Dissertation Proposal

# MULTIOMICS DATA INTEGRATION AND MULTIPLEX GRAPH NEURAL NETWORK APPROACHES

by

ZIYNET NESIBE KESIMOGLU

University of North Texas

Denton, Texas

November 2021

# Table of Contents

# 1. Dissertation Summary

With decreasing cost and increasing throughput of next-generation sequencing technologies, multiple types of omics data (e.g. transcriptomics, proteomics, etc.) from the same set of samples have been generated. Since each omic data modality contains a unique aspect of the underlying biology, multiple omics (multiomics) data is integrated to capture various biological information. Previous studies showed that integrating multiomics data improved the models to solve existing biological problems.

Multiomics integration is traditionally done by extracting features from multiomics data. However, biological networks (i.e. graphs) represent previously-inferred or data-driven associations between entities (i.e. nodes) such as genes and patients, and those additional associations are important to utilize. Recently, some advanced approaches have been developed, named Graph Neural Networks (GNN). GNN is a neural network that is applied directly on graph-structured data utilizing node features. GNN considers the local neighborhood of a node on the graph with additional node features. Leveraging node associations and features, the previous studies generated considerable GNN-based models even on a single graph. Some GNN-based models adapted attention mechanisms to adjust the associated neighbor's impact instead of assuming equal contributions. Even though GNN enables the utilization of node associations in addition to node features, integrative GNN-based approaches on multi-view heterogenous graphs are limited and have some limitations to address.

In this dissertation, first, I introduce a computational tool integrating multiomics data to infer gene regulatory networks. Second, I present a cancer subtype prediction methodology integrating multiomics features and patient associations from multiple graphs with GNN-based approaches (i.e. multiplex GNN approaches). Finally, I aim to develop a GNN-based architecture that operates on the multiplex, heterogenous, graph-structured data integrating multiomics datasets, utilizing node associations, incorporating edge and node features, and adapting attention mechanisms. This proposed dissertation will help to better characterize a given disease by integrating multiomics data and improving the exploitation from node associations with novel multiplex approaches.

# 2. Specific Aims

In this dissertation, I have three specific aims to solve by developing computational tools with multiomics integration and multi-view graph neural network utilization for some biological problems, namely gene regulation inference, cancer subtype prediction as node classification, and drug-side effect prediction as edge prediction.

## Aim 1. Develop a computational tool to infer genome-wide competing endogenous RNA interactions integrating multiomics data

CeRNA regulation has been recently discovered. Despite existing ceRNA inference tools, there are crucial drawbacks of improper utilization of expression datasets and preclusion of the biological processes that potentially play important roles in that regulation. To address the critical drawbacks in the current ceRNA inference methods, I developed a computational tool named CRINET (CeRNA Interaction Network). CRINET infers pairwise and groupwise ceRNA interactions leveraging multiomics datasets. CRINET-inferred interactions were validated with high-throughput protein expression-based benchmarks. Knockdown of the inferred ceRNAs highly affected the expression of the ceRNA partners. Inferred ceRNAs were significantly associated with the cancer-related genes and processes.

## Aim 2. Develop a node classification framework encoding multiomics features and applying convolutions on multiplex graph-structured data

Cancer subtype prediction is an important problem and assigning accurate labels could help the patients get the appropriate treatment. GNN-based models emerged including node associations and node features at the same time, but existing studies have some limitations at leveraging multiomics datasets and multiple networks. Addressing limitations of current approaches, I developed a cancer subtype prediction methodology, called SUPREME. SUPREME integrates multiomics datasets applying graph convolutions on multiplex networks along with multiomics node features in order to predict subtypes of cancer patients. SUPREME utilizes multiomics data and associations between patients with convolutions on *multiplex* patient similarity networks gathering *common* and *complementary* signals between multiple datatypes. SUPREME improved the prediction accuracy as compared to the baseline methods and the state-of-the-art methods. SUPREME results were consistent with the exclusion of single omics features and graphs, with overall consistent results for any combination of multiomics datasets. SUPREME-inferred predictions have significant survival differences consistently, indicating that leveraging multiomics features and multiplex graphs, SUPREME might give more accurate labels as compared to the ground truth, which depends mainly on one omics data.

## Aim 3. Develop a GNN-based architecture with convolutions on multiplex heterogeneous graphs with attention mechanisms

Deep learning on graph-structured data has been recently improved. Although there are some with considerable architectures that utilize node features and associations on the graph-structured data, applying them on multiplex heterogeneous networks is still under-explored. In addition to utilizing graph-structured data, there are attention-based mechanisms that adjust the importance of the neighbors of the nodes, and existing attention-based models showed improvements over the baseline models. To utilize those advances, I will develop a novel neural network architecture that operates on *multiplex heterogenous* graph-structured data integrating multiomics data and utilizing node associations. *Adapting attention mechanisms* that pay attention to the most relevant part of the input, the architecture will extract *more informative neighborhoods* from given networks. The multiplex approach will enable

*different impacts* for different associations between the same nodes. Considering the associations from multiple node types with *heterogeneous* networks, the utilization of node associations will be boosted. This novel, multiplex, heterogenous GNN architecture will utilize multiomics features and multiple node associations, where the associations might come from *multiple graphs* of *multiple node types*.

