genome Data Analysis Center) Firehose (`https://gdac.broadinstitute.org/`) or alternatively at Firebrowse (`http://firebrowse.org`). Level 3 data was used in this project. Level 1 refers to raw and controlled data, level 2 to processed and controlled data, level 3 to segmented or interpreted and open access data, and level 4 to high level genomic and open access data (Silva *et al.*, 2016). While the access to level 1 and level 2 data is controlled, level 3 and level 4 data is freely accessible as it is not individually identifiable. The data was downloaded using the firehose_get tool.

### 2.1.2 Data Preliminaries

Three omics, gene expression, DNA methylation, and microRNA expression, from the multi-omics data set were used. The number of features in the three omics is presented in table 2.1 below. There are 622 samples that have matched data across the three omics types. The samples are classified into five subtypes: LumA, LumB, Basal, Her2, and Normal-like. The distribution of samples across cancer subtypes is presented in table 2.2 below.

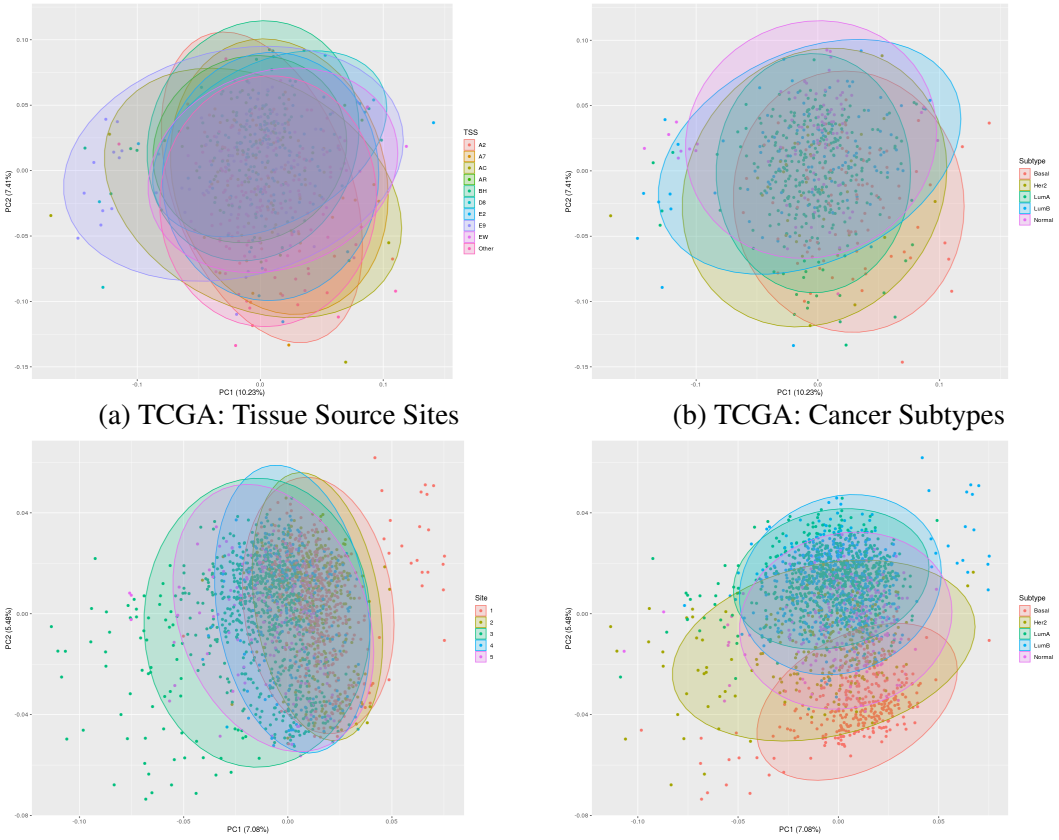
| mRNA | DNA methylation | miRNA |
|---|---|---|
| 18321 | 24039 | 387 |

Table 2.1: Number of features in the TCGA BRCA data set

| LumA | LumB | Basal | Her2 | Normal | **Total** |
|---|---|---|---|---|---|
| 298 | 107 | 95 | 36 | 86 | 622 |

Table 2.2: Cancer subtypes in the TCGA BRCA data set

The level 4 data used in this study has been processed and corrected for batch effects. Demographic and clinical variables like age, race, ethnicity, histological type, pathologic stage, therapy administered and more are available.




(a) TCGA: Tissue Source Sites

(b) TCGA: Cancer Subtypes

(c) METABRIC: Tissue Source Sites

(d) METABRIC: Cancer Subtypes

Fig. 2.1: **Data**: TCGA BRCA (a & b) and METABRIC (c & d) samples plotted based on principal components of scaled data and coloured based on potential covariates. There are no well-defined clusters.

Principal Component Analysis is performed to assess whether the data is clustering. The PCA clustering is visualized with different colour overlays to assess whether the data clusters based on technical or biological characteristics. As observable in figure

2.1a, samples do not cluster based on tissue source sites. Also, all tissues were taken out of cold storage and sequenced at the same center. No batch effects are observed. There are no well-defined clusters in figure 2.1b. The data does not naturally cluster based on cancer subtype and thus classifying the samples is not a simple task. PCA clustering for more covariates can be found in Appendix A. (Refer figure A.1)

### 2.1.3 Data Processing

The gene expression (or mRNA expression) data was obtained by mRNA-Seq as RSEM normalized log2 read counts. Missing values were eliminated, resulting in 12660 genes (or features). The DNA Methylation data was obtained from Infinium HumanMethylation450 arrays to measure the level of methylation at known CpG sites as beta values. The beta values were averaged over probes on the same gene and thus the data was reduced to average beta values of 24039 genes (or features). The miRNA expression data was obtained by miRNASeq and is available as read counts per million reads. Missing values were eliminated, resulting in 387 microRNAs (or features).

The samples were divided into train and validation sets such that both the sets had the same distribution of cancer subtypes. ANOVA and a greedy selection method were then used to reduce the number of features to a thousand per omics (except for miRNA which was reduced to 257 features based on significance of the ANOVA f-value after correction for multiple testing). A thousand features were selected because of the fall in F-values observed beyond the top thousand or so features. The greedy selection methods selected features with high F-values that were not significantly correlated with other selected features.

The features were then centered to 0 and scaled to have a standard deviation of 1. The validation set's features were centered and scaled using the train set's means and standard deviations.

## 2.2 Molecular Taxonomy of Breast Cancer International Consortium

### 2.2.1 Data Availability

The METABRIC multi-omics data set used in this project requires controlled access. A request to access the data needs be made by writing to the Data Access Committee and completing requisite forms and agreements. The data set is stored in the European Genome-Phenome Archive (EGA) and can be viewed here: https://ega-archive.org/studies/EGAS00000000083. Upon obtaining access, the data sets were downloaded using pyEGA3 - a python based EGA download client. The associated clinical data is made available as a supplementary table in the original publication (Curtis *et al.*, 2012).

### 2.2.2 Data Preliminaries

Only the gene expression data was used from the METABRIC multi-omics data set, since this was primarily used as an independent validation for the model trained on the TCGA BRCA data set. This data set does not contain DNA methylation or miRNA expression data.

The data set is divided into discovery and validation sets and were released at different times. Each of the gene expression sets contains values of over 48,000 probes covering 19,876 genes. There are 997 samples in the discovery set and 989 samples in the validation set. The samples are classified into five subtypes: LumA, LumB, Basal, Her2, and Normal-like. The distribution of samples across cancer subtypes is presented in table 2.3 below.

| Set | LumA | LumB | Basal | Her2 | Normal | **Total** |
|---|---|---|---|---|---|---|
| Discovery | 466 | 268 | 118 | 87 | 58 | 997 |
| Validation | 255 | 224 | 213 | 153 | 144 | 989 |
| **Total** | 721 | 192 | 331 | 240 | 202 | 1986 |

Table 2.3: Cancer subtypes in the METABRIC data set

The data has been pre-processed and normalized according to the protocols de-

scribed in the supplementary material of the original publication (Curtis *et al.*, 2012). The data has been corrected for batch effects. Clinical variable like age, tumour cellularity, treatment administered and more are available.

The PCA clustering illustrated in figure 2.1 help visualize whether the data clusters based on any technical or biological characteristics. As observable in figure 2.1c, samples do not cluster based on tissue collection site. This implies that there are no batch effects. Even in figure 2.1d, the clusters are not distinctly defined, but are better defined than in figure 2.1b. Classifying the samples is a hard problem especially considering that the two data sets do not seem to have the same distribution. PCA clustering for more covariates can be found in Appendix A. (Refer figure A.2)

### 2.2.3 Data Processing

The gene expression (or mRNA expression) data was obtained using a Illumina HumanHT-12 Expression BeadChip microarray which has over 48,000 probes covering 19877 genes. illuminaHumanv3.db R package was used to convert manufacturer identifiers to entrez gene identifiers (Mark Dunning, 2017). Six samples in the validation set were removed for not having a cancer subtype label.

The data was averaged over probes covering the same genes to get 19877 genes (or features). These were then reduced to a thousand top features using ANOVA and a greedy selection method. Features with high F-values and low correlation with one another were selected. Since, this processing involves the use of cancer subtype labels, only the discovery (or training) set was used.

The features were then centered to 0 and scaled to have a standard deviation of 1. The validation set's features were centered and scaled using the discovery set's means and standard deviations.

## 2.3 Harvard Brain Tissue Resource Center (HBTRC)

### 2.3.1 Data Availability

The Human Brain Agilent data set obtained from the Harvard Brain Tissue Resource Center is part of a bigger data set released by Zhang *et al.* (2013), with the GEO database accession number GSE44772. The data is also available on University of Tennessee's GeneNetwork with the GN Accession numbers GN326, GN327, and GN328 representing mRNA expression data for Cerebellum, Primary Visual Cortex, and Prefrontal Cortex tissues respectively.

### 2.3.2 Data Preliminaries

This data set is a multi-tissue matched data set, i.e., samples from different tissues of the same patient are taken and gene expression is measured. Gene expression data was obtained using a custom-made Agilent 44K microarray. Each tissue's gene expression set contains 39,280 probes covering 19,539 genes. There are 384 matched samples in the data set of the following three categories: Alzheimer's Disease (AD), Huntington's Disease (HD), and Normal Control (NC). The distribution of samples across the categories is presented below in table 2.4 below.

| Alzheimer's Disease (AD) | Huntington's Disease (HD) | Normal Control (NC) |
|---|---|---|
| 186 | 84 | 114 |

Table 2.4: Categories of patients in the HBTRC data set

The data has been pre-processed and normalized. Instead of the general Z-Score, which has mean 0f 0 with a standard deviation of 1 unit, this data is normalized to 2Z + 8, i.e., it has been re-scaled to have a mean of 8 units with a standard deviation of 2 units.

### 2.3.3 Data Processing

Each of the three tissues are processed separately. The values of the 39,280 probes are averaged over the genes they cover to reduce the feature set to 19,539 genes. Features showing zero readings for all samples were removed.

The samples were then divided into train and validation sets in a stratified manner. All further processing was done using only the train set's labels. Using ANOVA and a greedy selection methods, a thousand features with high ANOVA F-values and low correlation with one another were selected for each tissue.

The features were then centered to 0 and scaled to have a standard deviation of 1. The validation set's features were centered and scaled using the discovery set's means and standard deviations.

# CHAPTER 3

# METHODS

All of the data processing was done in R and RStudio Server (R Core Team, 2019; RStudio Team, 2018) primarily using Tidyverse (Wickham *et al.*, 2019). The deep learning model was coded in PyTorch (Paszke *et al.*, 2019). The data imputation and over-sampling was done in python using scikit-learn classes (Pedregosa *et al.*, 2011). The gene set enrichment analysis was done in R using the fgsea package (Korotkevich *et al.*, 2016). Feature importance was measured using LIME (Ribeiro *et al.*, 2016) and SHAP (Lundberg and Lee, 2017).

## 3.1 Data Pre-processing

Only samples that had matched data across omics or tissue types were kept. The rest were removed. The train-test split was done early in the pipeline and cross-validation was not done because ANOVA and greedy selection had to be done to reduce the feature set and this required label information. Thus, only the training set was used for this pre-processing. No data leakage was allowed.

The 80-20 train-test split was done in a stratified manner, i.e., it was ensured that both the sets had an equal proportion of cancer subtypes/disease types. ANOVA was performed on the train set, significant features (Bonferroni correction adjusted p-value $<$ 0.001) were kept, and the top 1000 features, sorted by ANOVA F-values, were selected such that no feature had a Pearson correlation of more than 0.7 with another. All of the microRNA features that had an ANOVA adjusted p-value $<$ 0.001 were kept.

The data was normalized to obtain z-scores using only the train set's means and standard deviations to prevent data leakage. The imbalanced training set was balanced using Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla *et al.*, 2002; Blagus and Lusa, 2013).

## 3.2 Pairwise Feature Interactions

Inter-modality and intra-modality information can be utilized as pairwise interaction features. These pairwise features can be obtained by multiplying the features from the same modality/data type (see fig. 3.1a or different modalities/data types (see fig. 3.1b together to obtain a new set of features that constitutes a new modality/data type. Here, modality/data type refers to different omics types in a multi-omics data set or different tissue types in a multi-tissue data set.

For intra-modality interactions, features within the same modality are multiplied pairwise and the most informative of these generated features are selected to constitute a new modality/data type. For inter-modality interactions, features of two different modalities are multiplied pairwise and the most informative of these generated features are selected to constitute a new modality/data type.

ANOVA and greedy selection can be subsequently used to condense the newly generated feature set as described in section 3.1 paragraph 2.

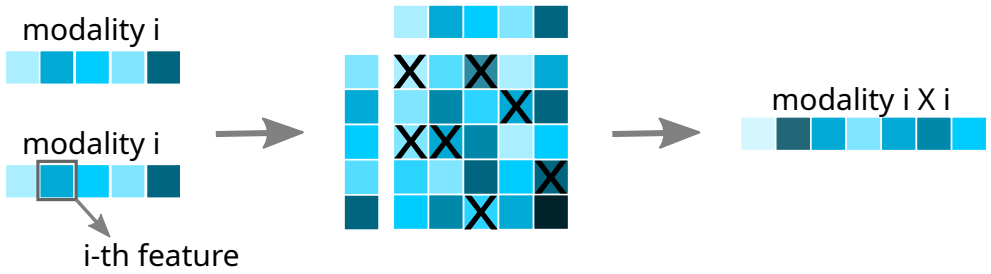
While the primary data set is analogous to the main effect terms in a linear regression equation, the intra-modality feature set is analogous to squared terms and the inter-modality feature set to interaction effect terms in a linear regression equation. These additional derived data sets aid in inferring relationships in the data better and subsequently in making better predictions.

