IIT MADRAS

Dual Degree Project Report 1 - BT5802

Cross-omic Deep Learning Networks for Identifying Disease Biomarkers and Pathways

Aditya Jeevannavar (BS16B001)
Guide : Prof. Manikandan Narayanan
November 4, 2020

1 ABSTRACT

The advent and widespread use of various technologies for the collection of big data in biotechnology have led to a variety of omics data types such as transcriptomics, proteomics, and epigenomics data on the same set of samples. Integrating such multi-omics, including multi-tissue data sets, can provide an unbiased approach to find biomarkers, causative variants and pathways for diseases ranging from Alzheimer's to Cancer.

Wang et al.[34] introduced a novel neural network based method named Multi-Omics gRaph cOnvolutional NETworks (MORONET) to integrate mRNA, miRNA, and methylation data for biomedical classification and biomarker identification. Tuncbag et al.[30] implemented the prize-collecting Steiner forest algorithm on a protein-protein interaction network to identify putative molecular pathways connecting a set of query proteins (such as protein biomarkers of a disease).

However, scholars in the field have not yet comprehensively addressed how interactions among features both within and across different omics data types can be harnessed to improve classification performance, or how to use the learned features (or feature-feature interactions) in the neural network to understand disease mechanisms. Additionally, many studies are unable to generate all the different omics data due to limitations of funds or equipment.

The objectives of the proposed project are to improve on the performance and interpretability of the graph convolutional network approach used in MORONET by using pairwise interactions of the features, to develop a framework to impute missing omics data types, and to predict biomarkers and causative pathways.

The following methods will be used to fulfil the objectives mentioned above:

- **Deep Learning methods** - To implement modified graph convolutional networks, perform hyperparameter tuning, and retrieve biomarkers using a feature importance measure.
- **Machine Learning methods** - To implement a framework for imputing a missing omics data type (tissue data set) using other available omic data types (tissue data sets).
- **Network Analysis** - To implement the prize-collecting Steiner forest algorithm on a heterogeneous multi layered network for improving the interpretability of the neural network.

2 INTRODUCTION

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, proteomics the study of the proteome, and so on.

The first whole genome of a bacterial strain, *Haemophilus influenzae*, was sequenced in 1995.[8] The first whole genome of a human was sequenced in 2007. [18] And subsequent genomic data sets sequenced have been used to elucidate gene functions, metabolic pathways, biomolecular networks, the basic functioning of an organism at the molecular level. But just the knowledge of the genome was not enough to decipher the complete working of an organism's molecular biology. More information was required. Thus, data began being collected on the gene expression and the proteins. Even this was not enough. Data began being collected on the epigenome and the metabolome. Thus, over the past decade, researchers have been trying to understand the biological system, its structure and functions, through the various omes and the corresponding omics. Apart from the now common genome, transcriptome, and proteome, there exists comprehensive knowledge in epigenome, glyome, lipidome, metabolome, phenome, etc. and their associated "omics" technologies and data analysis methods.

With the current high throughput nature of the "omics" technologies, researchers are able to collect several omics data sets on the same experimental samples, i.e., researchers are able to collect genomic, gene expression, and other information from the same set of sample tissues. This has led to the advent of what is termed "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.

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. [10]

There exist many tools that integrate multi-omics data, but challenges remain. The tools vary either in the method they use to integrate the data, for example, canonical correlation analysis or machine learning, or in the purpose of the integration, for example, for the diagnosis/prognosis of a disease or for the inference of a mechanistic network of molecular interactions, or whether they integrate the omics data early or late in the pipeline. There are also methods which overlay all of the omics data sets onto a single layer, like a genomic or metabolic network, and perform the analysis, as opposed to some methods considering the different omics as independent or heterogeneous data. There are many reviews on multi-omics integration methods and, consequently, many different ways and labels under which the methods have been categorized. For example, Nguyen and Wang [22] categorize multi-omics methods as factorization-based or alignment-based multi-view learning methods.

The existence of so many tools is also an indication that no universally applicable tool is available yet. Different tools and methods have been necessary for different types of data sets and endpoints. There are very few tools that can infer cross-omics relations for any given kind of omics input.

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[26] 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. This framework and other similar approaches can be used to integrate quantitative proteomics data using tissue-specific proteomics data to make cross-tissue predictions or by pairing tissue gene expression and proteomics correlations. Since the proteome is dependent on the genome, epigenome, and the other omes, those individual omes can also be integrated across the tissues. The integration of multi-omics and multi-tissue data are abundantly similar in that they both integrate the data across different phenotypes such as disease states or categories, for example Alzheimer's affected or normal control. They differ in what they integrate, i.e., multi-omics

integration involves integrating different omics data on the same tissue and multi-tissue integration involves integrating different tissue data on the same omics. Although they are very similar, there isn't any tool that performs both of the analyses.

Collecting multi-omics data is expensive. There are many cohorts and 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 is that integration models/methods are unable learn all the features from a limited set of samples, and therefore imputing missing data can increase the performance of the model/method. 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 will, therefore, implement an imputation framework before multi-omics (or multi-tissue) data enters the integration pipeline.

Critics of the systems biology field opine that multi-tissue multi-omics is not an independent research discipline because it is not hypothesis-driven. They undervalue the multi-omics approach as compared to traditional studies where one begins with a prior hypothesis focusing on a particular gene or protein or pathway.[39] But such a traditional approach has been insufficient to deal with complex diseases like cancer and concepts like personalized medicine. Multi-omics integration methods, like the one we present here, are essential exploratory analyses to elucidate inter-omics and inter-tissue interactions, to find biomarkers, and to make personalized medicine a reality.

We propose improvements to Wang et. al.'s Graph Convolutional Network based MORONET model[34] so that it can learn cross-omics relations better, irrespective of the omics data type, and thus improve the performance of the model. The proposed model will additionally be able to perform multi-tissue integration and omics imputation. Preliminary results show an improvement in performance over the MORONET model. Future work will include implementing other ways to integrate pairwise interaction of features and obtaining biomarkers from the model.