The proposed aims will result in a better characterization of a given disease integrating multiomics data and improving the exploitation from node associations with novel approaches. By integrating multiomics data properly, we will have a better understanding of gene regulatory circuitry addressing important drawbacks pertaining to ceRNA regulation. For clinical outcome prediction problems, GNN-based approaches could potentially help to get more consistent and clear subtype characteristics from available data and improve the ground truth's labeling that is mainly based on a single datatype. With novel, multi-view GNN-based architecture, the impact of known associations along with multiomics integration will be boosted with *multiplex heterogeneous* graphs with adapted *attention* mechanisms. Novel multiplex GNN-based approaches with proper multiomics integration will be important assets for various biological problems.

# 3. Innovation

CeRNA regulation as related to microRNA regulation has been recently discovered. MicroRNAs bind to RNAs and mostly repress their expression, so the expression of microRNAs and their target RNAs generally have a negative correlation. In ceRNA regulation, some target RNAs compete for their common regulating microRNAs. For instance, we can assume that there are two targets for $\mathrm{micro}RNA_a$: $RNA_x$ and $RNA_y$. With some tendency, when $\mathrm{micro}RNA_a$ binds to $RNA_x$, then $RNA_y$ is free from this microRNA's regulation. Since those competing RNAs indirectly positively regulate each other, potentially leading to a positive correlation among the expression of ceRNA partners. Several existing ceRNA inference tools use sequence-based and expression-based clues, however, they have some issues to address. Current approaches focus on inferring only pairwise interactions. However, *groupwise ceRNA interactions* should be considered among ceRNAs besides pairwise ceRNA interactions based on the premise that several ceRNAs could work together to sequester microRNA(s) targeting one or more key ceRNAs. MicroRNA expression should be at some level to be able to regulate many targets, and similarly, genes should be available with higher expression to be regulated by many microRNAs. In the current computational studies, a microRNA could be assigned to "mediate" thousands of ceRNA interactions without considering if the microRNA has *sufficient abundance*. Considering "effective expression" instead of using all expression abundance could be beneficial while integrating multiomics data. Furthermore, since ceRNAs positively regulate each other, two genes having many common ceRNA partners might be inferred just because of the *amplifying effect of regulation by common ceRNA partners*. Thus, excluding these false-positive ceRNA interactions is important. Also, different from the studies that analyze only differentially expressed (DE) genes, all long noncoding RNAs (lncRNAs) and pseudogenes should be included from available multiomics data in addition to coding RNAs (including non-DE mRNAs) to understand comprehensive ceRNA regulation.

To address those critical drawbacks, I developed a computational tool named CRINET to infer genome-wide ceRNA networks. CRINET considers all mRNAs, lncRNAs, and pseudogenes as potential ceRNAs and incorporates a network deconvolution method to exclude the spurious ceRNA pairs. Crinet excludes the effect of copy number aberration (the number of copies of each region of the genome) from expression datatype and measures the sufficiency of expression abundance considering the number of regulators and targets.

Deep learning has the ability to deal with high-dimensional data, and studies are getting the benefit of deep learning rapidly on graph-structured data. GNNs started to be employed with the application to biological problems such as cancer type/subtype prediction and drug response prediction [26][27]. Although there are considerable GNN-based models, they utilize only a single graph or they utilize omics-specific features independently. To apply the GNN-based model on multiplex networks with inferring associations from multiomics datasets, I present a cancer subtype prediction methodology, called SUPREME. SUPREME applies graph convolutions on patient similarity networks collecting common and complementary signals from *multiplex* graphs and *multiomics* features.

Along with GNN, attention-based mechanisms have recently been used on graph networks, adjusting the local neighbor's impact on each node. Attention-based mechanisms help handle the spurious interactions in the inferred networks, and adjust the neighbor's effect rather than assuming equal contribution from all. There are important advances in attention-based mechanisms and some studies are applying the attention mechanisms on GNN, however, GNN-based models with attention mechanisms need to be

integrated properly for multiplex graphs with edge and node features. Furthermore, graphs are based on one node type for most existing GNN models, however, integrating associations from multiple node types simultaneously could be important since the direct impact of associations from multiple nodes will be considered on heterogenous networks. For this purpose, I aim to develop a novel GNN-based architecture that is applicable on *multiplex heterogenous* networks adapting *attention* mechanisms on local neighbors.