## 3.3 Graph Convolution Network

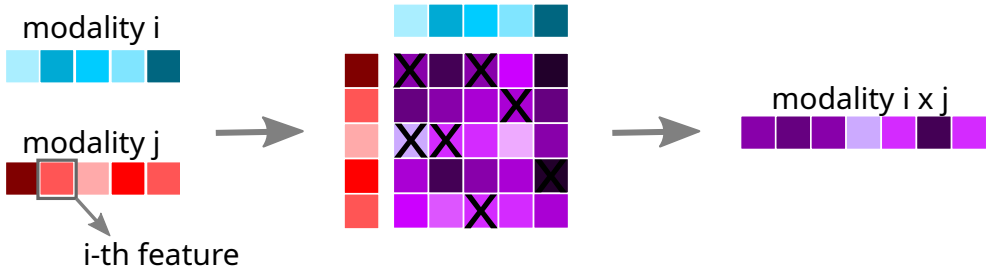
Graph Convolutional Networks, as described by Manessi *et al.* (2020), are a special class of neural networks whose goal is to learn a function of features on a graph $G = (V, E)$ which takes as input:

- $X$, an $N \times D$ feature matrix, that contains the feature descriptions for $N$ samples with $D$ features each, and
- $A$, an $N \times N$ adjacency matrix, which describes the graph structure as a matrix

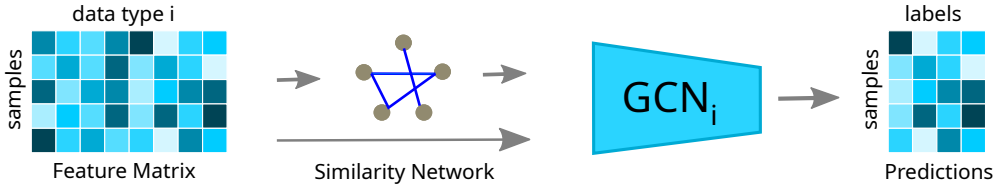
and generates $Z$, an $N \times C$ output matrix, where $C$ is the number of categories or labels. (Duvenaud *et al.*, 2015) Each of the individual modalities is input into a separate GCN. To convert the data into a GCN input, some processing needs to be done. We have



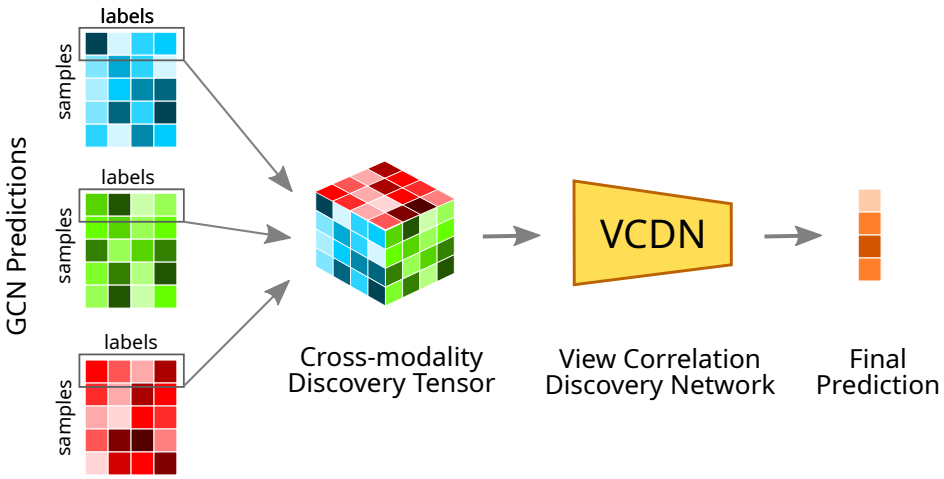

(a) Intra-modality interaction




(b) Inter-modality interaction




(c) Graph Convolutional Network (GCN)




(d) View Correlation Discovery Network

Fig. 3.1: **Methods**: (a) and (b) The pairwise features allow the inclusion of intra-modality and inter-modality interaction knowledge. X's mark the selected features. (c) The GCN utilizes sample similarity within each modality to predict cancer subtypes. (d) The VCDN combines individual GCN predictions and outputs final predictions.

the $N \times D$ feature matrix $X$ where $N$ is the number of patient or tissue samples and $D$ is the number of features. Thus, we now have a list of nodes, but no edges.

In order to get the adjacency matrix $A$, Wang *et al.* (2020) generated a patient similarity matrix. The patient similarity matrix is constructed using the cosine similarity measure. If the cosine similarity between a pair of nodes, here patients, is greater than a threshold $\epsilon$, then an edge is said be formed between the nodes. The weight of the edges is equal to the cosine similarity between the nodes. The adjacency between nodes $i$ and $j$, $A_{ij}$, is calculated as:

$$A_{ij} = \begin{cases} s(x_i, x_j), & \text{if } i \neq j \text{ and } s(x_i, x_j) \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where $x_i$ and $x_j$ are the feature vectors of node $i$ and node $j$ respectively, and $s(x_i, x_j)$ is the cosine similarity between node $i$ and $j$.

Now that we have the input for the graph convolutional network ready, we construct the network itself. Each GCN is made up of two to three layers. Each layer is defined as:

$$\begin{aligned} H^{(l+1)} &= f(H^{(l)}, A) \\ &= \sigma(AH^{(l)}W^{(l)}), \end{aligned} \tag{3.2}$$

where $\sigma(\cdot)$ is a non-linear activation function like sigmoid or relu, $H^{(l)}$ is the input of the l-th layer and $W^{(l)}$ is the weight matrix of the l-th layer.

As illustrated in figure 3.1c, the feature matrix consists of $N$ samples with $D$ features each. Cosine similarity between the features are used to construct a similarity network. The two inputs are used by the GCN to generate predictions.

This graph convolutional network is trained on all the training samples together so as to learn cross-sample relations. And while testing the model, the test sample is appended to the training sample set and submitted to the model. The model utilizes both the test features and relations between the test and train samples for classification.

## 3.4 View Correlation Discovery Network

As each of the graph convolutional networks considers individual modalities and makes the prediction, the most intuitive next step would be to take a linear combination of these labels to generate a final set of labels, which will then have been based on the complete multi-modality data set. But, that would be too simple and not consider any cross-modality correlations. To consider these, cross-modality correlations, a view correlation discovery network is used. (Wang *et al.*, 2020)

The original view correlation discovery network (Wang *et al.*, 2019) was used to generate views between discrete shots of an object. In simpler words, consider an object on a table and imagine a circle around it with a radius of one metre. Now take two pictures of the object from two nearby points on the circumference. A view correlation discovery network aims to integrate the features of the two pictures and thereby imagine a view of the object from a point between the two points from which the pictures were taken. This requires the network to learn intra-view and cross-view relations in the label space itself.

The original work on VCDN was designed for data with two views. MORONET (Wang *et al.*, 2020) extends the VCDN framework to three views. They hard-coded the model to take exactly three views' labels, i.e., mRNA expression data, DNA methylation data, and miRNA expression data (from the TCGA cohorts), but it was easily extended to a variable number of views.

Considering three views, $i$=1,2,3, let $\hat{y}_j^{(i)} \in \mathbb{R}^c$, represent the $j$-th training sample, where $c$ is the number of labels. A cross-modality discovery tensor $C_j \in \mathbb{R}^{c \times c \times c}$ is constructed, where each entry of $C_j$ is calculated as:

$$C_{j,abc} = \hat{y}_{j,a}^{(1)} \hat{y}_{j,b}^{(2)} \hat{y}_{j,c}^{(3)}, \tag{3.3}$$

where $\hat{y}_{j,c}^{(i)}$ is the $c$-th entry of $\hat{y}_j^{(i)}$. The tensor so obtained, $C_j$ is reshaped to a $c^3$ dimensional vector and forwarded to the VCDN($\cdot$) for final classification.

As illustrated in figure 3.1d, the GCNs' predictions are aggregated together to form a cross-modality discovery tensor which is then used by the VCDN to infer cross-modality relations and generate the final predictions.

There are two things of note here:

- The VCDN(·) itself is a two layer fully connected network which outputs the final label predictions, based on the cross-modality discovery tensor formed by integrating the labels predicted by the individual graph convolutional networks.
- The VCDN(·)'s input, the cross-modality discovery tensor scales exponentially, i.e., it is of the size $C^N$ where $C$ is the number of output labels and $N$ is the number of modalities included in the analysis.

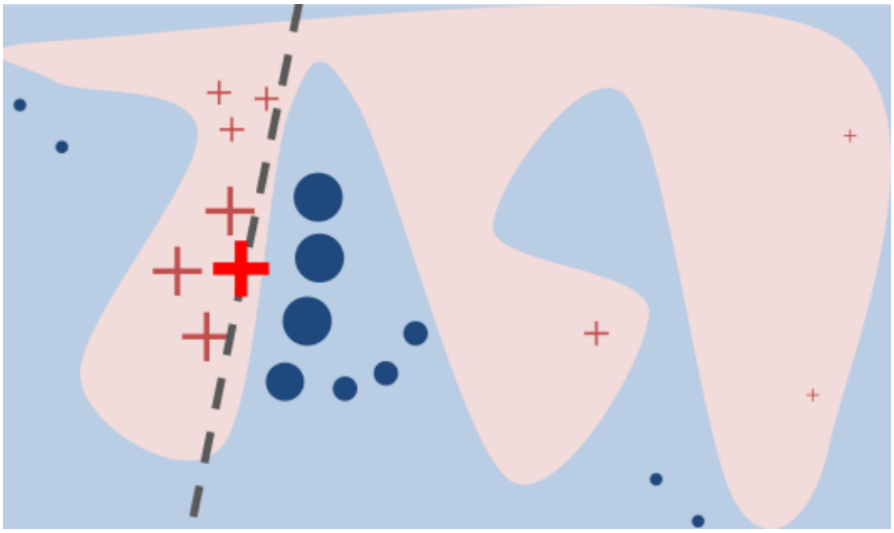
## 3.5 Feature Importance Measures

Biomarkers are objectively measurable features, either molecular like gene expression levels or clinical like body temperature or weight, that are indicative of disease processes. While predicting type or subtype of a disease is useful for diagnosis and consequently prognosis, identifying biomarkers is vital to understanding the disease process thoroughly enough to diagnose, prognose, and treat a disease completely.

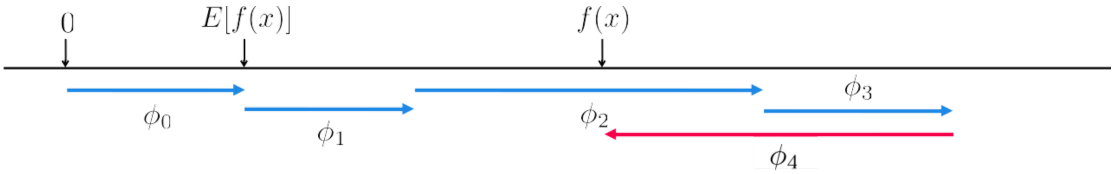
As our GCN-VCDN model is trained on multi-omics or multi-tissue data, it learns and infers relations within modalities, between modalities, and between the data modalities and the disease types or subtypes. Directly, this model can only be used to predict the disease type or subtype of a new sample. In order to obtain the learnings and inferences of the model about the disease, feature importance measures have to be used. Here, we use two existing model-agnostic methods to obtain feature importance.

### 3.5.1 Local Interpretable Model-agnostic Explanations (LIME)

Ribeiro *et al.* (2016) define an explanation as "a local linear approximation of the model's behaviour." While the complete model may be a black box and have a complex decision boundary in the feature space, LIME assumes that at any given point on the decision boundary, it is linear and can be approximated by a sparse linear model. In the fig 3.2a, the pink-blue boundary is the original model's complex decision boundary and the big red cross represents the sample point. It's feature values are perturbed and the model is used to generate predictions. LIME then uses these perturbed samples and the model's predictions to learn a sparse linear model (the dashed line) which explains the original model's behaviour in the sample point's locality.



(a) **LIME** takes a model's complex decision boundary and a single sample point (here, the big red cross) and perturbs the sample to learn a sparse linear model in the locality of that point. Image was obtained from LIME's github repository. ©Marco Tulio Correia Ribeiro




(b) **SHAP** has a base prediction $E[f(x)]$ for sample $x$. The individual feature contributions estimated as SHAP values, $\phi_i$s contribute toward estimating the true prediction $f(x)$ for the sample $x$. Since ordering matters, $\phi$s are obtained by averaging over all possible orderings of the subset of features used.

Fig. 3.2: **Methods**: Feature importance is measured using LIME and SHAP.

Given a sample, we can infer the features that were important in making the classifying decision from the weights of the approximated linear model. This local explanation can be a form of personalized medicine and allow the clinician to diagnose/treat the patient accordingly.

Also, as described above, LIME can be used to approximate the decision boundary at many such points and then at each point, we can infer the features that were important in making the decision from the linear model. Thus, LIME can be used to know features that are important to classification globally and these features can be used as biomarkers. The lime python package was used in this project.

### 3.5.2 SHapley Additive exPlanations (SHAP)

Shapley value is a concept from coalitional game theory that describes how to fairly distribute the "payout" among the players. Consider our model as the game. The features are players and they collude together to winning the game, i.e., making the correct prediction. Shapley value is the average marginal contribution of a feature value across all possible coalitions. While it is a theoretically sound concept, computing Shapley values is computationally expensive because for $k$ features, there are $2^k$ possible coalitions of features and each of these has to be computed to get the explanations.

Lundberg and Lee (2017) bring the concept of additive feature attribution methods and Shapley values together and propose SHAP values as a measure of feature importance. An explanation model for a model is an interpretable approximation of the original model. SHAP's explanation model has a base value that would be predicted if we did not know any features to a particular sample, and following that, SHAP values for each feature that explains how the feature when added to the mix aids to the classification of the sample. In a non-linear model, the order in which features are added matters and that is tended to by averaging over all possible orderings of the subset of features.