3 Literature Review

3.1 Multi-Omics Integration Methods

In the early years of the "multi-omics" era, many of the laboratories that generated the data tried to integrate them using basic methods like correlation or co-expression analysis alone. Since then, multi-omics integration methods have come a long way. The multitude of integration methods can be categorized based on when in the process the features from different -omics were brought together (early or late), what the integration level was (conceptual, statistical, or model-based), whether all types of -omics data can be integrated or not (complete or incomplete), and what the outcome of the method is (visualization, label prediction, biomarker prediction, etc.)

Jamil et. al. in their review of the systematic multi-omics integration (MOI) approach [13] categorise the various MOI workflows based on their integration levels. The three levels, in order of increasing complexity, are "element-based" integration, "pathway-based" integration, and "mathematical-based" integration. Each of the levels are further discussed in detail in the review.

- **Element-based Integration** - This level of integration encompasses correlation analysis, clustering analysis, and multivariate analysis. These methods are simple and intuitive.
- **Pathway-based Integration** - This level of integration encompasses pathway mapping and co-expression analysis. Pathway mapping maps the omics data sets to existing metabolic pathway databases like Kyoto Encyclopedia of Genes and Genomes (KEGG). Statistical correlations between omics data sets form the backbone of a co-expression analysis. Then, a weighted network of features is generated.
- **Mathematical-based Integration** - This level of integration encompasses differential analysis and genome-scale analysis. These are the most complex integration methods of all. Their primary goal is to construct a model for systems-level understanding.

Subramanian et. al. in their review of multi-omics data integration, interpretation, and its application [29] classify each of the integration tools or methods into of the following categories: network, Bayesian, fusion, similarity-based, correlation-based, and other multivariate methods. They also mention tools

like similarity network fusion (SNF) and PARADIGM which use a combination of the above mentioned approaches. The review however organises all the tools based on the following broad biological questions that they address:

- "**Disease subtyping and classification based on multi-omics profiles**" - Diseases like cancer are heterogeneous. Identifying the subtypes of the disease or classifying samples into known subgroups to understand the aetiology of the disease and altering the treatment based on the subgroup the patient belongs to is important. There are many tools like iCluster and mixOmics that perform such grouping of samples.
- "**Prediction of biomarkers for various applications including diagnostics and driver genes for diseases**" - Biomarkers are objectively measurable features, either molecular like gene expression level or clinical like body temperature, that are indicative of disease processes. The prediction of biomarkers is vital to diagnosis, prognosis, and treatment of disease. There are fewer methods like iClusterPlus and multi-omics factor analysis (MOFA) that perform such feature selection.
- "**Deriving insights into disease biology**" - This category incorporates a variety of tools that help in diagnosis and treatment by elucidating certain mechanistic details of disease biology. Tools like sparse multi-block partial least squares (sMBPLS) and thresholding singular value decomposition (T-SVD) are useful in deriving insights into disease biology.

A table listing various methods/frameworks of multi-omics integration can be found below. They have been categorised based on method, integration level, output generated, integration stage, and the omics types that can be combined.

Tool	Method	Output	Integration Stage	Omics Type Involved	Reference
Conceptual Integration					
3Omics	Correlation Analysis	Correlation Network and Ontology Enrichment	Late	Transcriptomics, Proteomics, Metabolomics	Kuo et.al[17]
Cytoscape + OmicsAnalyzer	Correlation analysis	Correlation network	Late	Transcriptomics, Proteomics, Metabolomics	Xia et. al. [38]
IMPaLA (Integrated Molecular Pathway Level Analysis)	Pathway enrichment	GO/KEGG enrichment	Late	Transcriptomics, Proteomics, Metabolomics	Kamburov et. al. [15]
PaintOmics	Pathway enrichment	GO/KEGG enrichment	Late	Proteomics, Metabolomics	Garcia-Alcalde et.al. [9]
mixOmics	Multivariate analysis, Clustering, PLS, rCCA, rGCCA, PLS-DA	Biomarkers	Late	Transcriptomics, Proteomics, Metabolomics	Rohart et.al. [24]
MORONET	Graph convolutional networks	Label prediction, Biomarkers	Late	Any three	Wang et. al. [34]
Statistical Integration					
mixOmics - DIABLO	Correlation analysis, Latent components, SVD	Label prediction, Biomarkers	Early	Any	Singh et. al. [28]
OmicsPLS	Two-way orthogonal partial least squares	Cross-omics relations	Early	Any two	el Bouhaddani et. al. [6]
OmicKrigging	Krigging or Gaussian process regression	Label prediction	Early	Any	Wheeler et. al. [35]

Continued on next page

Tool	Method	Output	Integration Stage	Omics Type Involved	Reference
MOTA (Multi-Omics inTegrative Analysis)	Co-expression analysis	Heterogeneous multi layered network, Biomarkers	Early	Any	Fan et. al. [7]
BIDIFAC+	Matrix factorization and decomposition	Cross-omics relations, Sample clusters	Early	Any	Lock et. al. [19]
INMEX (INtegrative Meta-analysis of EXpression data)	Meta-analysis based on p-values, effect sizes etc.	Biomarkers, GO/KEGG enrichment	Late	Trainscriptomics, Metabolomics	Xia et. al. [37]
maui (multi-omics autoencoder integration)	Deep Learning	Sample clusters	Early	Any	Ronen et. al. [25]
D-CCA	Decomposition-based canonical correlation analysis	Sample clusters	Early	Any two	Shu et. al. [27]
ATHENA (Analysis Tool for Heritable and Environmental Network Associations)	Grammatical evolution neural networks and symbolic regression	Label prediction	Early	Any	Holzinger et. al. [12]
Integrative Network Fusion	Similarity Network Fusion	Label prediction, Biomarkers	Late	Any	Chierici et. al. [3]
Spectrum	Fast density-aware spectral clustering	Sample clusters	Late	Any	John et. al. [14]
Model-based Integration					
GIM3E (Gene Inactivation Moderated by Metabolism, Metabolomics and Expression)	Genome-scale metabolic model construction	Metabolic model	Late	Transcriptomics, Metabolomics	Machado et. al. [20]
Ingenuity Pathway Analysis	Proprietary Information	Metabolic model, Causal network, Mechanistic network	-	Transcriptomics, Proteomics, Metabolomics	Krämer et. al. [16]
PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models)	Multivariate analysis	Label prediction, Patient-specific pathway model	Early	Metagenomics, Genomics	Vaske et. al. [31]
Lemon tree	Gibbs sampling	Biomarkers, Label prediction, Mechanistic network	Early	Metagenomics	Bonnet et. al. [1]