Utilizing current advances and properly improving advanced solutions, multiomics integration employed with node associations on multiplex graphs could be an important asset for existing biological problems and diseases.

# 4. Significance

Next-generation sequencing technologies have been improved and multiomics data from tissues have been generated. Utilizing the high dimensional biological datasets in public databases, several studies rely only on one type of biological datasets such as gene expression, mutation, microRNA expression, and DNA methylation [3][4][5][6][7][9]. However, biological datasets have some common and complementary signals to properly integrate in order to better characterize the diseases. Thus, integrative computational methods are needed to comprehensively solve biological problems.

There is a recently discovered gene regulation layer called competing endogenous RNA (ceRNA) regulation [20]. Given the enormous number of RNAs in the genomes, it is cost and labor-prohibitive to identify ceRNA interactions experimentally. Several computational tools have been published to infer ceRNA networks. Hermes [57] uses gene and microRNA expression profiles utilizing information theory. In that way, it infers the ceRNA interactions considering their impact on each other and also considering impact via common regulators. MuTaME [58] integrates sequence patterns and coexpression between candidate ceRNA pairs, while cefinder [59] uses sequence information alone. Cupid [19] integrates sequence-based evidence reevaluating binding sites of microRNAs and functional evidence derived from RNA and miRNA expression profiles. Using ensemble machine learning classifiers, Cupid predicts candidate miRNA-target interactions by training them on the validated interactions. Then some candidate miRNA-target interactions are excluded for ceRNA inference if inferred targets compete for their predicted miRNA regulators. In that way, Cupid infers ceRNA interactions and microRNA-target interactions simultaneously. Keeping the ceRNA inference algorithm the same, an extension is added to Hermes [16] identifying specific microRNAs that mediate each interaction. In that study, most of the ceRNA pairs did not have significant coexpression and having significant coexpression did not implicate genes as ceRNA partners. This finding contradicted some existing approaches [58][59]. Extended Hermes [16] significantly outperformed MuTaME and cefinder with high-throughput validation. Cancerin [17] integrated transcription factors, DNA methylation, and copy number alteration in addition to expression profiles and considered those as additional regulators to microRNAs for a LASSO-based feature selection. Despite existing tools, there is a crucial need for more advanced methods to infer ceRNA interactions. Current tools attempt to infer only pairwise ceRNA interactions, however, groupwise ceRNA interactions could exist based on the premise that several ceRNAs could work together to sequester microRNA(s) targeting key ceRNA(s). Furthermore, in most of the existing tools, a microRNA could be assigned to mediate many ceRNA interactions without considering whether microRNA abundance is sufficient to target all these molecules. Most studies did not involve copy number aberration that potentially has regulation on expression profiles. There is a gap for ceRNA inference tools in terms of inference method and also the proper multiomics utilization.

Recent significant advances in the GNN-based methodologies should be utilized for existing biological problems. Cancer subtype prediction is an important problem for obtaining proper treatment in particular cancer. Although some of the important factors differentiating subtypes are discovered from clinical implications and genomic datasets, still, the differentiating subtype is not clear either way. Using current omics datasets, cancer subtype labels are assigned for some existing cancer types. This kind of model uses mainly one or few datatypes and features. Ignoring available datatypes from the analysis, we might miss some signals having an impact on subtype characteristics. On the other hand, when we analyze different omics datasets, they will not give consistent results, requiring an advanced integration of multiple datatypes for complex diseases like cancer.

Recently several studies have developed computational tools to integrate multiomics datasets to discover cancer subtypes. Most of the previous studies focus on unsupervised multiomics data integration. One popular unsupervised approach is iCluster [33], which uses a joint latent variable model. iCluster concatenates all datasets with dimension reduction and then clusters the patients. Another tool, Similarity Network Fusion (SNF), uses patient networks from different datatypes, and integrates those networks applying a nonlinear fusion step, and obtains the fused patient network to do the clustering on [34]. In PINSPLUS, the small changes in the datatypes will be stable in the same subtype to cluster the cancer patients [35]. MuNCut performs clustering within and across datasets by connecting them in one-to-one order based on dependency between datasets [37]. These aforementioned unsupervised studies do not use the additional patient label information that is widely collected from existing studies, however, predefined labels could help develop supervised approaches with improved accuracy. Also, as compared to the existing studies, there are some recent architectures utilizing network associations in addition to the omics features.