SHAP's model-agnostic explanation method is called Kernel SHAP which is a combination of LIME as described in subsection 3.5.1 and Shapley values. Kernel SHAP like LIME provides local explanations and local feature importances which can subsequently be summarised to get global feature importances. The shap python package was used in this project.

## 3.6 Enrichment Analysis

Gene set enrichment analysis (GSEA) is a statistical method to identify sets of genes that are over represented in a larger ranked list genes. GSEA involves calculating an enrichment score for each gene set. A high enrichment score implies over representation of the gene set at the top of the ranked list and a high negative score implies over representation of the gene set at the bottom of the ranked list. Then the significance level of the enrichment score is calculated and the calculated p-value is adjusted for

multiple hypothesis testing. (Subramanian *et al.*, 2005)

The gene set collections used in this project include: Molecular Signatures Database (MSigDB) (Liberzon *et al.*, 2011, 2015), DisGeNET database 7.0 (Piñero *et al.*, 2019), DriverDBv3 (Liu *et al.*, 2019), and miRCancer (Xie *et al.*, 2013). Gene set collections H, C2, C3, C4, C5, and C6 of MSigDB were were used. The R package `fgsea` was used to perform the gene set enrichment analysis (Korotkevich *et al.*, 2016). Fgsea performs fast pre-ranked gene set enrichment analysis.

To quantify the enrichment analysis in the context of comparing the ranked lists obtained from multiple models, the number of significantly enriched gene sets was used. Higher the number of significantly enriched gene sets, better the model.

## 3.7 Modality Imputation

The multi-modality data set was divided into train and test sets. One or more modality's values were removed and then imputed based on the remaining modality's values and the train set. Spearman correlation coefficient was used to quantify the imputation because we are interested more in the position of a feature with respect to other features than it's value alone.

Many methods were tried for the imputation and it was found that K-Nearest Neighbours (KNN) was the best methods to perform the imputation. Other methods tried include imputing with mean, median, elastic net regression, k nearest neighbours regression (iterative), random forest regression (iterative). Iterative here means that values of each feature were imputed using all other features in an iterative manner until no more change was observed in an iteration.

The analysis was performed in python and the following scikit-learn classes were used: SimpleImputer, Iterative Imputer, and KNNImputer from sklearn.impute, ElasticNet from sklearn.linear_model, KNeighborsRegressor from sklearn.neighbors, and RandomForestRegressor from sklearn.ensemble. (Pedregosa *et al.*, 2011)

# CHAPTER 4

# RESULTS

## 4.1 The GCN-VCDN model classifies test set the best

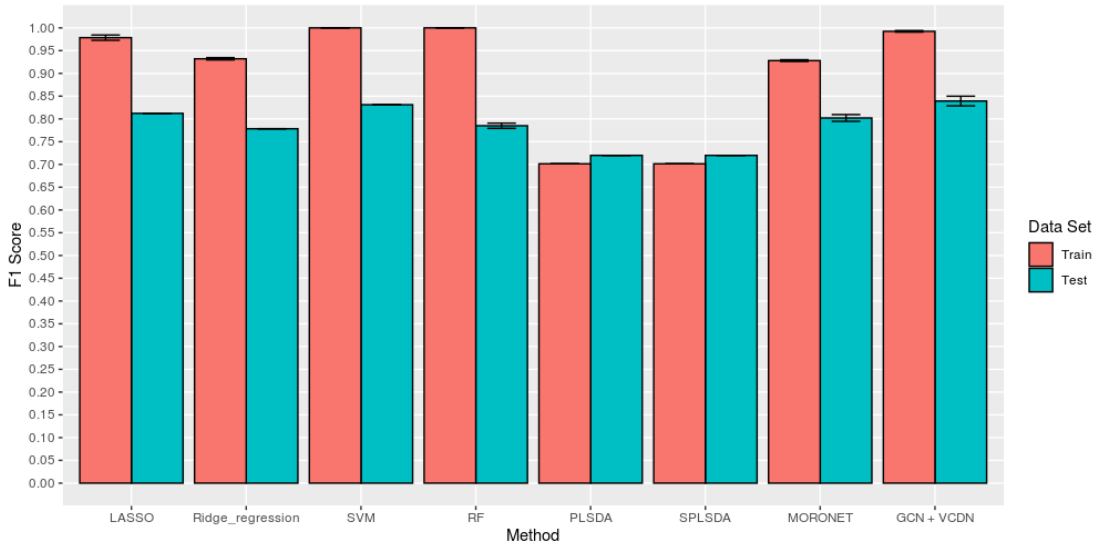
The GCN-VCDN model was tested against simple regression methods (LASSO and ridge regression), machine learning models (SVM and random forest), state-of-the-art multi-omics integration methods from the mixOmics DIABLO framework (PLSDA and SPLSDA) (Singh *et al.*, 2019), and the original GCN+VCDN multi-omics integration study MORONET Wang *et al.* (2020) in predicting the cancer subtypes of the TCGA BRCA multi-omics data set. Our model had the highest test set F1 scores as seen in figure 4.1a and the supplementary table B.1.

Inferring sample-sample relations using the patient similarity matrix appears to aid the model in classifying the data better than other models. It performs better than MORONET because of better pre-processing and due to the use of Synthetic Minority Over-sampling TEchnique (SMOTE) to improve learning on the imbalanced data set (Chawla *et al.*, 2002; Blagus and Lusa, 2013). All models were trained thrice and the resulting F1 scores' mean and standard deviation are seen illustrated in figure 4.1a and the supplementary table B.1.

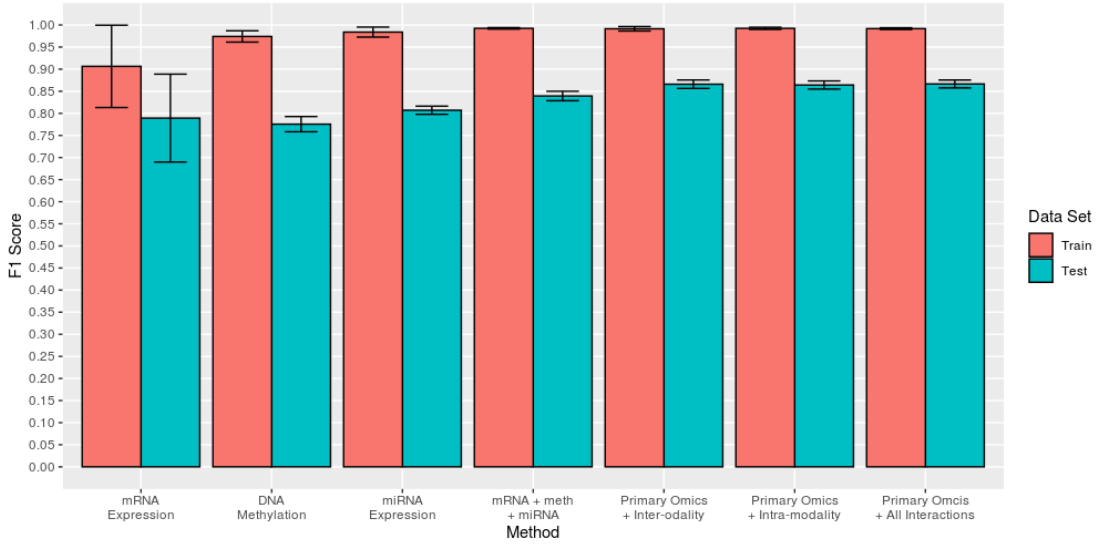
## 4.2 Pairwise feature interactions improve model performance and interpretability

To understand what (sub)sets of the data set, the individual omics, multi-omics, or multi-omics along with pairwise feature interactions (intra-modality, inter-modality, or both), are important, the GCN-VCDN model was trained with different inputs. While training with a single omics, only a GCN was used.

While individual omics sets themselves can classify the test set with F1 scores nearing 0.80, inclusion of other omics as well as pairwise interaction features improves




(a) Comparison, measured by F1 score, of different methods' cancer subtype classification on the TCGA BRCA data set. (LASSO = Least Absolute Shrinkage and Selection Operator, SVM = Support Vector Machine, RF = Random Forest, PLSDA = Partial Least Squares Discriminant Analysis (Singh *et al.*, 2019), SPLSDA = Sparse Partial Least Squares Discriminant Analysis, MORONET = Multi-Omics gRaph cOnvolutional NETworks (Wang *et al.*, 2020))




(b) Comparison, measured by F1 score, of cancer subtype classification on the TCGA BRCA data set based on different omics used. Here, Primary Omics refers to the three individual omics: mRNA Expression, DNA Methylation, and miRNA Expression, Intra-modality refers to mRNA X mRNA, meth X meth, and miRNA X miRNA interactions, Inter-modality refers to mRNA X meth, meth X miRNA, and miRNA X mRNA interactions, and All Interaction refers to Intra-modality and Inter-modality interactions. (meth = DNA Methylation)

Fig. 4.1: **Results: The GCN-VCDN model classifies test set the best.** (a) Test set F1 score for our model is better than other models. (b) Test set F1 score for our model is best when using all the primary omics and the pairwise interactions.

the classification. F1 score for classification with all omics and pairwise interactions included was 0.87, significantly higher than with just the primary omics without interaction features. All models were trained thrice and the resulting F1 scores' mean and standard deviation are seen illustrated in figure 4.1b and the supplementary table B.2.

Consider a heterogeneous multi-layered network with features as nodes and omics types as layers. The feature values represent the node description and the pairwise feature interactions represent the edge description or edge weights. Including the pairwise feature interactions in the model and then obtaining LIME and SHAP scores can be used to effectively quantify these edges and improve interpretability of the data and disease.

## 4.3 Local explanations enable personalized medicine

Both SHAP and LIME can be used to interpret the GCN-VCDN model and its output in order to learn which features' values are most important in classifying the cancer as a particular subtype. In figure 4.2, SHAP explanation for classifying the patient TCGA-D8-A1XU-01 with a Luminal A type breast cancer is visualised in multiple forms. The decision plot in figure 4.2a shows how much the most important features (illustrated in a descending order on the y-axis) influence the classification. Each of the features attribution is additive in nature. The solid black horizontal line indicates the base prediction (or base value) of the model. As one moves upward along the y-axis, it is observable that the features' shap values add up to the model output for the TCGA-D8-A1XU-01 sample.

The model's local explanations can also be visualized as force plots as observable in figures 4.2b and 4.2c. The additive attribution of individual features is more prominently observable in these force plots. $f(x)$ here the model's output for sample TCGA-D8-A1XU-01. In 4.2c, the top features are visualized along with their normalized values (z-scores). The force plots only showcase the features' contribution toward the predicted subtype as opposed to all the subtypes in the decision plots.

Such plots can be used by clinicians to understand a patient's cancer better, and thus improve diagnosis and treatment. Clinicians can also learn about the most impor-

tant predictive features in a population using stacked bar plots as exemplified in the supplementary figure A.3.

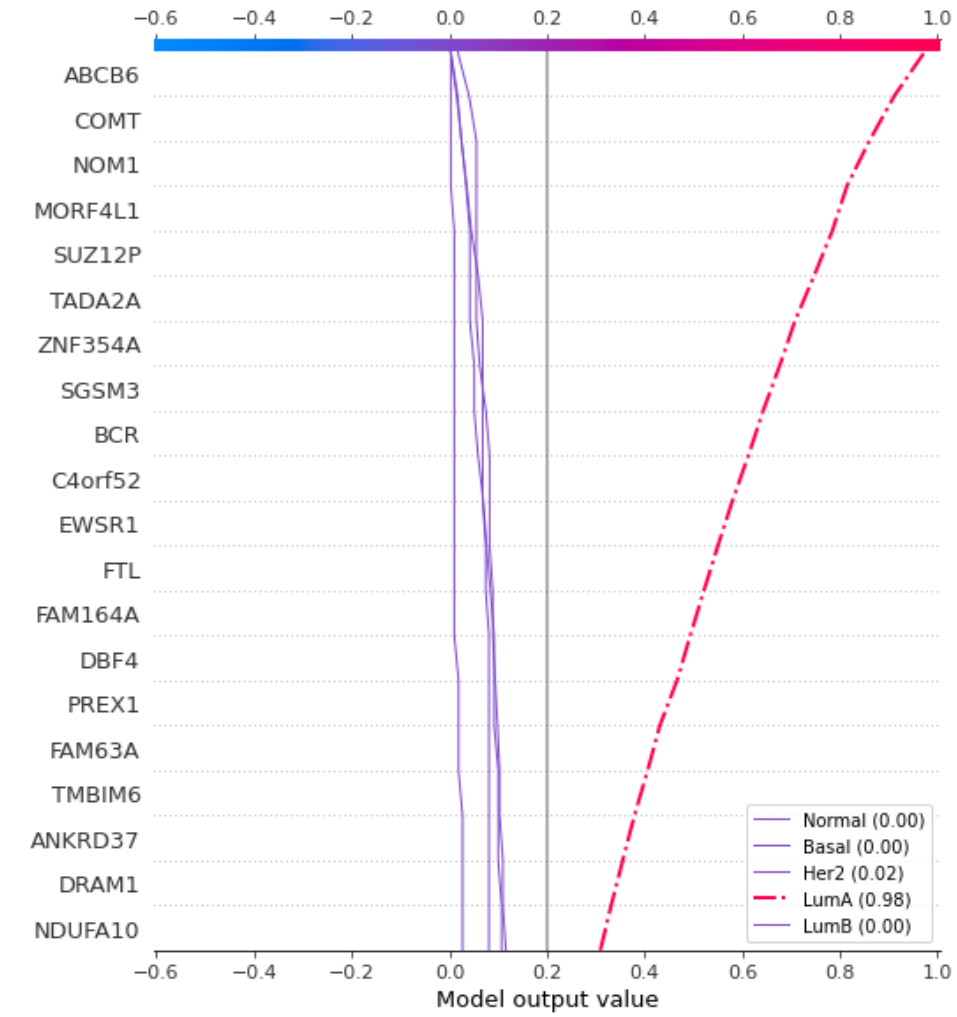
## 4.4 DACH1 suppresses breast cancer

Dachshund homolog 1 (DACH1) is a well-studied breast cancer associated cell fate determination factor (Popov *et al.*, 2009; Powe *et al.*, 2014; Zhao *et al.*, 2015; Zhang *et al.*, 2015; Xu *et al.*, 2017). DACH1 is expressed higher in normal tissue or Luminal A type breast cancer, which has a good prognosis, while it is expressed lesser in more aggressive Basal-like breast cancer. In mouse models, DACH1 has been demonstrated to be suppressed in cancer stem cells and that their expression is inversely proportional to epithelial-mesenchymal transition. In humans, higher DACH1 expression is associated with prolonged survival time.