Continued on next page

Tool	Method	Output	Integration Stage	Omics Type Involved	Reference
I-BOOST	Elastic net with boosting	Label prediction, Biomarkers	Late	Any, Clinical data	Wong et. al. [36]
COMBI (Compositional Omics Model-Based Integration)	Regression with latent variables	Biomarkers, Sample clusters	Early	Any	Hawinkel et. al. [11]
msPLS (multiset sparse Partial Least Squares)	Penalized partial least squares path modeling	Biomarkers, Cross-omics relations	Early	Any	Csala et. al. [4]

Table 3.1: A list of various tools used for multi-omics integration, the method they employ, and some other assorted characteristics of the tools. The tools have been ordered based on their multi-omics integration level. Conceptual integration tools tend to perform late integration and have fixed input types. Statistical integration tools tend to perform early integration and have unlimited input types. Model-based tools tend to perform early integration but have fixed input types.

3.2 Categorization of Multi-Omics Integration Tools

The multi-omics integration tools can be categorized based on integration level as follows:

- **Conceptual Integration** - In this type of integration, the individual omics data sets are analysed separately and the resulting findings are matched or combined. The data set is not analysed as a whole. While useful, these methods can miss out on valuable cross-omics relations and associations.
- **Statistical Integration** - In this type of integration, the individual omics data sets are analysed together with the aim to find statistical associations between features present in different omics types. Most multi-omics integration methods fall under this category.
- **Model-based Integration** - In this type of integration, the individual omics data sets are combined to form or refine a computational or mathematical model of the system, for example, a metabolic model. A complete model is unobtainable in most situations, bu there exist many tools that approximate most parameters of a model.

The multi-omics integration tools can be categorized based on integration stage as follows:

- **Early Integration** - In this type of integration, the features from different omics types are brought together and analysed together early in the integration pipeline. This allows for better discovery of cross-omics relations and associations.
- **Late Integration** - In this type of integration, the features from different omics types are either brought and analysed together very late in the integration pipeline or not at all. The omics types may be analysed independently and the resulting features or inferences combined. This does not allow for discovery of cross-omics relations and associations.

The multi-omics integration tools can be categorized based on their input as follows:

- **Fixed** - This type of integration tools can integrate only a specific combination of omics types. Many of the early tools belong to this category. For example, 3Omics can only integrate transcriptomics, proteomics, and metabolomics data.
- **Limited** - This type of integration tools can integrate any type of omics types but are limited to how many of the omics they can integrate. For example, OmicsPLS can integrate any and*only two* omics data sets.
- **Unlimited** - This type of integration tools can integrate any type of omics types and are not limited to the number of omics type inputs either. For example, MORONET can can take any number of varied omics types as inputs.

3.3 Meta-analysis

There are also methods that integrate data across cohorts/studies. This is generally called meta-analysis. Combining data across cohorts/studies increases the sample size and aids in the construction of reliable models.

This integration of data across cohorts or studies is also called horizontal integration. Similarly, integration of data across different omics types in the same study or cohort is called vertical integration. A few tools like IMPaLA[15] and mixOmics[24] combine both horizontal and vertical integration within the same tool.

3.4 Deep Learning Based Methods

Deep learning is a subfield of the broader familiy of machine learning that is characterized by artificial neural networks and representation learning. The adjecitve "deep" in deep learning signifies the multiple layers of artificial neurons used in these models. Representational learning or feature learning referes to the ability of the neural networks to discover features or combinations of features in the input data set that are important to the regression or classification task at hand.

For both of these factors to be useful, a large quantity of samples are required in the inputs. But there are very few samples present in multi-omics data sets. Thus, one, the use of too many layers leads to overfitting, i.e., the model becomes too specific to the input data and loses generalizability, and two, the model is unable to learn all the features present in the data from the limited amount of samples.

Thus, a multi-layered perceptron or a simple fully connected neural network is not ideal for multi-omics integration. Studies have showed that while deep learning can be used to combine different omics data, in order to significantly better the performance over simple fully-connected networks, a good framework is necessary. Researchers have done this by either implementing more elaborate frameworks of neural networks or feature engineering, so as to help the model know what features to learn. For example, ATHENA (Analysis Tool for Heritable and Environmental Network Association) uses grammatical evolution neural networks. In the following section we will look at a tool called MORONET that uses a complex framework.

4 Background

4.1 MORONET

MORONET combines features from different omics-specific data, patient similarity matrices, and cross omics learning in the label space for supervised learning and classification.[34] MORONET uses Graph Convolution Networks (GCNs) for learning individual omics-specific features and then uses a View Correlation Discovery Network (VCDN) to combine the individual GCN outputs for learning cross-omics features in the label space. The structure of the MORONET framework is illustrated in figure 4.1 and described in detail in the subsections below.