Biological networks (i.e. graphs) are mostly preferred to encode the relation between nodes such as patients, genes, diseases, and drugs. Graph representation learning caught the attention by encoding the graph-structured data into a low-dimensional representation. Encoding unstructured networks is a hard problem since the graph is not grid-like (like images) and thus the relations are more complex due to the different number of neighbors and lack of neighbor positions. Some approaches utilize graph-structured data [42][43], however, they do not utilize the node features of the local neighborhood that we have plenty of. Recently, there are improvements in graph-structured data, and Graph Neural Network (GNN) started to employ node features and local graph structure at the same time, and they are recently started to be used in biological problems. Although some studies showed the benefit of using network associations in addition to node features, advanced approaches are needed to have integrative analysis with GNNs using multiplex graphs. Some existing studies used limited datatypes and associations, some of them fused multiple networks into one, or used multiple predictions from separate models [32][50][51], causing missing information. Also, patient similarity networks are important to get the associations of one omics data between patients, but integrating each patient network with node features of other datatypes is crucial to get the associations between different omics datasets. To utilize the existing GNN-based architectures more, it is needed to get the important signals within and between omics datasets of multiple graphs.

In addition to existing GNN-based architectures, there is a gap for GNN-based models applying on multiplex heterogenous networks. As an example, for the cancer subtype prediction problem, we have nodes as patients. Patients generate a graph representing their associations and their node features in order to integrate into a GNN-based model. However, each relation from a different dataset could add complementary signals independently. Therefore, multiple graph relations with node and edge features could be involved in GNN-based models properly. Moreover, GNN-based architecture should learn the contribution of local neighbors' associations instead of sticking to the local neighborhood of an inferred or published network. Adapted attention-based mechanisms will learn the impact of each local neighbor, having adjusted neighborhood rather than a fixed input network with equal contributions from all neighbors. Furthermore, there are multiple graphs from multiple datatypes for multiple node types such as patients and genes, so heterogenous networks with multiple nodes should be considered while we are embedding the graph structure. Since we cannot apply the current GNN models to involve these advanced features collectively, developing novel multiplex heterogenous GNN-based architectures is important.

The proposed dissertation will establish novel integrative computational tools using multiomics data applying to other organisms and diseases, flexible to add any dataset into the model. I will utilize recent

gene regulatory networks and new deep learning-based approaches on graphs while integrating multiomics data including common and complementary signals from multiplex networks.

# 5. Approach

## Aim 1.Develop a computational tool to infer genome-wide competing endogenous RNA interactions integrating multiomics data

To infer competing endogenous RNA (ceRNA) interactions integrating multiomics data properly, I developed a computational tool named CRINET [56]. CRINET infers a genome-wide ceRNA network with pairwise and groupwise associations addressing critical drawbacks in the existing studies.

### Aim 1. a. Background

The underlying biological processes in complex organisms are governed by multilayered gene regulatory networks. Dysregulation in these networks causes diseases such as cancer. Among these layers, microRNA-gene regulatory networks are one of the important and well-studied layers. MicroRNAs are small non-coding RNAs that bind to other RNAs such as mRNA, long non-coding RNA (lncRNA), and circular RNA to regulate their expression post-transcriptionally. MicroRNAs generally bind to the 3'-UTR (untranslated region) of their target mRNAs and repress their expression. Recently, a new regulatory layer related to microRNAs has been discovered where certain RNAs targeted by common microRNAs "compete" for these microRNAs and thereby regulate each other indirectly by making the other RNA(s) free from microRNA regulation [20]. Such indirect interactions between RNAs are called competing endogenous RNA (ceRNA) interactions.

It has been shown that ceRNA interaction has important roles in diseases including cancer [21][22][23][24]. There is a regulation multiplicity between microRNAs and RNAs, meaning that a microRNA could have multiple RNA targets, and an RNA could be targeted by multiple microRNAs. Given the enormous number of RNAs and difficulty of deciphering microRNA binding targets accurately, identifying ceRNA interactions experimentally is cost- and labor-prohibitive. Therefore, computational tools are crucial to infer ceRNA interactions in complex genomes like human.

Despite efforts in this new area [16][17][18][19][57], there exist some limitations in the current computational approaches. Current tools attempt to infer only pairwise ceRNA interactions, however, groupwise ceRNA interactions could exist based on the premise that several ceRNAs could work together to sequester microRNA(s) targeting key ceRNA(s). Furthermore, in most of the existing tools, a microRNA could be assigned to mediate many ceRNA interactions without considering whether microRNA abundance is sufficient to target all these molecules. In this study, I present a computational pipeline to infer cancer-associated ceRNA interactions genome-wide addressing these critical drawbacks by inferring groupwise and pairwise ceRNA interactions considering microRNA abundance.

CRINET was applied to the breast cancer dataset from the Cancer Genome Atlas (TCGA) to infer ceRNA interactions and groups. I evaluated CRINET-inferred microRNA-target interactions with protein expression-based benchmarks having increased performance following filtering steps. Expression change in the inferred ceRNAs was highly affected following the knockdown of their ceRNA partners. Inferred ceRNAs, hub ceRNAs, and some ceRNA groups were significantly associated with the cancer-related genes and processes, and consistently involved in the immune system-related processes, thus CRINET-inferred ceRNAs could be important in the view of the studies validating the relation between immunotherapy and cancer.