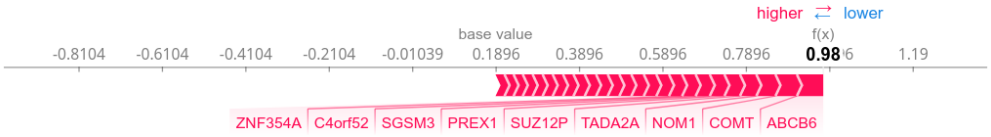
Aggregating the feature importances across all the samples yields an approximation of globally important features in the classification of breast cancer. Biomarkers have been thus obtained using both LIME and SHAP. DACH1 was found to be the most important gene expression feature by both LIME and SHAP (see table 4.1). Likewise SNORA23;IPO7 and MIR342 have been found to be the most important DNA Methylation and microRNA biomarkers respectively.

Small Nucleolar RNA, H/ACA Box 23 (SNORA23) is a small nucleolar RNA, a class or RNAs that primarily accumulate in the nucleoli and are involved in post translational modification and maturation of rRNAs and snRNAs. SNORA23 has been observed to accumulate in highly metastatic cells but not in normal tissue (Cui *et al.*, 2017). SNORA1 levels are inversely correlated with patients' survival time post diagnosis. Cui *et al.* (2017) also showed that administering SNORA23 in mice slowed xenograft tumours. In pancreatic adenocarcinoma, tumours with high SNORA23 levels were found to respond well to the drug axitinib (Liu *et al.*, 2020*b*).

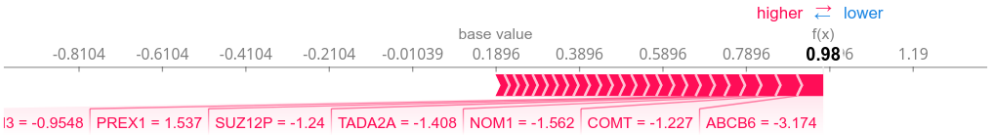
Importin 7 (IPO7) facilitates nuclear protein import, either as an autonomous nuclear transport receptor or in association with the importin-beta subunit KPNB1. IPO7 is often overexpressed in cancer. It is upregulated by c-Myc and downregulated by p53 (Golomb *et al.*, 2012). Overexpression of IPO7 has been particularly noted and thought




(a) Decision plot




(b) Force plot




(c) Force plot with normalized feature values (z-scores)

Fig. 4.2: **Results: Local explanations enable personalized medicine.** Visualisation of model explanations using SHAP for the patient TCGA-D8-A1XU-01. (a) Decision plot has features in the decreasing order of importance with the solid black horizontal line indicating the base prediction. (b) & (c) The force plots only indicate the features' contribution toward the model's top prediction.

| Omics Data Type | Top Biomarkers (LIME) | Top Biomarkers (SHAP) |
|---|---|---|
| mRNA Expression | DACH1, NTRK2, FAM107A, FGD3, C8orf84, FGF2, LRRN1, PIF1, CLSTN2, DNAJC12, MMP11, ANKRD29, C4A, CDH1, TINAGL1, ATAD3C, PHGDH, CNN1, CACNA2D2, SLC4A8, C16orf71, RAI2, INPP5J, SLC7A5, SHISA2 | DACH1, NTRK2, CLSTN2, LRRN1, DNAJC12, FAM107A, FGF2, C4A, PIF1, SLC1A1, C8orf84, MMP11, FGD3, ANKRD29, PTHLH, MYB, CACNA2D2, L3MBTL4, INPP5J, TINAGL1, ZNF135, SLC4A8, LMX1B, SHISA2, PHGDH |
| DNA Methylation | SNORA23;IPO7, DOCK2;FAM196B, C21orf130, HIST3H2BB;HIST3H2A, DLX6;DLX6AS, LOC727677, NAP1L6, CDC123, OR8U8;OR5AR1, ZIC4;ZIC1, LBX2;LOC151534, OR10G9, LY6E, KLC2;RAB1B, DCT, PCDHGA4;PCDHGA6;, MIR195, ZNF492, MIR1304, ESPNL;SCLY, UTS2D;CCDC50, FBXO47, POU3F3, DERL3, VPS28;NFKBIL2 | SNORA23;IPO7, DOCK2;FAM196B, C21orf130, HIST3H2BB;HIST3H2A, LOC727677, DLX6;DLX6AS, NAP1L6, ZIC4;ZIC1, LBX2;LOC151534, ZNF492, OR8U8;OR5AR1, KLC2;RAB1B, C21orf29;KRTAP10-5, CDC123, DCT, PCDHGA4;PCDHGA6;, MIR195, C9orf66;DOCK8, ESPNL;SCLY, LY6E, ICAM5;ICAM4, POU3F3, OR10G9, LRRC36;KCTD19, FBXO47 |
| miRNA Expression | MIR342, MIR101-2, MIR30C2, MIR29C, MIR576, MIR130B, MIRLET7D, MIR3690, MIR345, MIR28, SNORD138, MIR3682, MIR3912, MIR450A2, MIR101-1, MIR196B, MIR425, MIR3074, MIR3922, MIR3605, MIR200B, MIR639, MIR1306, MIR15B, MIR26A2 | MIR342, MIR101-2, MIR29C, MIR30C2, MIR196B, MIR3912, MIR28, MIR581, MIR425, MIRLET7D, MIR345, MIR3610, MIR3605, MIR296, MIR1291, MIR15B, MIR3690, MIR450A2, MIR1245A, MIR887, MIR101-1, MIR193B, MIR3922, SNORD138, MIR944 |

Table 4.1: Top 25 biomarkers of each omics type selected by LIME and SHAP on the TCGA BRCA data set

to be of importance in breast, prostate, and lung cancers (Smith *et al.*, 2010; Szczyrba *et al.*, 2012; Lee *et al.*, 2017).

MIR342 expression in breast cancer has been associated with better prognosis. It is

significantly associated with estrogen receptor levels (Crippa *et al.*, 2014). Lindholm *et al.* (2019) found that HER2 signalling was regulated by MIR342 and that overexpression of MIR342 renders HER2 breast cancer cells less proliferative and susceptible to cellular stress. Downregulation of MIR342 leads to tamoxifen resistance in breast tumors and restoration of MIR342 levels sensitises cells to tamoxifen-induced apoptosis and reduces cell growth (Cittelly *et al.*, 2010). MIR342 targets and modulates Cofilin 1 (CFL1) which is involved in the cofilin signalling pathway, a therapeutic target in breast cancer treatment (Liu *et al.*, 2020*a*).

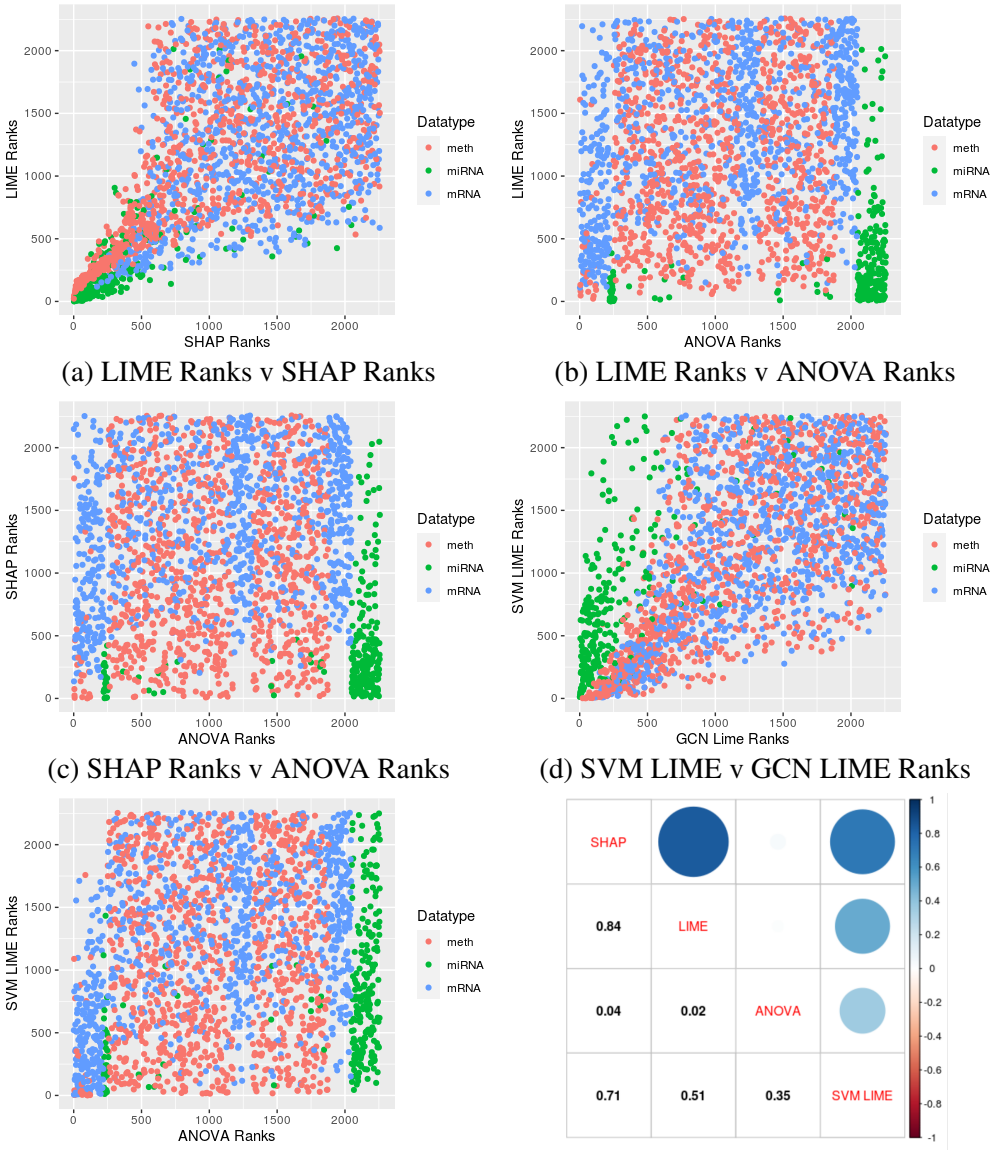
## 4.5 Global interpretability yields biomarkers

Global interpretability was achieved through both LIME and SHAP by inferring the most predictive features for each of the samples and then aggregating them. Thus, we obtained a ranked list of features each from LIME and SHAP, the top 25 of which can be seen in the table 4.1. The top features for each of the different omics type are displayed. This can be thought of as a biomarker panel.

### 4.5.1 LIME and SHAP feature rankings corroborate each other

As observable in figure 4.3a and table 4.1, the top predictable features are ranked similarly. This is as expected since both of them explain the same model and data. Significant deviation from each other's ranking would have laid doubt on the validity of the explanations themselves.

Also, both LIME and SHAP ranks differ significantly from the ANOVA ranks (fig. 4.3b and 4.3c). This implies that all this analysis was not in vain. If LIME and SHAP ranks aligned with ANOVA ranks, one could just use ANOVA to get biomarkers. Additionally, as seen in fig. 4.1a and table B.1, SVM performed on par with the GCN-VCDN model. So similar biomarker analyses were conducted for the SVM model too. SVM ranks do not align well with our model's LIME ranks or ANOVA ranks (fig. 4.3d and 4.3e). A summary of all the figures mentioned in this subsection can be seen in figure 4.3f.




(a) LIME Ranks v SHAP Ranks

(b) LIME Ranks v ANOVA Ranks

(c) SHAP Ranks v ANOVA Ranks

(d) SVM LIME v GCN LIME Ranks

(e) SVM LIME Ranks v ANOVA Ranks

(f) Feature importance scores' correlation

Fig. 4.3: **Results: Global interpretability yields biomarkers.** Visualizing the correlation of TCGA BRCA multi-omics' feature importance ranking as quantified by ANOVA, LIME, and SHAP on the GCN + VCDN model as well as SVM.

### 4.5.2 The GCN-VCDN model ranks features better than ANOVA

Differing of our model's feature ranking with ANOVA does not itself imply which of the rankings is better. Thus, gene set enrichment analysis was performed. Our GCN-VCDN model's feature ranking enriched for >100 gene sets each from the molecular signatures database (MSigDB), while ANOVA ranking enriched for only 49 gene sets and SVM enriched for none.

Table 4.2 lists the number of gene sets enriched for in MSigDB collections or sub-

collections. Both LIME and SHAP rankings enrich for gene sets in Gene Ontology collections. ANOVA rankings enrich particularly for gene sets in the chemical/genetic perturbations sub-collection which is curated manually. These gene sets come in pairs representing genes that are induced or repressed by the perturbation.