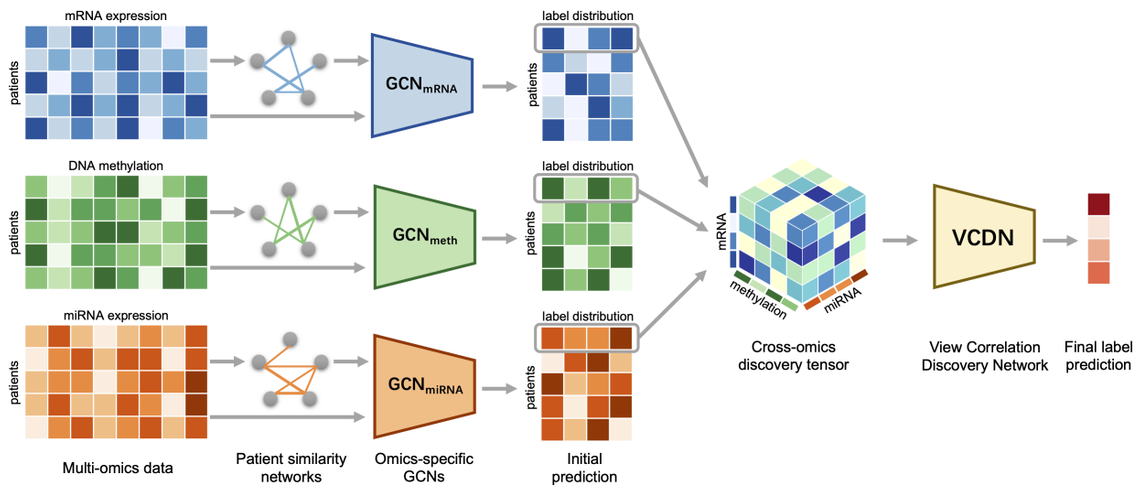


Figure 4.1: The MORONET Model[34]. (Image obtained from https://github.com/txWang/MORONET) GCNs perform omics-specific learning. They take the individual omics features and the generated patient similarity network based on these features as input, and generates labels as output. The VCDN takes these labels from each of the GCNs as input and generates a final set of labels as output. The GCNs and the VCDN are trained together.

While the primary purpose of MORONET is biomedical classification, an interpretation of feature importance can be used to identify biomarkers. If we set a particular feature's value to zero, then the performance

of the model drops. More the performance drops, more important the feature is to the classification. Thus, this performance drop provides a way to quantitatively know the importance of the biomarkers in the classification task.

4.2 Graph Convolution Networks

Graph Convolutional Networks, as described by Manessi et. al.[21], 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:

- "A feature description x_i for every node i, summarized in a $N \times D$ feature matrix X where N is the number of nodes and D is the number of input features."
- "A representative description of the graph structure in matrix form, typically in the form of an adjacency matrix A."

and generates Z, an $N \times C$ output matrix, where C is the number of categories or labels.[5]

Each of the individual omics data sets is input into a separate GCN. To convert the omics data into a GCN input, some processing needs to be done. We have the $N \times D$ feature matrix X where N is the number of patient or tissue samples and D is the number of omics features. Thus, we now have a list of nodes, but no edges.

In order to get the adjacency matrix A, Wang et. al.[34] generated a patient similarity network. 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 ϵ, then the edge is retained are retained. The adjacency between nodes i and j, A_{ij}, is calculated as:

$$A_{ij} = \begin{cases} s(x_i, x_j), & \text{if } i \neq j \text{ and } s(x_i, x_j) \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where x_i and x_j are the feature vectors of node i and node j respectively, $s(x_i, x_j)$ 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:

$$\begin{aligned} H^{(l+1)} &= f(H^{(l)}, A) \\ &= \sigma(AH^{(l)}W^{(l)}), \end{aligned} \tag{4.2}$$

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.

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 to produce an output label matrix in which the last column of labels corresponds to the test sample. Therefore, "both the features of the test sample and the correlations between the test sample and the training samples are utilized in predicting the label of the new test sample."

4.3 View Correlation Discovery Network

As each of the graph convolutional networks considers individual omics data and outputs the required label, 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-omics data set. But, that would be too simple and not consider any cross-omics correlations. To consider these, cross-omics correlations, a view correlation discovery network is used. [34]

The original view correlation discovery network[33] 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[33] was designed for data with two views. MORONET[34] 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 extendable 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-omics discovery tensor $C_j \in \mathbb{R}^{c \times c \times c}$ is constructed, where each entry of C_j is calculated as:

$$C_{j,abc} = \hat{y}_{j,a}^{(1)} \hat{y}_{j,b}^{(2)} \hat{y}_{j,c}^{(3)}, \tag{4.3}$$

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.

There are two things of note here:

- The VCDN$(\cdot)$ itself is a two layer fully connected network which outputs the final label predictions, based on the cross-omics discovery tensor formed by integrating the labels predicted by the individual graph convolutional networks.
- The VCDN$(\cdot)$'s input, the corss-omics 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 omics types included in the analysis.

4.4 MORONET Critique

The MORONET framework as mentioned in the paper[34] and their code available here (`https://github.com/txWang/MORONET`) have the following drawbacks:

- While pre-processing the data, the top 200 features were selected using their ANOVA F-values. While it is necessary to remove redundancy and reduce the number of features, the number 200 arbitrary. Also, in the process, feature associations across omics are lost.
- The multi-omics integration takes place only in the label space. Individual omics types are analysed separately in the GCNs and then the labels are brought together in the VCDN
- The VCDN scales exponentially. For a data set with 5 labels and 3 GCNs, the VCDN input has $5^3 = 125$ features, but if the number of GCNs is increased to 6 or 9, then the number of input features increases to $5^6 = 15625$ and $5^9 = 1953125$ respectively.

5 Methods

This section consists of methods that have been proposed for this study.

5.1 Pairwise Interactions of Features

In the MORONET framework, each of the GCNs models individual omics' features. This is analogous to the main effects in a linear regression. The pairwise interactions of features is then analogous to the interaction effects in a linear regression.