## Aim 1. b. Methodology

CRINET is a computational tool to infer genome-wide ceRNA interactions and groups (Figure 1). Briefly, the first step is the data preparation step. In the second step, miRNA-target interactions are computed by incorporating expression datasets and considering expression abundance sufficiency. Starting with final miRNA-target interactions, ceRNA interactions are inferred in the third step. In the last step, ceRNA groups are inferred from the ceRNA network and integrated into the final network. In the following, each step of CRINET is explained in more detail.

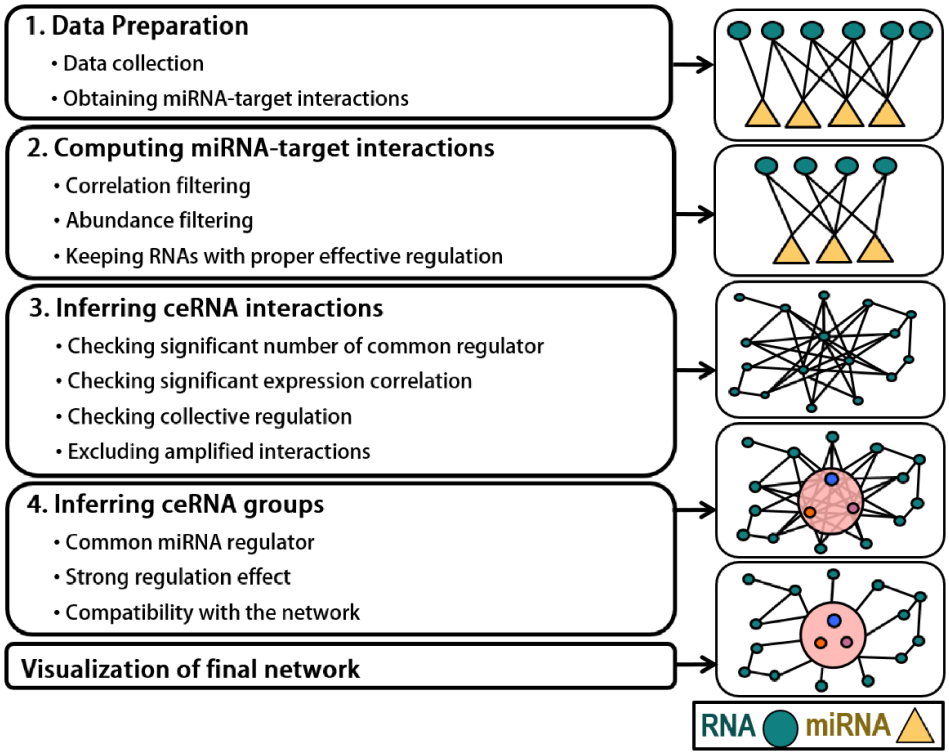


*Figure 1 CRINET pipeline*

**1. Data Preparation:** CRINET incorporates miRNA-target interactions with binding scores, gene-centric copy number aberration (CNA), and expression datasets. If binding scores are not available, the same score for all interactions could be used.

We obtained the datasets from the TCGA project including gene expression, miRNA expression, and CNA for a total of 1107 breast (BRCA) tumor samples (available at https://www.cancer.gov/tcga). We preprocessed each datatype separately obtaining normalized expression values (gene expression as FPKM, miRNA expression as RPM, and gene-centric CNA using [11]). We obtained all conserved and nonconserved miRNA-target interactions with weighted context++ scores from TargetScan [12]. To compute pseudogene and lncRNA targets of miRNAs, we ran TargetScan separately for lncRNAs and pseudogenes. After normalization, we used these normalized scores as the weight for each miRNA-target interaction assuming that these scores show the binding strength between miRNA and its gene target.

**2. Computing miRNA-target interactions:** In this step, we computed final miRNA-target interactions leveraging expression datasets and considering expression abundance sufficiency.