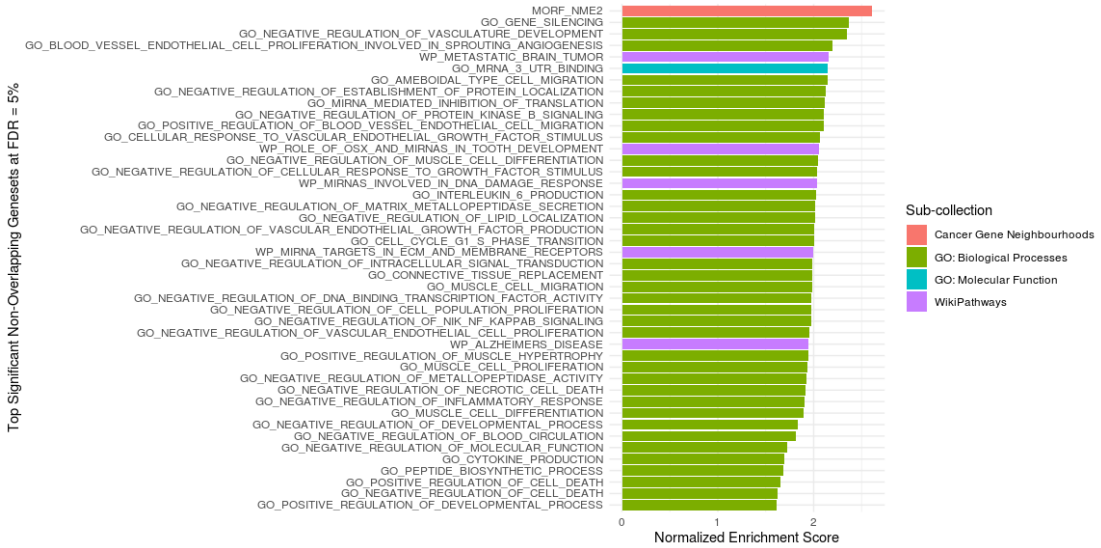
| Gene Set Collection | LIME | SHAP | ANOVA | SVM LIME |
|---|---|---|---|---|
| GO: Biological Processes | 120 | 92 | 9 | 0 |
| WikiPathways | 5 | 5 | 1 | 0 |
| GO: Molecular Function | 4 | 4 | 0 | 0 |
| Cancer Gene Neighbourhoods | 2 | 0 | 0 | 0 |
| Chemical/Genetic Perturbations | 1 | 0 | 35 | 0 |
| GO: Cellular Components | 1 | 0 | 0 | 0 |
| Reactome Pathways | 0 | 1 | 0 | 0 |
| Legacy Transcription Factor Targets | 0 | 1 | 0 | 0 |
| Hallmark | 0 | 0 | 2 | 0 |
| Cancer Modules | 0 | 0 | 1 | 0 |
| miRDB Targets | 0 | 0 | 1 | 0 |
| **Total** | 133 | 103 | 49 | 0 |

Table 4.2: Number of gene sets enriched for by the feature importance measures in different MSigDB (sub-)collections. FDR = 5%.

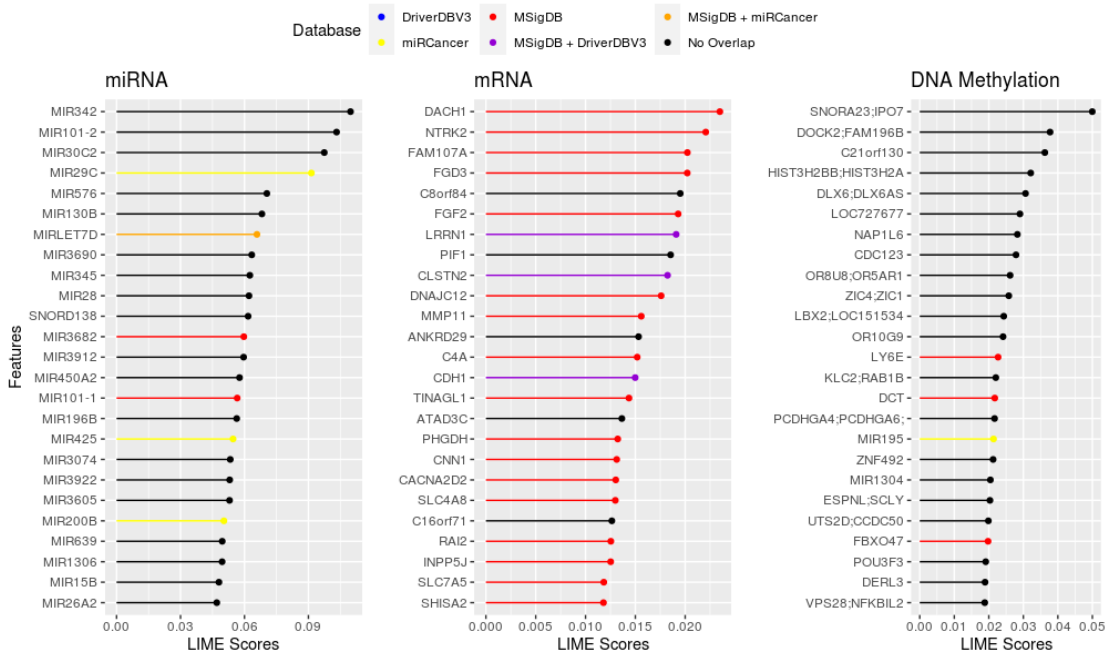
### 4.5.3 Top enriched gene sets are cancer associated

While the gene sets enriched for by the feature ranking are not directly associated with cancer or do not involved manual curation from observing cancer microarrays, the gene sets refer to biological processes and molecular functions that would be aberrant in tumours. Many of the enriched gene sets like those involved in regulation of vasculature development, sprouting angiogenesis, metastasis, DNA damage response, vascular endothelial cell proliferation, and more, clearly indicate toward tumour and metastasis.

Figure 4.4a illustrates the gene sets that were enriched for by LIME on our model. Benjamini-Hochberg procedure was applied to correct for multiple hypothesis testing reaching an FDR level of 0.05. However, the number of resulting gene sets was too high for easy visualization, so gene sets with over 20% overlap were removed. A similar illustration for gene sets enriched for by SHAP can be seen in supplementary fig. A.4a.




(a) Top gene sets enriched for using the feature rank list from LIME on the TCGA BRCA data set. Many of the enriched gene sets like those involved in regulation of vasculature development, sprouting angiogenesis, metastasis, DNA damage response, vascular endothelial cell proliferation, and more, clearly indicate toward tumour and metastasis.




(b) Top features selected by LIME and their membership in various gene set collections. Here, MSigDB refers to gene sets that are associated with cancer, a subset of the larger database. For more information on the databases/gene set collections used, refer section 3.6.

Fig. 4.4: **Results: Global interpretability yields biomarkers. (SHAP)** (a) Top enriched gene sets are cancer associated. (b) Model predicts novel biomarkers.

### 4.5.4 Model predicts novel biomarkers

Many of the top features, of each omics type, are not part of popular cancer databases or gene set collections, as seen in figure 4.4b. Fig. 4.4b shows the top 25 features for each omics type coloured by their membership in the databases: MSigDB, DriverDBv3, miRCancer, and combinations of them. This figure uses LIME rankings. For a similar figure with SHAP rankings, refer supplementary figure A.4b.

The absence of many of these top features in cancer associated data bases indicates that the databases are incomplete and also that these features represent novel biomarkers. For example, in section 4.4, association of MIR342, SNORA23, and IPO7 with cancer was made evident, but they do not appear in any of the cancer associated data bases. The features in black in fig. 4.4b and A.4b represent novel biomarkers.

## 4.6 The model generalizes well to unseen data set

In order to independently validate our model, we used the METABRIC data set. While it does not contain DNA methylation and microRNA expression data, two of the omics used from the TCGA BRCA data set, it contains over 1900 breast cancer samples with gene expression data and also the same set of cancer subtype labels. The independent validation was performed in a variety of ways as described and quantified in the subsections below. Since the METABRIC set contains only gene expression data, all mentions of features in the following subsections refer to gene expression features unless stated otherwise.

### 4.6.1 Features pre-selected from TCGA BRCA are also predictive for the METABRIC data set

Over 95% of the features pre-selected from the TCGA data set were present in the METABRIC data set. The median ANOVA p-value of these features in the METABRIC data set was 1.37e-18. This implies that the features selected from the TCGA set are useful for the classification task in the METABRIC set too. This median p-value is three orders of magnitude lesser than the median ANOVA p-value for a random sampling of

features. The pre-selected features from TCGA BRCA generalize well to the unseen METABRIC data set.

Additionally, when a GCN is trained on the METABRIC discovery set using only the features pre-selected from TCGA BRCA and then tested on the METABRIC validation set, the model achieves a test F1 score of 0.71. This again validates that features pre-selected from TCGA BRCA are useful for classifying cancer subtypes in the METABRIC data set as well.

### 4.6.2 Biomarkers from independent training on different data sets are correlated

Pre-processing and training was done on TCGA BRCA and METABRIC gene expression data independently and LIME was used to obtain feature importance. Since pre-processing selects only a thousand features from a population of >17,000 features, the input features for both the models were not the same. Regardless, F1 scores of 0.80 and 0.74 were achieved for the TCGA BRCA and METABRIC test sets respectively.

There was an overlap of 94 genes between the two 1000 feature samplings (TCGA BRCA and METABRIC) from a common set of 17,206 gene expression features. Considering that 47 of the TCGA pre-selected features were not present in the METABRIC set and 17 of the METABRIC pre-selected features were not present in the TCGA BRCA set, a hypergeometric test was performed. The 94 gene overlap is significant ($p = 8.11e\text{-}07$).

In the 94 feature overlap, the LIME scores from the two data sets were significantly correlated, with a Pearson correlation coefficient of 0.58.

### 4.6.3 Model trained on TCGA generalizes to METABRIC data

The GCN-VCDN model was trained on TCGA BRCA gene expression and then tested on the METABRIC gene expression data. This yielded a test set F1 score of 0.71. In this scenario, the TCGA features and METABRIC features were normalized separately. When the test set, i.e., METABRIC was normalized using the TCGA BRCA set's means and standard deviations, the test set F1 score achieved was 0.63.

While 0.63 is a good classification score on a completely unseen data set possibly from a different distribution, it was improved when the test set was normalized independently. There is no data leakage though. The only concern is that at time of deployment, the model may be classifying only a single sample and there will be no mean and standard deviation to normalize with. In this case, it is advised to combine the two distributions of data to obtain the means and standard deviations.

## 4.7 Omics imputation adds predictive power as well as interpretability

In many scenarios, like the METABRIC set, data for all the requisite omics is unavailable. In such a scenario, imputing the missing omics can add to the data's predictive power as well as interpretability in the form of feature importance scores as obtained from LIME and SHAP. Many methods were tried to impute different sets of missing omics. Figure 4.5 illustrates many of the methods tried like imputing with mean, elastic net regression, k-nearest neighbours, and random forest. For illustration of more methods including imputing with median, and KNN with more values of $K$, refer fig. A.5.

### 4.7.1 KNN performs imputation the best

Spearman correlation was used to quantify the different methods' imputed values with the true values. Spearman was used because we are more interested in the relative ranks of the features than the values themselves.

In the figures 4.5 and A.5, X-axis indicates the omics that were removed in the test set and subsequently imputed using the remaining omics. Each point is an individual omics feature. As seen on the x-axis of the figure 4.5, different omics types were imputed in the test set. KNN ($K = 50$) was the best method to impute the missing omics. While KNN (iterative) and Random Forest impute some omics well, they perform particularly poorly while imputing both DNA methylation and gene expression omics from just microRNA expression.

In the figures, random refers to imputing with random values form a uniform distribution with the same range as that of the respective features in the training set. This acts as a baseline. Imputing with the same value across samples for each feature, like mean and median, results in zero correlation with true values adding no value to the classification task. Elastic net also fails to impute values well.

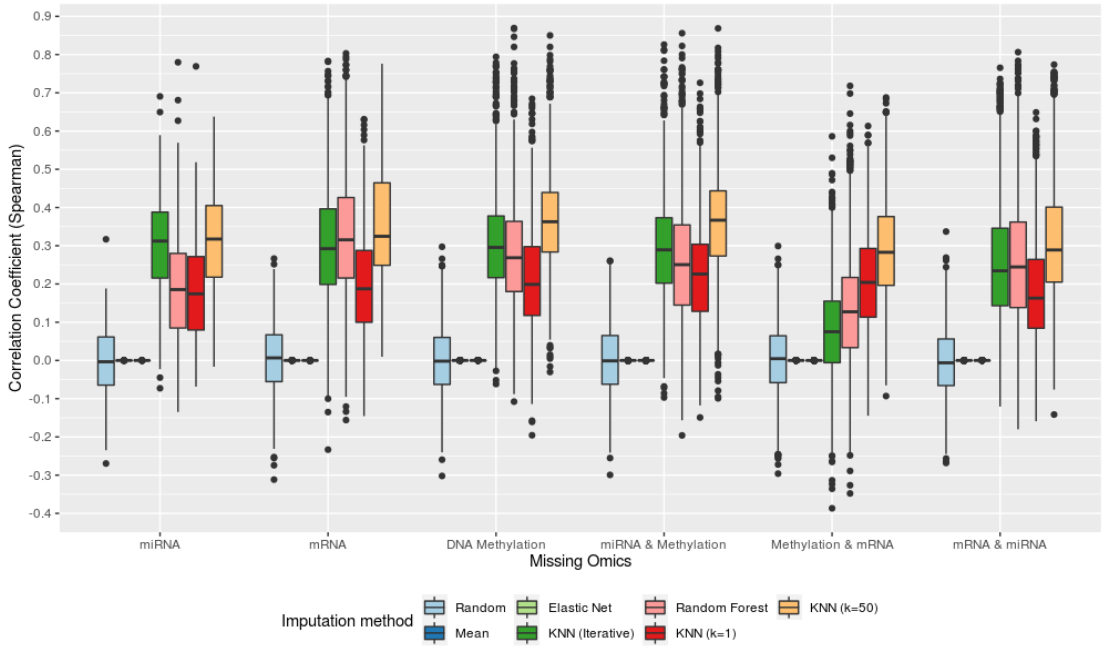



Fig. 4.5: **Results: Omics imputation adds predictive power.** Distribution of Spearman correlation coefficient of the features imputed by various methods with the true values.

KNN was tested with various values of $K$. While testing with the TCGA BRCA data set, $K = 50$ was found to be the best choice overall, while $K = 25$ was better some of the time. This could be because the smallest class in the training set contained less than 50 samples. Increasing $K$ beyond that would likely result in including too many samples from other classes in the averaging over process and thus reduce correlation and usefulness of the imputation in the classification downstream. This also implies that in case one trains the imputer on data sets containing more samples in the smallest class, one can increase $K$ accordingly for best results.

### 4.7.2 Imputed data set is just as useful for classification

In order to further quantify the imputation, test set features of the DNA methylation and miRNA expression were nullified and imputed using the training set and KNN ($K = 50$). The GCN-VCDN model was trained on the train set and then tested on the imputed multi-omics test set. An F1 score of 0.821 was achieved. This is just as much as with the true test set. Imputation of missing omics is just as useful for classification as the true set.