Consider a typical linear regression equation *without* an interaction:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 \tag{5.1}$$

where $\hat{y}$ is the predicted value of the dependent variable, b_0, b_1, and b_2 are regression coefficients, and X_1 and X_2 are independent variables. The regression coefficients, b_1 and b_2 are said to model the main effects of the independent variables.

Now consider a linear regression equation *with* an interaction:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 \tag{5.2}$$

where b_3 is a regression coefficient, and $X_1 X_2$ is the interaction. This two-way interaction is an interaction between two independent variables.

Now consider a linear regression equation with a squared term or a self-interaction:

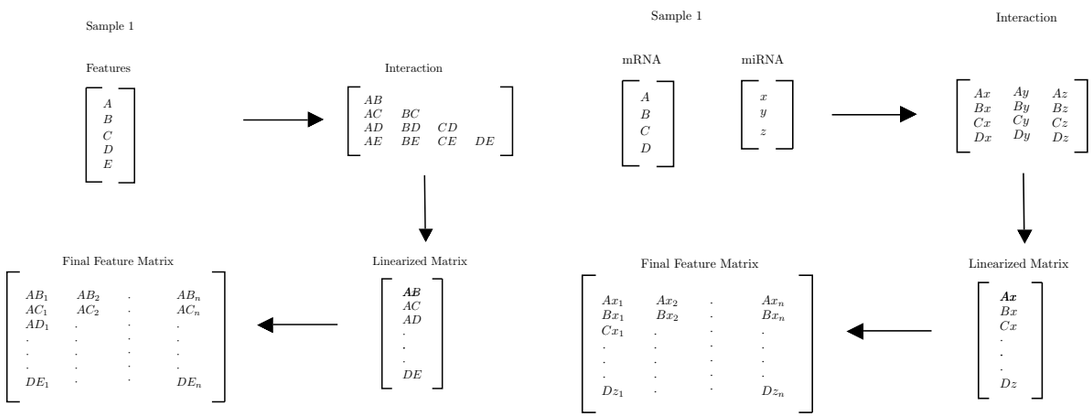
$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1^2 \tag{5.3}$$

where X_1^2 is a squared term. This two-way interaction is an interaction between a variable and itself. Here, X_i is a scalar variable. In the context of an omics data set, X_i can be the i-th patient/sample's vector of features, where X_i^2 would then be a vector of products of each feature with every other feature in X_i.

Now, based on the ensuing discussion, let the independent variables X_is be the different omics types in a multi-omics data set. Then each GCN is trying to model the main effects of the independent variables. We propose to use the interaction effects to be input to the GCN so as to perform early integration and learn valuable cross-omics interaction.

5.1.1 Naive Interactions

The pre-processing of the omics data before being input to the graph convolution networks selects the 200 most variant features in each omics type and scales the feature values to [0,1]. This leads to a simple way of introducing pairwise feature interactions into the existing MORONET model. This has been illustrated in the figure 5.1.



(a) Self-Interactions

(b) Cross-omics interaction

Figure 5.1: Naive pairwise interactions

As illustrated in figure 5.1a, the squared terms can be introduced as follows:

- **Multiplication** - For every sample, each feature of the omics type is multiplied with every other feature of the same omics type.
- **Linearization** - The products thus formed are linearized such that for an omics type input of N features, we now have a vector of $\frac{N\times(N+1)}{2}$ interaction features.
- **Similarity Matrix Preparation** - After linearization, the interaction feature matrix looks similar to the initial feature matrix, albeit with a much bigger size. This is then used to generate a similarity matrix based on cosine similarity.

This is then added to the model as a separate GCN. This process can be done for each of the omics data.

As illustrated in figure 5.1b, the interaction terms can be introduced as follows:

- **Multiplication** - For every sample, each feature of one omics type is multiplied with every feature of another omics type.
- **Linearization** - The products thus formed are linearised such that for an omics type input of M and N features each, we now have a vector of $M \times N$ interaction features.
- **Similarity Matrix Preparation** - After linearization, the interaction feature matrix looks similar to the initial feature matrix, albeit with a much bigger size. This is then used to generate a similarity matrix based on cosine similarity.

This is then added to the model as a separate GCN. This process can be done for each pair of omics data.

5.1.2 Pruning edges

There are three disadvantages of introducing pairwise interactions as described in the previous subsection on naive interactions. They are:

- **Size increase** - The input to the GCN and the GCN itself becomes too large. For instance, the standard MORONET selects 200 features per omics type to be input to the GCN. When interactions are introduced naively, each of the new GCN input has 20,100 or 40,000 features. This greatly increases the computational space and time requirements.
- **Redundancy** - The interaction features we are creating are not novel features, in the sense that, they are not obtained experimentally. This can introduce a lot of redundant features that that can reduce classification performance.
- **Biological significance** - Beyond drawing analogies to interaction effects in linear regression, there is little meaning to what the product of two features can signify. While some of the products can be used to indicate molecular interactions or other similar associations between the features, many of the products will be meaningless.

These disadvantages can be overcome by pruning the edges of the constructed interaction matrices using the following methods:

- **Utilizing previous knowledge** - We can utilize existing knowledge present in the form of molecular interaction networks or co-expression networks to select which of the interaction terms to include. For example, mir2gene[23] network obtained from miRNet[2] contains the molecular interactions between miRNAs and mRNAs.
- **Selecting top variants** - We can perform some more processing on these interactions before we form a similarity matrix or input to a GCN. We can use ANOVA F-values to find top variants among the interaction terms and select only those terms for further processing.
- **Finding eigengenes** - We can use eigengenes to represent sub-networks or co-expression modules as shown by Wang et. al.[32] After selecting all the interaction terms based on molecular interaction networks or co-expression, edges starting from the same node can be averaged over and all the nodes connecting to this node can be replaced by one eigengene. This way the sub-network is reduced to a single edge, reducing redundancy, and assuming the noise in the feature quantification was balanced, signal-to-noise ratio is also improved.