- Integrating interactions with expression datasets (correlation filtering): Since miRNAs are known to repress their target genes, we kept only miRNA-target pairs having a negative correlation between their expressions. We also applied random sampling with replacement to compute the correlation coefficient of each miRNA-target pair 1000 times and required that the threshold was satisfied for ≥99% of the samplings.
- Getting interactions with sufficient abundance and binding probabilities (abundance filtering): In the miRNA-target interaction sets, a miRNA could be assigned to mediate thousands of RNAs, and similarly, an RNA could be assigned to many miRNAs as a potential target. To quantify the expression sufficiency for the putative interactions, I introduced Interaction Regulation ($IR$) formulated as:

$$IR(r,t) = \frac{Exp(r).Exp(t).score_{rt}}{\sum_{j \in r\text{'s targets}} Exp(j).score_{rj}} * \frac{Exp(t).Exp(r).score_{rt}}{\sum_{j \in t\text{'s regulators}} Exp(j).score_{jt}} \quad (1)$$

where $\mathrm{IR}(\mathrm{r},\mathrm{t}) \in \mathbb{R}^n$ is the IR of the regulator r and the target t across samples, $\mathrm{Exp}(.) \in \mathbb{R}^n$ is the expression vector across samples, $\mathrm{score}_{\mathrm{rt}} \in \mathbb{R}^n$ is the normalized binding score for the interaction between the regulator r and the target t. Multiplications are element-wise for the vectors. Using this formula, we kept the final miRNA-target interactions having high IRs.

- Keeping genes with proper effective regulation: To exclude genes from analysis if they were not under strong miRNA regulation based on the final miRNA-target interactions, we introduced Effective Regulation ($ER$) formulated as:

$$ER(t) = \sum_{r \in t\text{'s regulators}} \frac{Exp(r)score_{rt}}{\sum_{j \in r\text{'s targets}} score_{rj}} \quad (2)$$

To keep genes with proper effective regulation by miRNAs, I filtered out genes without a strong negative correlation between their expression and $ER$, assuming that they did not have strong miRNA regulation for the used dataset. We also applied random sampling with replacement to compute correlation for 1000 times and required that the threshold was satisfied for ≥99% of the samplings. I used the remained genes (called candidate genes) for further analysis.

**3. Inferring ceRNA interactions:** To infer ceRNA interactions, I generated all possible gene-gene combinations using candidate genes and filtered them based on the following criteria:

- Checking the significant number of the common regulator: Since ceRNAs should have common miRNAs to compete for, I kept gene pairs having a significant number of common miRNA regulators.
- Checking significant expression correlation: Since the ceRNA pairs indirectly positively regulate each other and CNA considerably affects the expression values, I kept the gene pairs having significant partial correlation when excluding the CNA effect from the gene expression.
- Checking collective regulation: If there exists ceRNA regulation between two genes then both genes compete for common miRNAs and those miRNAs affect both genes simultaneously. We called this regulation Collective Regulation ($CR$) formulated as:

$$CR(S) = corr\left(\sum_{s \in S} ER(s), \sum_{s \in S} Exp(s)\right) \quad (3)$$

where $S$ is a set of genes having ceRNA interactions and $corr()$ is the Pearson correlation function. We kept the ceRNA pairs if they had a low $CR$.

We applied random sampling with a replacement for both partial correlation and $CR$ measurements (last two steps) 100 times separately and kept the interactions when the threshold was satisfied for ≥99% of the samplings.

- Excluding amplified interactions: Having common ceRNA partners between any two genes will increase the correlation between their expression. If a gene pair has too many common ceRNA partners, then some of the ceRNA interactions could be spurious due to the high number of common ceRNA partners. To exclude such spurious interactions from the inferred network, I employed a network deconvolution algorithm [13] and kept the top one-third of ranked interactions as pairwise ceRNA network.

**4. Inferring ceRNA groups:** In ceRNA regulation, each ceRNA pair compete for common miRNA(s) and act as a decoy to make the other RNA free from miRNA regulation. This competition could occur among more than two RNAs, or between two groups of RNA, too. Based on this premise, CRINET inferred ceRNA groups in addition to ceRNA interactions.

To obtain ceRNA groups, I utilized one of the popular community detection algorithms named Walktrap [14] on the weighted ceRNA network where weights were normalized partial correlation coefficient. We kept the groups satisfying all the group conditions, otherwise split them iteratively. Three group requirements of CRINET are listed as follows:

- Common miRNA regulator: To be able to compete, all the group members were required to have at least one common miRNA regulator.
- Strong regulation effect: CeRNAs in a group were expected to have a stronger miRNA regulation effect as a group than as individual ceRNAs. Thus, I required that the $CR$ of a group must be stronger (i.e. reduced) than the correlation between expression and ER of ≥90% of the group members. Moreover, to ensure that most of the group members would be under strong collective regulation effect, I required that an average difference between $CR$ of the group and $corr(Exp(g), ER(g))$ for each gene $g$ in the group was >0.
- Compatibility with the network: To hold inference consistency of the ceRNA network, for a given group, I called each ceRNA partner of group members as a neighbor and expected the group to satisfy any two of the three conditions for at least 90% of its neighbors. The conditions are i) at least one common miRNA regulator between the group and the neighbor, ii) a strong Pearson correlation based on expression, and iii) strong collective regulation between the group and the neighbor.