### 4.7.3 Imputation of missing omics improves classification

As another avenue for independent validation of the model, the entire TCGA BRCA multi-omics data set was used for training and the the model so trained was tested on the METABRIC data set. First, the TCGA BRCA data and KNN ($K = 50$) was used to impute the missing omics, DNA methylation and miRNA expression, in the METABRIC data set. An F1 score of 0.74 was achieved on this imputed set, a 4.2% improvement over using just gene expression data. Aside from just the classification accuracy, the imputation of the missing omics also increases interpretability of the model as feature importance scores for the missing omics can also be obtained.

## 4.8 Model performs multi-tissue integration as well

The GCN-VCDN was trained to integrate multi-tissue (cerebellum, primary visual cortex, and prefrontal cortex) data and classify samples as being normal, Huntington's disease, or Alzheimer's disease samples. The model achieved F1 scores of $0.98 \pm 0.004$ on the training set and $0.91 \pm 0.010$ on the test set.

Integration of multi-tissue data is similar to integration of multi-omics data. Both of them are data with multi-modality and our GCN-VCDN model is robust to integrate any kind of multi-modality data set and classify the samples. Similar analysis of feature importance can be performed on the multi-tissue data too.

# CHAPTER 5

# CONCLUSION

A framework has been created to integrate multi-modality data. Pairwise interaction features are generated as a feature engineering step to learn intra-modality and inter-modality relations. GCN-VCDN model utilizes multi-modality data and inter-sample relations to classify samples. SHAP and LIME provide feature importance scores to enable personalized medicine and predict biomarkers.

# CHAPTER 6

# FUTURE WORK

## 6.1 Prize-collecting Steiner Forest

Consider heterogeneous multi-layered network where different omics data types for separate layers, individual omics features (like genes, proteins, miRNA) form nodes, and inter-feature relations (like co-expression, pairwise feature importance) for the edges. Such a network is a superset of all causative disease pathways (Lee *et al.*, 2020; Hammoud and Kramer, 2020). Tuncbag *et al.* (2016) solves the prize collecting Steiner forest problem on a protein-protein interaction network to identify putative underlying molecular pathways. Similarly, a heterogeneous multi-layered network can be constructed using the multi-omics data and then use the feature importance scores obtained from LIME/SHAP to formulate a prize collecting Steiner forest problem solving which would yield minimal sub-networks/pathways that would explain underlying molecular pathways involved in the disease. (Tuncbag *et al.*, 2013; Gitter *et al.*, 2013)

## 6.2 Simple GCNs

Wu *et al.* (2019) opine that graph convolutional networks have unnecessary complexity and redundant computation. They present Simple GCN, a linear model constructed by successively removing non-linearity between and collapsing individuals layers inside a conventional graph convolutional network. They also demonstrate that this model is just as good at classifying as conventional GCNs while being orders of magnitude faster. Hence, I would like to experiment with implementing a simple GCN in place of the conventional GCN to make the model more efficient.

## 6.3 Graph Transformers

With the introduction of Attention mechanism and Transformer models, there have been many different models like Graph Attention Networks and Graph Transformers that have replaced graph convolutional networks for certain tasks. In the near future, I want to experiment with implementing these model in place of the conventional GCN to classify better. (Vaswani *et al.*, 2017; Velicković *et al.*, 2018; Hu *et al.*, 2020)

# CHAPTER 7

# DISCUSSION

## 7.1 Challenges

We have presented and demonstrated the use of a framework for the integration and imputation of high throughput multi-modality data in the biological context. With the improved performance and interpretability, it is arguable that domain knowledge is a strong aid. Creating better models for multi-omics/multi-tissue integration and imputation requires specific knowledge in data science as well as bioinformatics. Capitalizing on knowledge of biology and bioinformatics, the relevant domain knowledge, has helped this data science and disease modelling inter-disciplinary project.

While the classification of disease or disease subtype appears trivial, modelling the data for the classification allows the model to learn useful relationships between the data and disease and among different modalities of the data itself. Translating this knowledge from the model to human-readable form is a challenge.

We have used SHAP and LIME for better interpretation of the data in a global sense to predict biomarkers as well as in a local sense to enable personalized medicine. A complete understanding of the disease is still elusive. We propose to use heterogeneous multi-layered networks and formulate the inference of molecular pathways related to the disease as a prize-collecting Steiner forest problem. Feature importance scores obtained from SHAP and LIME will aid in inferring these molecular pathways.

## 7.2 Code Availability

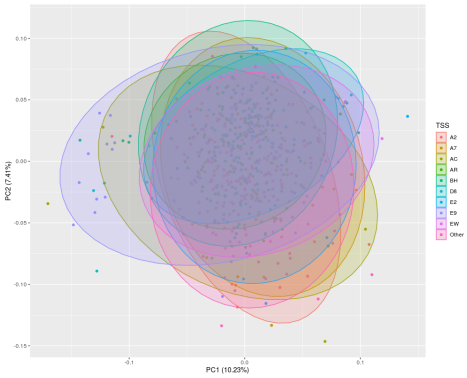
All of the code used to generate the results, figures and tables, is available on my public repository on GitHub.

## 7.3 Reproducibility of Code

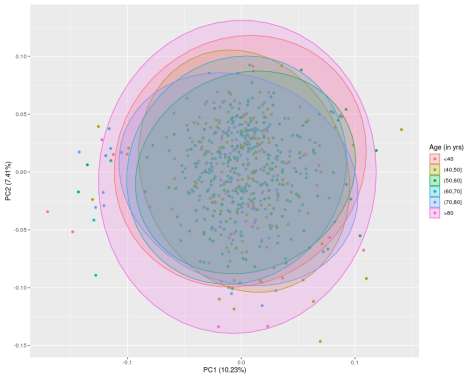
In order to ensure reproducibility, the code is made available in a complete and well-documented form. Pre-processed data as well as R code for pre-processing the data obtained from the sources mentioned in chapter 2 are present in a separate folder named R. The GCN-VCDN model and all the accessory functions needed to run the model are present in the home directory. Multiple Jupyter Noteboooks present in the repository clearly demonstrate how to use the code. Logs of previous runs have also been stored to document the hyperparameters used to obtain certain results.

# APPENDIX A

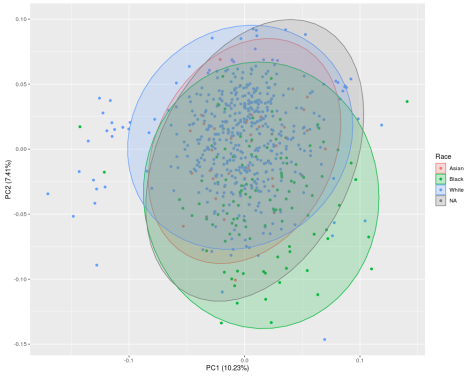
# SUPPLEMENTARY FIGURES



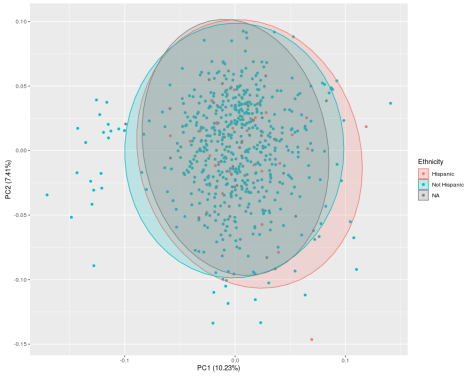

(a) Tissue Source Sites



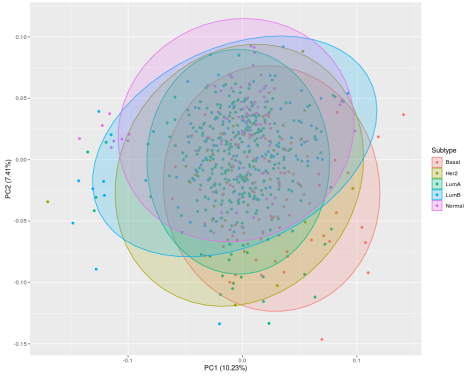

(b) Age




(c) Race



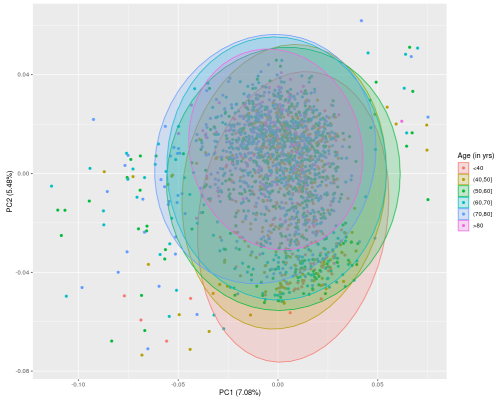

(d) Ethnicity




(e) Cancer Subtypes

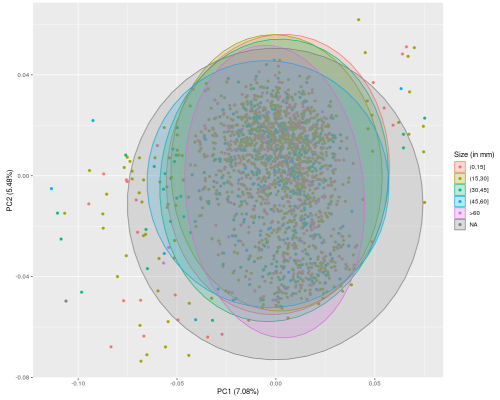
Fig. A.1: **Data**: TCGA BRCA samples plotted based on principal components of scaled data and coloured based on potential covariates. No well-defined clusters are observed.


(a) Tissue Source Sites

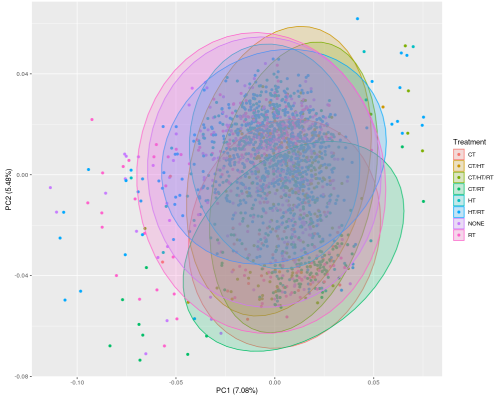



(b) Age

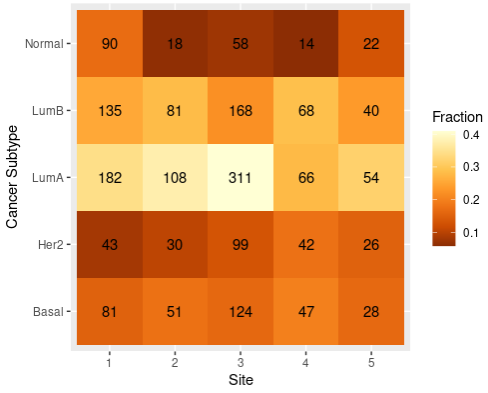

(c) Tissue Cellularity




(d) Size (in mm)




(e) Treatment Administered




(f) Heatmap of TSS v Cancer subtype

Fig. A.2: **Data**: METABRIC samples plotted based on principal components of scaled data and coloured based on potential covariates.