5.1.3 Concatenating Omics

Intuitively, there should be a another way to model these pairwise interactions than to individually multiply the features together, and that is to concatenate the features of the different omics together and then input them into a single GCN. The neural network, on its own, learns all the interaction terms of significance, This should ideally work, but the number of training samples is generally insufficient to learn all the interaction features. Thus, feature engineering, as mentioned in the previous subsection, become necessary to improve the accuracy of classification.

5.2 Baseline Methods

For the preliminary results presented in this paper, the performance of the original MORONET model as mentioned in the paper by Wang et. al.[34] is used as a baseline. Upon obtaining further results, these preliminary results will then be used as baselines for newer results.

Wang et. al. use two methods from mixOmics-DIABLO [28]: block PLSDA and block SPLSDA on the same data set as baselines because they are at the cutting-edge of multi-omics integration and classification. These and other classification tools like ATHENA[12] and OmicKrigging[35] can be used as baselines for our study too.

6 Data

6.1 BRCA Data

6.1.1 Data preliminaries

The Breast Invasive Carcinoma (BRCA) multi-omics data set was obtained from MORONET's Github who obtained it from The Cancer Genome Atlas Program (TCGA) through the Broad GDAC Firehose

(https://gdac.broadinstitute.org/). The data set consists of gene expression, DNA methylation, and microRNA data of 770 patients and 115 normal controls. The samples fall under the following 5 categories: Normal, Basal, Her2, LumA, and LumB. 'Normal' represents the control samples, while the other 4 labels represent the 4 sub-categories of breast invasive carcinoma.

Normal	Basal	Her2	LumA	LumB
115	131	46	436	147

Table 6.1: Categories in the BRCA data set

mRNA	DNA methylation	miRNA
20531	20206	403

Table 6.2: Number of features in the BRCA data set

The multi-omics data set, consisting of three omics types, mRNA expression, DNA methylation, and miRNA expression, are used to classify the samples into the five categories. The mRNA expression, DNA methylation, and miRNA expression data are matched data, i.e., each of them were obtained from the same set of samples.

6.1.2 Data processing

The data was divided randomly into training and testing data using a 70:30 split, while maintaining the label distribution. For DNA methylation data, only the features corresponding to coding region of the DNA were kept. Then the data was filtered. Features with zero mean signal and low variances were removed.

A variance threshold of 0.1 was applied for mRNA expression data, 0.002 for DNA methylation data, and 0 for miRNA expression data. The variance filtering thresholds were determined based on the ranges of the data. Zero variance threshold was applied to miRNA expression data because of the limited number of miRNAs present.

Redundant features can have a negative effect on the classification performance. Thus, ANOVA F-values were used to select features significantly different across the different categories. Only 200 significant features were selected for each omics data type.

Finally, each omics data was linearly scaled to [0,1]. The pre-processed data set is available at https://github.com/txWang/MORONET/tree/master/BRCA.

6.2 Brain Tissue Data

6.2.1 Data preliminaries

The Human Cerebellum Agilent data set was obtained from the Harvard Brain Tissue Resource Center (HBTRC) using University of Tennessee's genenetwork.org. The data set is available at http://gn1.genenetwork.org/webqtl/main.py?FormID=sharinginfo&GN_AccessionId=326. The data set consists of mRNA expression data from three brain regions, cerebellum, visual cortex, and dorsolateral prefrontal cortex. This is a matched data set, i.e., the three brain regions were profiled from the same individuals.

This multi-tissue data set contains mRNA expression data for three brain regions, cerebellum, visual cortex, and dorsolateral prefrontal cortex. It has 803 individuals in the data set across three categories: Alzheimer's disease (AD) cases, Huntington's disease (HD) cases, and normal controls (N).

Alzheimer's Disease (AD)	Huntington's Disease (HD)	Normal Control (NC)
388	220	195

Table 6.3: Categories of patients in the data set

6.2.2 Data processing

Many of the samples had missing SNP data. These samples were dropped. After the dropping, 323 Alzheimer's disease samples, 145 Huntington's samples, and 123 normal samples remained.

Cerebellum	Visual Cortex	Prefrontal cortex
19529	19529	19529

Table 6.4: Number of features across the tissues in the data set

7 Results

7.1 Multi-omics integration with naive self-interactions improves testing accuracy

The models were trained independently three times for 2000 epochs each and the resulting metrics were averaged and are stated below.

Metric	MORONET Baseline	Naive Self-Interactions
Training Accuracy	0.9325 (0.002)	0.9243 (0.002)
Training F1 score	0.9307 (0.002)	0.9214 (0.002)
Testing Accuracy	0.7947 (0.005)	0.8112 (0.002)
Testing F1 Score	0.8032 (0.006)	0.8107 (0.003)

Table 7.1: Performance of the model with naive self-interactions of the three omics included in comparison with the MORONET model baseline with only three omics GCNs. All values were averaged over three independent training sessions. The standard deviation of the values are presented in parenthesis.

The MORONET baseline model is the standard model as mentioned in the MORONET study[34]. It has 3 GCNs corresponding to the mRNA expression, DNA methylation, and miRNA expression data. The naive self-interaction model is a modified MORONET model. It has 6 GCNs corresponding to mRNA expression, DNA methylation, miRNA expression, mRNA-mRNA interaction, methylation-methylation interaction, and miRNA-miRNA interaction data.

As can be seen from the MORONET baseline metrics, there is significant overfitting by the model, i.e., the model fits well to the training data but generalizes poorly to unseen data. Compared to the baseline model, the model with naive self-interactions has lower training accuracy and F1 scores, but it also has lesser overfitting. There is a **2% improvement** in the testing accuracy and F1 score in the new model.

Another important observation worth mentioning is that the baseline model takes only 5 minutes to train for 2000 epochs, but the model with naive self-interactions takes over 15 hours to train for 2000 epochs. This is a very big difference in time complexity and I believe that it can be reduced by pruning the edges without reducing the testing accuracy.