## Aim 1. c. Results

I tested CRINET on breast tumor samples from TCGA (See Data preparation section for details). I computed miRNA-target interactions (Table 1) and used them to infer 17,443 pairwise ceRNA interactions (Table 2). Using this pairwise ceRNA network, I obtained 81 ceRNA groups after applying 1508 iterations of Walktrap. Thirty-five of these groups were connected to at least one node in the final network, while the others had interactions only within the group. After this step, I had the grouped ceRNA network with 4352 nodes (4317 individual genes and 35 groups of genes) and 17,274 edges between inferred nodes.

| Step | # of Interactions | # of miRNAs | # of Genes |
|---|---|---|---|
| ALL | 8,193,904 | 888 | 28,129 |
| ALL40 | 3,285,523 | 888 | 27,999 |
| CORR | 535,284 | 888 | 23,099 |
| CORR+ABUN | 165,937 | 888 | 21,261 |

ALL: All obtained miRNA-target interactions; ALL40: Top 40% of ALL interactions based on weighted context+ + score; CORR: After correlation-based filtering on ALL40 interactions ($< -0.1$); CORR+ABUN: After abundance-based filtering on CORR interactions (log IR $> -4.89$ for $>80\%$ of samples)

*Table 1 Number of miRNA-target interactions after each miRNA-target interaction filtering step in CRINET.*

| Step | # of Pairs | # of Genes |
|---|---|---|
| ALL | 212,994,480 | 20,640 |
| Step1 | 16,082,000 | 20,640 |
| Step2 | 247,885 | 11,910 |
| Step3 | 209,220 | 11,726 |
| Step4 | 52,858 | 7,263 |
| Step5 | 17,443 | 4,494 |

ALL: All candidate pairs after keeping proper genes with effective regulation; Step1: pairs with a significant overlap for common miRNAs; Step2: pairs after filtered based on partial correlation between gene expressions excluding copy number aberration effect; Step3: pairs after filtered based on collective regulation; Step4: pairs after applying random sampling for Step2 and Step3; Step5: pairs after applying the network deconvolution method to exclude the spurious ceRNA pairs.

*Table 2 Number of remaining ceRNA pairs after each ceRNA interaction filtering step in CRINET.*

To check the reproducibility based on different datasets, the robustness to different hyperparameters, and the effect of each step in ceRNA inference, I conducted a more detailed analysis of CRINET results. To check the reproducibility of CRINET based on different datasets, I ran CRINET on two equal-sized random samplings of the breast cancer dataset multiple times. To avoid bias in the comparisons, I ensured that both samplings had similar subtype distribution (namely Basal-like, Normal-like, Luminal-A, Luminal-B, and Her2-enriched). I observed highly overlapping interactions and ceRNAs among different runs. I checked the distribution of consistently overlapping ceRNAs and observed that the mean degree (i.e. number of associated nodes) of these ceRNAs was much higher as compared to the overall mean degree (p-value <2.10−16) suggesting that consistently inferred ceRNAs were the hub ceRNAs highly involved in my inferred ceRNA network. Moreover, to examine the effect of each step in CRINET, I disabled major steps in ceRNA inference and evaluated the results. Disabling individual steps made a substantial difference in the inferred results. However, when I modified the hyperparameters in each of these steps, I observed highly overlapping interactions suggesting that CRINET is robust to different choices of hyperparameters.

Since I built a ceRNA network relying on miRNA-target interactions, proper selection of these interactions is important; therefore, I evaluated each filtering step of miRNA-target interactions using protein expression-based benchmarks. I utilized a Reverse Phase Protein Array (RPPA) dataset for the MDA-MB-231 breast cancer cell line from The Cancer Proteome Atlas (TCPA) database (accession number:

TCPA00000001) [21] to assess the inferred miRNA-target interactions as in [9]. I used 104 antibodies, their fold-change for 141 transfected miRNAs, and mock controls. For each miRNA-target interaction, I measured the expression fold ratio of each antibody of the target for the miRNA transfection relative to the average mock transfections. Table 3 confirms the preferential down-regulation of predicted miRNA targets, getting higher after each consecutive filtering step showing the positive effect of filtering for each independent interaction. I also checked the average of all targeting miRNAs per gene relative to average mock transfections and observed a similar down-regulation tendency. Although the ratio did not increase for the last step, it was due to few genes. As a negative control, I used non-inferred interactions and did not observe any strong down-regulation tendency for all the filtering steps for both phases.

| Step | # of Interactions | Interaction Phase | Gene Phase |
|---|---|---|---|
| ALL | 8,193,904 | 6,248/4,932 ≈ 1.3 | 90/74 ≈ 1.2 |
| ALL40 | 3,285,523 | 613/403 ≈ 1.5 | 97/59 ≈ 1.6 |
| CORR | 535,284 | 183/112 ≈ 1.6 | 71/37 ≈ 1.9 |
| CORR+ABUN | 165,937 | 179/111≈ 1.6 | 68/39 ≈ 1.7 |

Interaction phase shows the expression fold reduction of each antibody of target for its transfected miRNA regulator relative to mock transfection. Gene phase shows average expression fold reduction of each antibody of target for all transfected miRNA regulators relative to mock transfection. Down-regulated over up-regulated numbers along with the ratio are shown (ratio is expected to be more than 1 to have down-regulation tendency. Higher is better). See Table 1's caption for the definition of row labels.