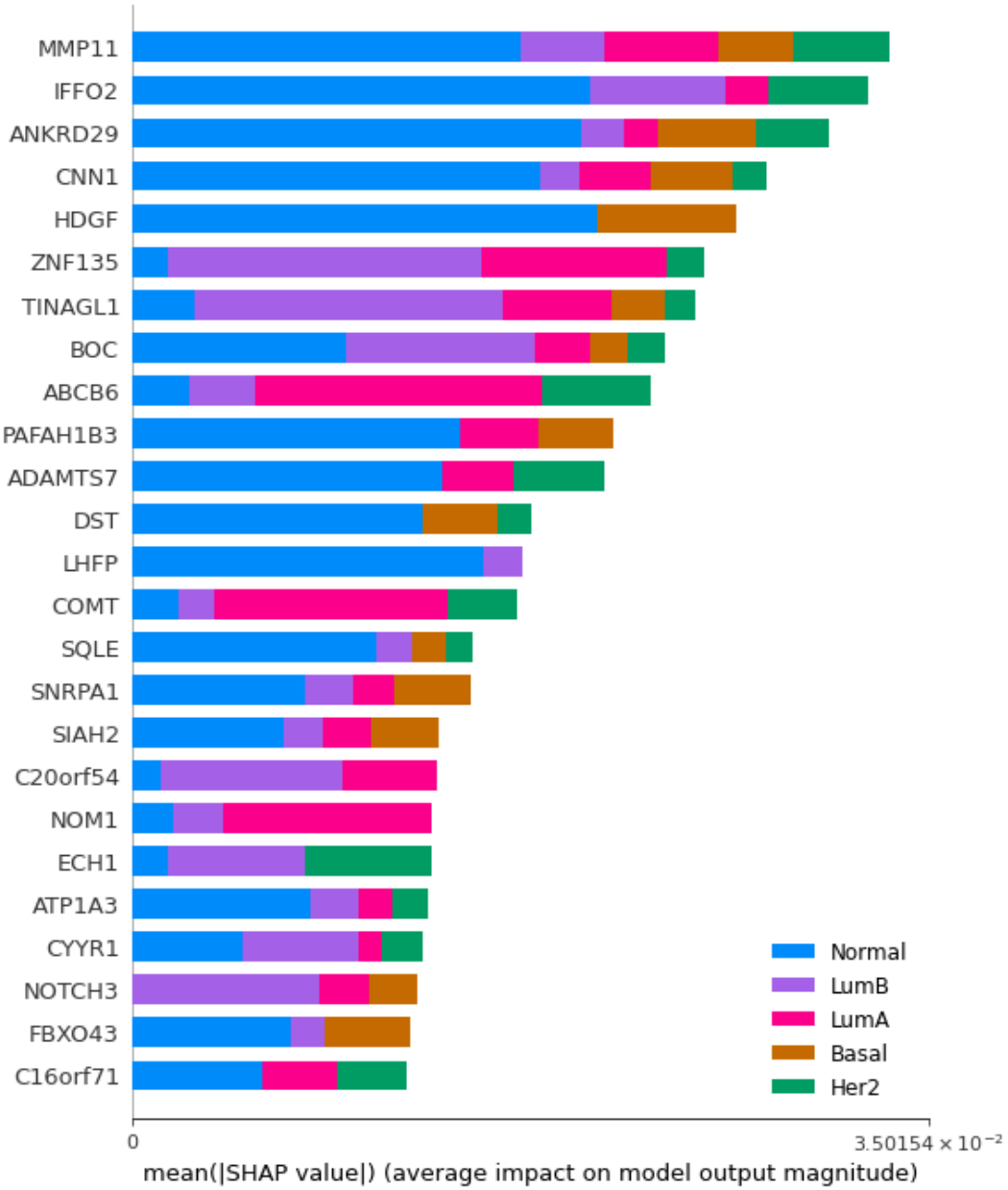
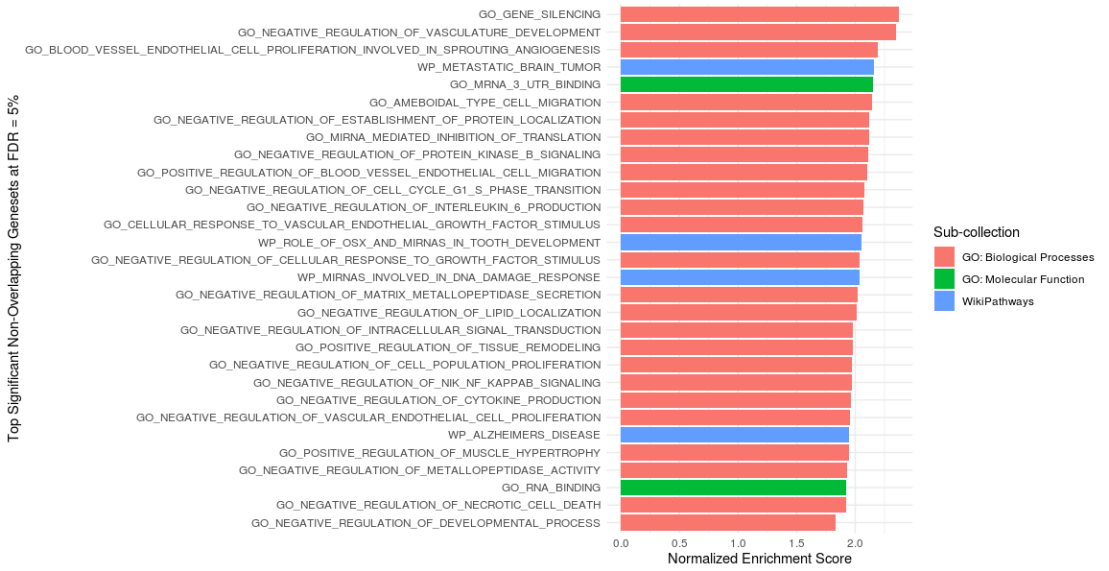
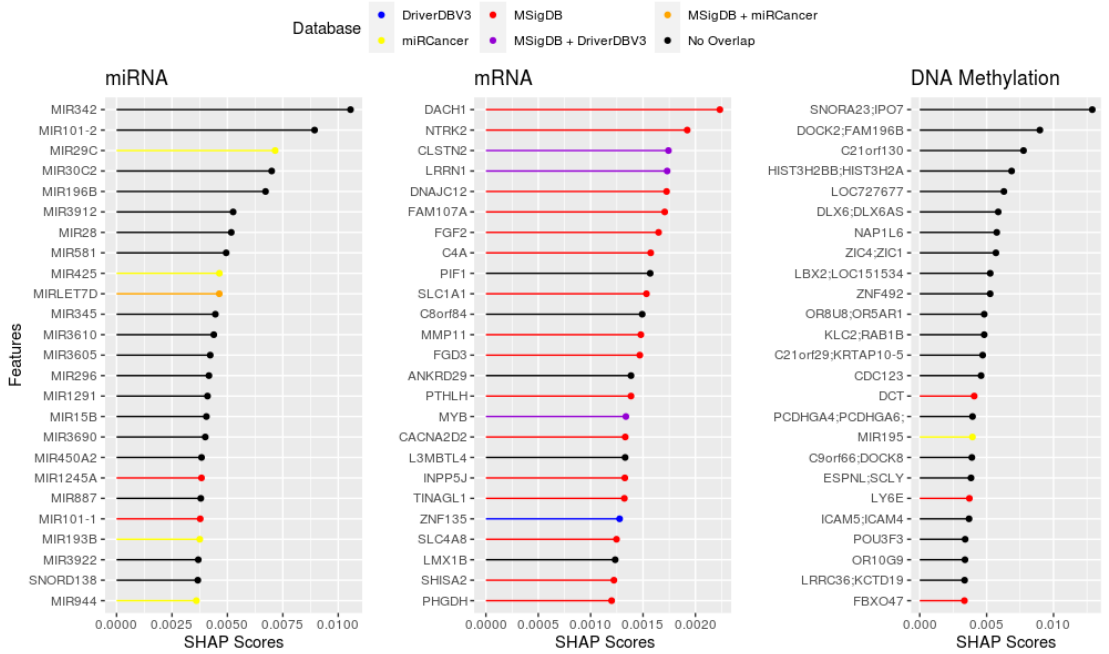



Fig. A.3: **Results: Local explanations enable personalized medicine.** SHAP summary plot of top features and their contribution toward the different breast cancer subtypes. This summary plot was calculated over five patients: TCGA-D8-A1XU-01, TCGA-D8-A1XV-01, TCGA-EW-A1P1-01, TCGA-BH-A1EV-11, TCGA-BH-A1FJ-11. Similar plots can be made over a single patient for local interpretability and over all patients for global interpretability.




(a) Top gene sets enriched for using the feature rank list from SHAP on the TCGA BRCA data set. Many of the enriched gene sets like those involved in regulation of vasculature development, sprouting angiogenesis, metastasis, vascular endothelial growth factor stimulus, DNA damage, and more, clearly indicate toward tumour and metastasis.




(b) Top features selected by SHAP and their membership in various gene set collections. For more information on the databases/gene set collections used, refer section 3.6.

Fig. A.4: **Results: Global interpretability yields biomarkers (SHAP).** (a) Top enriched gene sets are cancer associated. (b) Model predicts novel biomarkers.

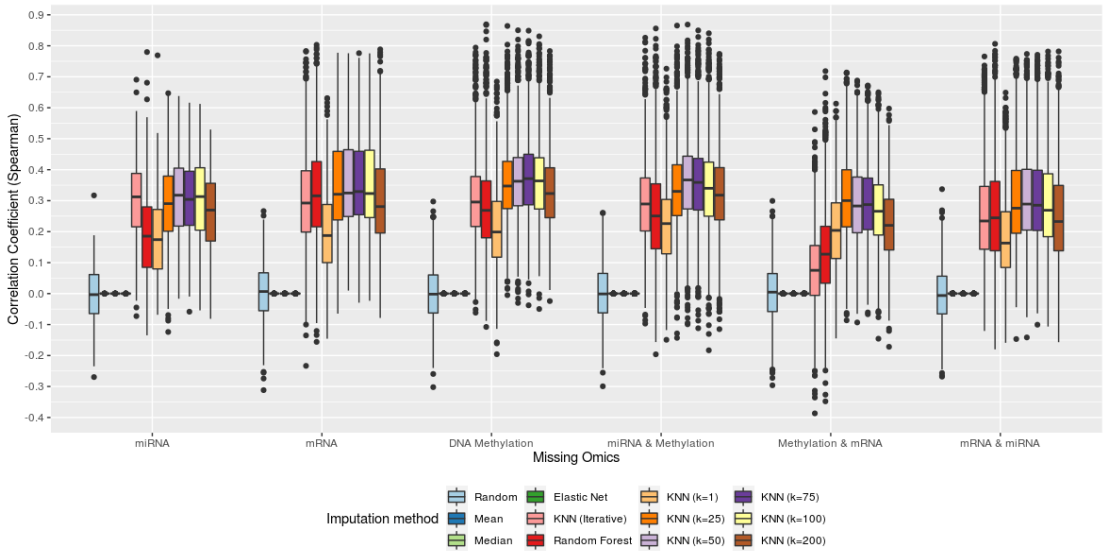



Fig. A.5: **Results: Omics imputation adds predictive power.** Distribution of spearman correlation coefficient of the features imputed by all tested methods with the true values.

# APPENDIX B

# SUPPLEMENTARY TABLES

| Method | Training Set F1 Score | Test Set F1 Score |
|---|---|---|
| LASSO | $0.978 \pm 0.006$ | $0.812 \pm 0$ |
| Ridge Regression | $0.932 \pm 0.002$ | $0.778 \pm 0$ |
| SVM | $1 \pm 0$ | $0.831 \pm 0$ |
| RF | $1 \pm 0$ | $0.785 \pm 0.006$ |
| PLSDA | $0.701 \pm 0$ | $0.72 \pm 0$ |
| SPLSDA | $0.701 \pm 0$ | $0.72 \pm 0$ |
| MORONET | $0.928 \pm 0.002$ | $0.802 \pm 0.007$ |
| **GCN + VCDN** | $0.992 \pm 0.002$ | $\mathbf{0.839 \pm 0.011}$ |

Table B.1: Comparison of different methods' cancer subtype classification on the TCGA BRCA data set. (LASSO = Least Absolute Shrinkage and Selection Operator, SVM = Support Vector Machine, RF = Random Forest, PLSDA = Partial Least Squares Discriminant Analysis (Singh *et al.*, 2019), SPLSDA = Sparse Partial Least Squares Discriminant Analysis, MORONET = Multi-Omics gRaph cOnvolutional NETworks (Wang *et al.*, 2020))

| Data Used | Training Set F1 Score | Test Set F1 Score |
|---|---|---|
| mRNA Expression Only | $0.906 \pm 0.093$ | $0.789 \pm 0.099$ |
| DNA Methylation Only | $0.974 \pm 0.013$ | $0.776 \pm 0.017$ |
| miRNA Expression Only | $0.984 \pm 0.011$ | $0.807 \pm 0.009$ |
| mRNA + meth + miRNA | $0.992 \pm 0.002$ | $\mathbf{0.839 \pm 0.011}$ |
| Primary Omics + Intra-modality | $0.992 \pm 0.002$ | $0.864 \pm 0.009$ |
| Primary Omics + Inter-modality | $0.991 \pm 0.005$ | $0.866 \pm 0.009$ |
| **Primary Omics + All Interaction** | $0.992 \pm 0.002$ | $\mathbf{0.866 \pm 0.009}$ |

Table B.2: Comparison of cancer subtype classification on the TCGA BRCA data set based on different omics used. Here, Primary Omics refers to the three individual omics: mRNA Expression, DNA Methylation, and miRNA Expression, Intra-modality refers to mRNA X mRNA, meth X meth, and miRNA X miRNA interactions, Inter-modality refers to mRNA X meth, meth X miRNA, and miRNA X mRNA interactions, and All Interaction refers to Intra-modality and Inter-modality interactions. (meth = DNA Methylation)

# REFERENCES


1. **Blagus, R.** and **L. Lusa** (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **14**(1), doi:10.1186/1471-2105-14-106.

2. **Chawla, N. V.**, **K. W. Bowyer**, **L. O. Hall**, and **W. P. Kegelmeyer** (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357, doi:10.1613/jair.953.

3. **Cittelly, D. M.**, **P. M. Das**, **N. S. Spoelstra**, **S. M. Edgerton**, **J. K. Richer**, **A. D. Thor**, and **F. E. Jones** (2010). Downregulation of miR-342 is associated with tamoxifen resistant breast tumors. *Molecular Cancer*, **9**(1), 317, doi:10.1186/1476-4598-9-317.

4. **Crippa, E.**, **L. Lusa**, **L. D. Cecco**, **E. Marchesi**, **G. A. Calin**, **P. Radice**, **S. Manoukian**, **B. Peissel**, **M. G. Daidone**, **M. Gariboldi**, and **M. A. Pierotti** (2014). miR-342 regulates BRCA1 expression through modulation of ID4 in breast cancer. *PLoS ONE*, **9**(1), e87039, doi:10.1371/journal.pone.0087039.

5. **Cui, L.**, **K. Nakano**, **S. Obchoei**, **K. Setoguchi**, **M. Matsumoto**, **T. Yamamoto**, **S. Obika**, **K. Shimada**, and **N. Hiraoka** (2017). Small nucleolar noncoding RNA SNORA23, up-regulated in human pancreatic ductal adenocarcinoma, regulates expression of spectrin repeat-containing nuclear envelope 2 to promote growth and metastasis of xenograft tumors in mice. *Gastroenterology*, **153**(1), 292–306.e2, doi:10.1053/j.gastro.2017.03.050.

6. **Curtis, C.**, , **S. P. Shah**, **S.-F. Chin**, **G. Turashvili**, **O. M. Rueda**, **M. J. Dunning**, **D. Speed**, **A. G. Lynch**, **S. Samarajiwa**, **Y. Yuan**, **S. Gräf**, **G. Ha**, **G. Haffari**, **A. Bashashati**, **R. Russell**, **S. McKinney**, **A. Langerød**, **A. Green**, **E. Provenzano**, **G. Wishart**, **S. Pinder**, **P. Watson**, **F. Markowetz**, **L. Murphy**, **I. Ellis**, **A. Purushotham**, **A.-L. Børresen-Dale**, **J. D. Brenton**, **S. Tavaré**, **C. Caldas**, and **S. Aparicio** (2012). The genomic and transcriptomic architecture of 2, 000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352, doi:10.1038/nature10983.

7. **Duvenaud, D. K.**, **D. Maclaurin**, **J. Iparraguirre**, **R. Bombarell**, **T. Hirzel**, **A. Aspuru-Guzik**, and **R. P. Adams** (2015). Convolutional networks on graphs for learning molecular fingerprints. *In* **C. Cortes**, **N. D. Lawrence**, **D. D. Lee**, **M. Sugiyama**, and **R. Garnett** (eds.), *Advances in Neural Information Processing Systems 28*, 2224–2232. Curran Associates, Inc. URL http://papers.nips.cc/paper/5954-convolutional-networks-on-graphs-for-learning-molecular-fingerpri
pdf.

8. **Gitter, A.**, **A. Braunstein**, **A. Pagnani**, **C. Bakdassi**, **C. Borgs**, **J. Chayes**, **R. Zecchina**, and **E. Fraenkel** (2013). Sharing information to reconstruct patient-specific pathways in heterogeneous diseases. *In Biocomputing 2014*. World Scientific, doi:10.1142/9789814583220_0005.