7.2 Multi-omics integration with concatenated GCN input increases overfitting

The models were trained independently three times for 2000 epochs each and the resulting metrics were averaged and are stated below.

Metric	MORONET Baseline	Concatenated GCN
Training Accuracy	0.9325 (0.002)	0.9423 (0.002)
Training F1 score	0.9307 (0.002)	0.9407 (0.002)
Testing Accuracy	0.7947 (0.005)	0.7922 (0.010)
Testing F1 Score	0.8032 (0.005)	0.7984 (0.009)

Table 7.2: Performance of the model with an additional concatenated input of the three omics included in comparison with the MORONET baseline model with only three omics GCNs. All values were averaged over three independent training sessions. The standard deviation of the values are stated in parenthesis.

The baseline MORONET model is the standard model as mentioned in the MORONET study[34]. It has 3 GCNs corresponding to the mRNA expression, DNA methylation, and miRNA expression data. The model with the concatenated GCN is a modified model with 4 GCNs corresponding to mRNA expression, DNA methylation, miRNA expression data, and a concatenated input of the three omics data.

While the training accuracy and F1 scores have improved for the model with concatenated GCN, the testing accuracy and F1 score haven't. This implies that overfitting has increased, i.e., the model is fitting better to the training data but is not generalizing better to the testing data.

Combined with the results from the previous subsection, we can see that while there is scope for improvement of the model's performance on the testing data, just providing a concatenated input is not helpful. The model is not able to learn all the pairwise interaction features on its own. Thus, feature engineering is necessary.

7.3 Multi-omics integration with pruned interaction input increases overfitting

The models were trained independently three times for 2000 epochs each and the resulting metrics were averaged and are stated below.

Metric	MORONET Baseline	Pruned miRNA-mRNA Interaction
Training Accuracy	0.9325 (0.002)	0.9444 (0.001)
Training F1 score	0.9307 (0.002)	0.9430 (0.001)
Testing Accuracy	0.7947 (0.005)	0.8010 (0.012)
Testing F1 Score	0.8032 (0.005)	0.8018 (0.006)

Table 7.3: Performance of the model with an additional pruned miRNA-mRNA interaction input in comparison with the baseline MORONET model with only three omics GCNs. (All values were averaged over three independent training sessions. The standard deviation of the values are stated in parenthesis.)

The baseline MORONET model is the standard model as mentioned in the MORONET study[34]. It has 3 GCNs corresponding to the mRNA expression, DNA methylation, and miRNA expression data. The model with the pruned miRNA-mRNA interaction is a modified model with 4 GCNs corresponding to mRNA expression, DNA methylation, miRNA expression data, and a miRNA-mRNA interaction input whose edges have been pruned using the mir2gene data set[23].

The pruned miRNA-mRNA interaction data set had only 134 features. This was not enough to increase the model's performance on the test data set significantly.

8 Future Work

Moving forward, I will immediately work toward the tasks stated in subsections 8.1, 8.2, and 8.3.
Later, I will work toward the other goals.

8.1 Pre-processing

- Rather than select exactly 200 features, which is an arbitrary number, we select a percentage of features like the top 50% or such.

8.2 Pruning features

- Find pre-existing molecular interaction and co-expression networks, like miRNet, for each pair of omics data. Then use this to prune the features input into the GCN.
- Rather than select for the interactions from pre-selected 200 features, we select for the interactions from the raw data, and then prune them.

8.3 Eigengenes or Representative features

- As exemplified by Wang et. al.[32], and mentioned in section 5.1.2, we can use representative features in place of sub-networks or co-expression modules. This could reduce redundancy and noise.

8.4 Omics Imputation

- Implement a machine learning framework for imputing missing omics data.

8.5 Biomarker and Pathway Discovery

- Implement a feature selection procedure to obtain biomarkers from the trained model.
- Construct a Heterogeneous Multi-Layered Network, where each omics type represents a layer, and use Prize-collecting Steiner Forest algorithm to find molecular pathways across omics.

9 Acknowledgments

I would like to thank *Dr. Manikandan Narayanan* for his guidance and patience, and all the members of the BIRDS (Bioinformatics and Integrative Data Science) group for their valuable input.

References

[1] Eric Bonnet, Laurence Calzone, and Tom Michoel. Integrative multi-omics module network inference with lemon-tree. *PLOS Computational Biology*, 11(2):e1003983, February 2015.

[2] Le Chang, Guangyan Zhou, Othman Soufan, and Jianguo Xia. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Research*, 48(W1):W244–W251, June 2020.

[3] Marco Chierici, Nicole Bussola, Alessia Marcolini, Margherita Francescatto, Alessandro Zandonà, Lucia Trastulla, Claudio Agostinelli, Giuseppe Jurman, and Cesare Furlanello. Integrative network fusion: A multi-omics approach in molecular profiling. *Frontiers in Oncology*, 10, June 2020.

[4] Attila Csala, Aeilko H. Zwinderman, and Michel H. Hof. Multiset sparse partial least squares path modeling for high dimensional omics data analysis. *BMC Bioinformatics*, 21(1), January 2020.

[5] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2224–2232. Curran Associates, Inc., 2015.

[6] Said el Bouhaddani, Hae-Won Uh, Geurt Jongbloed, Caroline Hayward, Lucija Klarić, Szymon M. Kiełbasa, and Jeanine Houwing-Duistermaat. Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics*, 19(1), October 2018.

[7] Ziling Fan, Yuan Zhou, and Habtom W. Ressom. MOTA: Network-based multi-omic data integration for biomarker discovery. *Metabolites*, 10(4):144, April 2020.

[8] R. Fleischmann, M. Adams, O White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, July 1995.

[9] F. Garcia-Alcalde, F. Garcia-Lopez, J. Dopazo, and A. Conesa. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics*, 27(1):137–139, November 2010.