9. **Gligorijević, V.** and **N. Przulj** (2015). Methods for biological data integration: perspectives and challenges. *Journal of The Royal Society Interface*, **12**(112), 20150571, doi:10.1098/rsif.2015.0571.

10. **Golomb, L.**, **D. R. Bublik**, **S. Wilder**, **R. Nevo**, **V. Kiss**, **K. Grabusic**, **S. Volarevic**, and **M. Oren** (2012). Importin 7 and exportin 1 link c-myc and p53 to regulation of ribosomal biogenesis. *Molecular Cell*, **45**(2), 222–232, doi:10.1016/j.molcel.2011.11.022.

11. **Hammoud, Z.** and **F. Kramer** (2020). Multilayer networks: aspects, implementations, and application in biomedicine. *Big Data Analytics*, **5**(1), doi:10.1186/s41044-020-00046-0.

12. **Hu, Z.**, **Y. Dong**, **K. Wang**, and **Y. Sun** (2020). Heterogeneous graph transformer. *In Proceedings of The Web Conference 2020*. ACM, doi:10.1145/3366423.3380027.

13. **Jeevannavar, A.** (2020). *Cross-omic Deep Learning Networks for Identifying Disease Biomarkers and Pathways: A Preliminary Report*. Master's thesis, Department of Biotechnology, IIT-Madras, Chennai – 600036. URL https://www.jeevannavar.com/assets/files/DDP_Preliminary_Report.pdf.

14. **Korotkevich, G.**, **V. Sukhov**, **N. Budin**, **B. Shpak**, **M. N. Artyomov**, and **A. Sergushichev** (2016). Fast gene set enrichment analysis. doi:10.1101/060012.

15. **Lee, A. Y.**, **S. Kim**, **S. Lee**, **H.-L. Jiang**, **S.-B. Kim**, **S.-H. Hong**, and **M.-H. Cho** (2017). Knockdown of importin 7 inhibits lung tumorigenesis in k-rasLA1 lung cancer mice. *Anticancer Research*, **37**(5), 2181–2386, doi:10.21873/anticanres.11576.

16. **Lee, B.**, **S. Zhang**, **A. Poleksic**, and **L. Xie** (2020). Heterogeneous multi-layered network model for omics data integration and analysis. *Frontiers in Genetics*, **10**, doi:10.3389/fgene.2019.01381.

17. **Liberzon, A.**, **C. Birger**, **H. Thorvaldsdóttir**, **M. Ghandi**, **J. P. Mesirov**, and **P. Tamayo** (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, **1**(6), 417–425, doi:10.1016/j.cels.2015.12.004.

18. **Liberzon, A.**, **A. Subramanian**, **R. Pinchback**, **H. Thorvaldsdottir**, **P. Tamayo**, and **J. P. Mesirov** (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**(12), 1739–1740, doi:10.1093/bioinformatics/btr260.

19. **Lindholm, E. M.**, **S.-K. Leivonen**, **E. Undlien**, **D. Nebdal**, **A. Git**, **C. Caldas**, **A.-L. Børresen-Dale**, and **K. Kleivi** (2019). miR-342-5p as a potential regulator of HER2 breast cancer cell growth. *MicroRNA*, **8**(2), 155–165, doi:10.2174/2211536608666181206124922.

20. **Liu, C.**, **H. Xing**, **X. Luo**, and **Y. Wang** (2020*a*). MicroRNA-342 targets cofilin 1 to suppress the growth, migration and invasion of human breast cancer cells. *Archives of Biochemistry and Biophysics*, **687**, 108385, doi:10.1016/j.abb.2020.108385.

21. **Liu, S.-H.**, **P.-C. Shen**, **C.-Y. Chen**, **A.-N. Hsu**, **Y.-C. Cho**, **Y.-L. Lai**, **F.-H. Chen**, **C.-Y. Li**, **S.-C. Wang**, **M. Chen**, **I.-F. Chung**, and **W.-C. Cheng** (2019). DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Research*, doi:10.1093/nar/gkz964.

22. **Liu, Y.**, **H. Ruan**, **S. Li**, **Y. Ye**, **W. Hong**, **J. Gong**, **Z. Zhang**, **Y. Jing**, **X. Zhang**, **L. Diao**, and **L. Han** (2020*b*). The genetic and pharmacogenomic landscape of snoRNAs in human cancer. *Molecular Cancer*, **19**(1), doi:10.1186/s12943-020-01228-z.

23. **Lundberg, S. M.** and **S.-I. Lee** (2017). A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems*.

24. **Manessi, F.**, **A. Rozza**, and **M. Manzo** (2020). Dynamic graph convolutional networks. *Pattern Recognition*, **97**, 107000, doi:10.1016/j.patcog.2019.107000.

25. **Mark Dunning, A. L.** (2017). illuminahumanv3.db. doi:10.18129/B9.BIOC.ILLUMINAHUMANV3.DB.

26. **Network, T. C. G. A.** (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70, doi:10.1038/nature11412.

27. **Paszke, A.**, **S. Gross**, **F. Massa**, **A. Lerer**, **J. Bradbury**, **G. Chanan**, **T. Killeen**, **Z. Lin**, **N. Gimelshein**, **L. Antiga**, **A. Desmaison**, **A. Kopf**, **E. Yang**, **Z. DeVito**, **M. Raison**, **A. Tejani**, **S. Chilamkurthy**, **B. Steiner**, **L. Fang**, **J. Bai**, and **S. Chintala** (2019). Pytorch: An imperative style, high-performance deep learning library. *In* **H. Wallach**, **H. Larochelle**, **A. Beygelzimer**, **F. d'Alché-Buc**, **E. Fox**, and **R. Garnett** (eds.), *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-librar pdf.

28. **Pedregosa, F.**, **G. Varoquaux**, **A. Gramfort**, **V. Michel**, **B. Thirion**, **O. Grisel**, **M. Blondel**, **P. Prettenhofer**, **R. Weiss**, **V. Dubourg**, **J. Vanderplas**, **A. Passos**, **D. Cournapeau**, **M. Brucher**, **M. Perrot**, and **E. Duchesnay** (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

29. **Piñero, J.**, **J. M. Ramírez-Anguita**, **J. Saüch-Pitarch**, **F. Ronzano**, **E. Centeno**, **F. Sanz**, and **L. I. Furlong** (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, doi:10.1093/nar/gkz1021.

30. **Popov, V. M.**, **J. Zhou**, **L. A. Shirley**, **J. Quong**, **W.-S. Yeow**, **J. A. Wright**, **K. Wu**, **H. Rui**, **R. K. Vadlamudi**, **J. Jiang**, *et al.* (2009). The cell fate determination factor dach1 is expressed in estrogen receptor-$\alpha$–positive breast cancer and represses estrogen receptor-$\alpha$ signaling. *Cancer research*, **69**(14), 5752–5760.

31. **Powe, D. G.**, **G. K. R. Dhondalay**, **C. Lemetre**, **T. Allen**, **H. O. Habashy**, **I. O. Ellis**, **R. Rees**, and **G. R. Ball** (2014). Dach1: its role as a classifier of long term good prognosis in luminal breast cancer. *PloS one*, **9**(1), e84428.

32. **R Core Team** (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

33. **Ribeiro, M. T.**, **S. Singh**, and **C. Guestrin** (2016). "why should I trust you?": Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*.

34. **RStudio Team** (2018). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA. URL http://www.rstudio.com/.

35. **Seldin, M. M.** and **A. J. Lusis** (2019). Systems-based approaches for investigation of inter-tissue communication. *Journal of Lipid Research*, **60**(3), 450–455, doi:10.1194/jlr.s090316.

36. **Silva, T. C.**, **A. Colaprico**, **C. Olsen**, **F. D'Angelo**, **G. Bontempi**, **M. Ceccarelli**, and **H. Noushmehr** (2016). TCGA workflow: Analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Research*, **5**, 1542, doi:10.12688/f1000research.8923.2.

37. **Singh, A.**, **C. P. Shannon**, **B. Gautier**, **F. Rohart**, **M. Vacher**, **S. J. Tebbutt**, and **K.-A. L. Cao** (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, **35**(17), 3055–3062, doi:10.1093/bioinformatics/bty1054.

38. **Smith, E. R.**, **K. Q. Cai**, **J. L. Smedberg**, **M. M. Ribeiro**, **M. E. Rula**, **C. Slater**, **A. K. Godwin**, and **X.-X. Xu** (2010). Nuclear entry of activated MAPK is restricted in primary ovarian and mammary epithelial cells. *PLoS ONE*, **5**(2), e9295, doi:10.1371/journal.pone.0009295.

39. **Subramanian, A.**, **P. Tamayo**, **V. K. Mootha**, **S. Mukherjee**, **B. L. Ebert**, **M. A. Gillette**, **A. Paulovich**, **S. L. Pomeroy**, **T. R. Golub**, **E. S. Lander**, and **J. P. Mesirov** (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. ISSN 0027-8424, doi:10.1073/pnas.0506580102.

40. **Szczyrba, J.**, **E. Nolte**, **M. Hart**, **C. Döll**, **S. Wach**, **H. Taubert**, **B. Keck**, **E. Kremmer**, **R. Stöhr**, **A. Hartmann**, **W. Wieland**, **B. Wullich**, and **F. A. Grässer** (2012). Identification of ZNF217, hnRNP-k, VEGF-a and IPO7 as targets for microRNAs that are downregulated in prostate carcinoma. *International Journal of Cancer*, **132**(4), 775–784, doi:10.1002/ijc.27731.

41. **Tuncbag, N.**, **A. Braunstein**, **A. Pagnani**, **S.-S. C. Huang**, **J. Chayes**, **C. Borgs**, **R. Zecchina**, and **E. Fraenkel** (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of Computational Biology*, **20**(2), 124–136, doi:10.1089/cmb.2012.0092.

42. **Tuncbag, N.**, **S. J. C. Gosline**, **A. Kedaigle**, **A. R. Soltis**, **A. Gitter**, and **E. Fraenkel** (2016). Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLOS Computational Biology*, **12**(4), e1004879, doi:10.1371/journal.pcbi.1004879.

43. **Vaswani, A.**, **N. Shazeer**, **N. Parmar**, **J. Uszkoreit**, **L. Jones**, **A. N. Gomez**, **L. Kaiser**, and **I. Polosukhin** (2017). Attention is all you need.

44. **Velicković, P.**, **G. Cucurull**, **A. Casanova**, **A. Romero**, **P. Liò**, and **Y. Bengio** (2018). Graph attention networks.

45. **Wang, L.**, **Z. Ding**, **Z. Tao**, **Y. Liu**, and **Y. Fu** (2019). Generative multi-view human action recognition. *In 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:10.1109/ICCV.2019.00631.

46. **Wang, T.**, **W. Shao**, **Z. Huang**, **H. Tang**, **J. Zhang**, **Z. Ding**, and **K. Huang** (2020). MORONET: Multi-omics integration via graph convolutional networks for biomedical data classification. doi:10.1101/2020.07.02.184705.

47. **Weinstein, J. N.**, , **E. A. Collisson**, **G. B. Mills**, **K. R. M. Shaw**, **B. A. Ozenberger**, **K. Ellrott**, **I. Shmulevich**, **C. Sander**, and **J. M. Stuart** (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, **45**(10), 1113–1120, doi:10.1038/ng.2764.

48. **Wickham, H.**, **M. Averick**, **J. Bryan**, **W. Chang**, **L. D. McGowan**, **R. François**, **G. Grolemund**, **A. Hayes**, **L. Henry**, **J. Hester**, **M. Kuhn**, **T. L. Pedersen**, **E. Miller**, **S. M. Bache**, **K. Müller**, **J. Ooms**, **D. Robinson**, **D. P. Seidel**, **V. Spinu**, **K. Takahashi**, **D. Vaughan**, **C. Wilke**, **K. Woo**, and **H. Yutani** (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686, doi:10.21105/joss.01686.

49. **Wu, F.**, **A. Souza**, **T. Zhang**, **C. Fifty**, **T. Yu**, and **K. Weinberger** (2019). Simplifying graph convolutional networks. *In* **K. Chaudhuri** and **R. Salakhutdinov** (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*. PMLR. URL http://proceedings.mlr.press/v97/wu19e.html.

50. **Xie, B.**, **Q. Ding**, **H. Han**, and **D. Wu** (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics*, **29**(5), 638–644, doi:10.1093/bioinformatics/btt014.

51. **Xu, H.**, **S. Yu**, **X. Yuan**, **J. Xiong**, **D. Kuang**, **R. G. Pestell**, and **K. Wu** (2017). DACH1 suppresses breast cancer as a negative regulator of CD44. *Scientific Reports*, **7**(1), doi:10.1038/s41598-017-04709-2.

52. **Zhang, B.**, **C. Gaiteri**, **L.-G. Bodea**, **Z. Wang**, **J. McElwee**, **A. A. Podtelezhnikov**, **C. Zhang**, **T. Xie**, **L. Tran**, **R. Dobrin**, **E. Fluder**, **B. Clurman**, **S. Melquist**, **M. Narayanan**, **C. Suver**, **H. Shah**, **M. Mahajan**, **T. Gillis**, **J. Mysore**, **M. E. MacDonald**, **J. R. Lamb**, **D. A. Bennett**, **C. Molony**, **D. J. Stone**, **V. Gudnason**, **A. J. Myers**, **E. E. Schadt**, **H. Neumann**, **J. Zhu**, and **V. Emilsson** (2013). Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, **153**(3), 707–720, doi:10.1016/j.cell.2013.03.030.

53. **Zhang, Q.**, **Y. Yuan**, **J. Cui**, **T. Xiao**, and **D. Jiang** (2015). Mir-217 promotes tumor proliferation in breast cancer via targeting dach1. *Journal of Cancer*, **6**(2), 184.

54. **Zhao, F.**, **M. Wang**, **S. Li**, **X. Bai**, **H. Bi**, **Y. Liu**, **X. Ao**, **Z. Jia**, and **H. Wu** (2015). Dach1 inhibits snai1-mediated epithelial–mesenchymal transition and represses breast carcinoma metastasis. *Oncogenesis*, **4**(3), e143–e143.