[10] Vladimir Gligorijevic and Natasa Przulj. Methods for biological data integration: perspectives and challenges. *JOURNAL OF THE ROYAL SOCIETY INTERFACE*, 12(112), NOV 6 2015.

[11] Stijn Hawinkel, Luc Bijnens, Kim-Anh Lê Cao, and Olivier Thas. Model-based joint visualization of multiple compositional omics datasets. *NAR Genomics and Bioinformatics*, 2(3), July 2020.

[12] Emily R. Holzinger, Scott M. Dudek, Alex T. Frase, Sarah A. Pendergrass, and Marylyn D. Ritchie. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics*, 30(5):698–705, October 2013.

[13] Ili Nadhirah Jamil, Juwairiah Remali, Kamalrul Azlan Azizan, Nor Azlan Nor Muhammad, Masanori Arita, Hoe-Han Goh, and Wan Mohd Aizat. Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology. *FRONTIERS IN PLANT SCIENCE*, 11, JUN 26 2020.

[14] Christopher R John, David Watson, Michael R Barnes, Costantino Pitzalis, and Myles J Lewis. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, September 2019.

[15] Atanas Kamburov, Rachel Cavill, Timothy M. D. Ebbels, Ralf Herwig, and Hector C. Keun. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 27(20):2917–2918, September 2011.

[16] Andreas Krämer, Jeff Green, Jack Pollard, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, December 2013.

[17] Tien-Chueh Kuo, Tze-Feng Tian, and Yufeng Tseng. 3omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology*, 7(1):64, 2013.

[18] Samuel Levy, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, Brian P. Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F. Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R. MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B. Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul A. Kravitz, Dana A. Busam, Karen Y. Beeson, Tina C. Mclntosh, Karin A. Remington, Josep F. Abril, John Gill, Jon Borman, Yu-Hui Rogers, Marvin E. Frazier, Stephen W. Scherer, Robert L. Strausberg, and J. Craig Venter. The diploid genome sequence of an individual human. *PLOS BIOLOGY*, 5(10):2113–2144, OCT 2007.

[19] Eric F. Lock, Jun Young Park, and Katherine A. Hoadley. Bidimensional linked matrix factorization for pan-omics pan-cancer analysis, 2020.

[20] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Computational Biology*, 10(4):e1003580, April 2014.

[21] Franco Manessi, Alessandro Rozza, and Mario Manzo. Dynamic graph convolutional networks. *Pattern Recognition*, 97:107000, January 2020.

[22] Nam D. Nguyen and Daifeng Wang. Multiview learning for understanding functional multiomics. *PLOS COMPUTATIONAL BIOLOGY*, 16(4), APR 2020.

[23] Chengxiang Qiu, Juan Wang, and Qinghua Cui. miR2gene: pattern discovery of single gene, multiple genes, and pathways by enrichment analysis of their microRNA regulators. *BMC Systems Biology*, 5(Suppl 2):S9, 2011.

[24] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An r package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, November 2017.

[25] Jonathan Ronen, Sikander Hayat, and Altuna Akalin. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Science Alliance*, 2(6):e201900517, December 2019.

[26] Marcus M. Seldin and Aldons J. Lusis. Systems-based approaches for investigation of inter-tissue communication. *Journal of Lipid Research*, 60(3):450–455, January 2019.

[27] Hai Shu, Xiao Wang, and Hongtu Zhu. D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, 115(529):292–306, April 2019.

[28] Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, January 2019.

[29] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14:117793221989905, January 2020.

[30] Nurcan Tuncbag, Sara J. C. Gosline, Amanda Kedaigle, Anthony R. Soltis, Anthony Gitter, and Ernest Fraenkel. Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLOS Computational Biology*, 12(4):e1004879, April 2016.

[31] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, June 2010.

[32] Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio C. P. Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Yucheng T. Yang, Min Xu, Michael J. Gandal, Shaoke Lou, Jing Zhang, Jonathan J. Park, Chengfei Yan, Suhn Kyong Rhie, Kasidet Manakongtreecheep, Holly Zhou, Aparna Nathan, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, Yan Jiang, Kiran Girdhar, Gabriel E. Hoffman, Selim Kalayci, Zeynep H. Gümüş, Gregory E. Crawford, Panos Roussos, Schahram Akbarian, Andrew E. Jaffe, Kevin P. White, Zhiping Weng, Nenad Sestan, Daniel H. Geschwind, James A. Knowles, and Mark B. Gerstein and. Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420):eaat8464, December 2018.

[33] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu. Generative multi-view human action recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6211–6220, 2019.

[34] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. MORONET: Multi-omics integration via graph convolutional networks for biomedical data classification. July 2020.

[35] Heather E. Wheeler, Keston Aquino-Michaels, Eric R. Gamazon, Vassily V. Trubetskoy, M. Eileen Dolan, R. Stephanie Huang, Nancy J. Cox, and Hae Kyung Im. Poly-omic prediction of complex traits: OmicKriging. *Genetic Epidemiology*, 38(5):402–415, May 2014.

[36] Kin Yau Wong, Cheng Fan, Maki Tanioka, Joel S. Parker, Andrew B. Nobel, Donglin Zeng, Dan-Yu Lin, and Charles M. Perou. I-boost: an integrative boosting approach for predicting survival time with multiple genomics platforms. *Genome Biology*, 20(1), March 2019.

[37] Jianguo Xia, Christopher D. Fjell, Matthew L. Mayer, Olga M. Pena, David S. Wishart, and Robert E. W. Hancock. INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Research*, 41(W1):W63–W70, June 2013.

[38] Tian Xia, John V. Hemert, and Julie A. Dickerson. OmicsAnalyzer: a cytoscape plug-in suite for modeling omics data: Fig. 1. *Bioinformatics*, 26(23):2995–2996, October 2010.

[39] Xia Yang. Multitissue multiomics systems biology to dissect complex diseases. *Trends in Molecular Medicine*, 26(8):718–728, August 2020.