*Table 3 Evaluation of miRNA-target interaction filtering steps for the computed miRNA-target interactions using miRNA transfection data.*

To evaluate predicted miRNA-target interactions in [9], the authors focused on the ESR1 protein, showing that ESR1 protein expression in TCGA breast cancer tumors (profiled by RPPA using the antibody ER.alpha.R.V_GBL.9014870) had a strong negative correlation with the expression of predicted miRNA regulators. Ranking samples based on miRNA expression, the top 10%, and the bottom 10% samples were compared based on ESR1 protein expression. Similarly, I generated a heatmap showing protein expression for CRINET-inferred miRNAs regulating ESR1. Figure 2 shows nine CRINET- inferred and 12 Cupid-inferred miRNAs regulating ESR1, having five miRNAs as common. I quantified the anticorrelation between miRNA and protein expression by measuring the fold-change of mean protein expression for the top 10% samples with respect to the bottom 10%. Our results indicated that the expression of CRINET- inferred miRNAs for ESR1 had high anticorrelation with protein expression with high fold-change consistently while Cupid had some low fold-change such as hsa-mir-381.

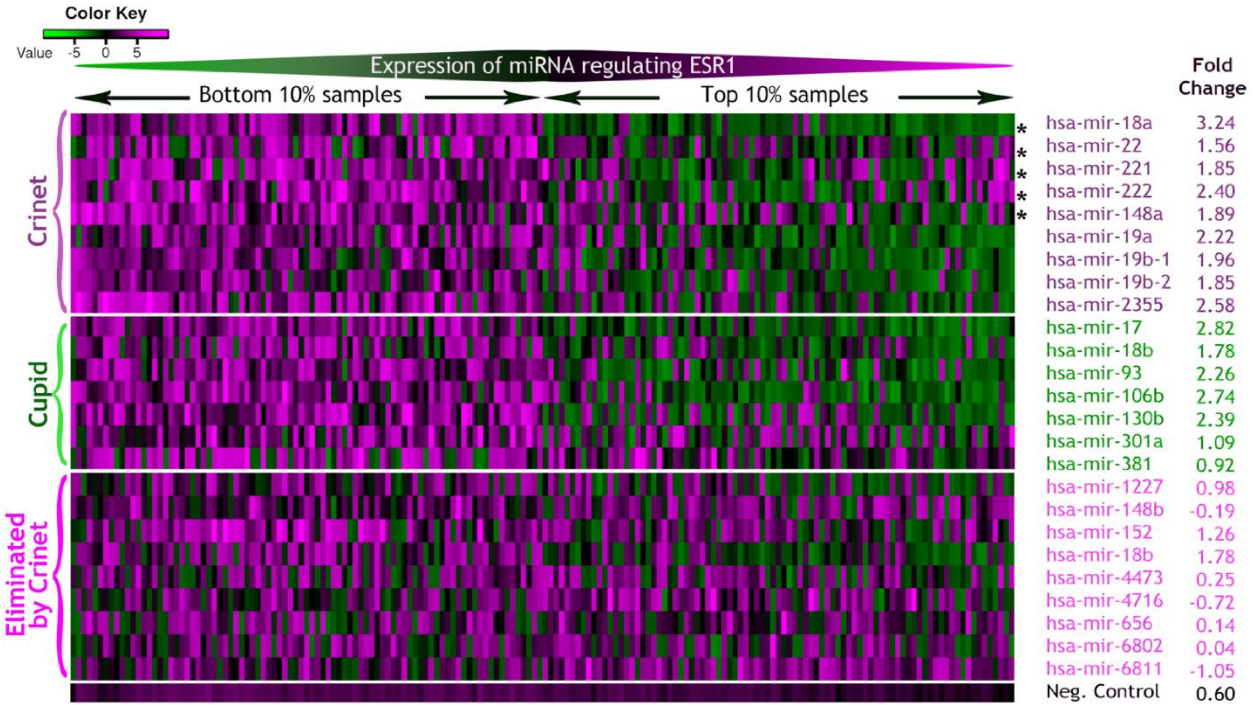


*Figure 2 Heatmap showing protein expression of ESR1 for the top and bottom 10% ranked samples with respect to miRNA expression. Protein expression is shown for the top and bottom 10% samples ranked with respect to miRNA expression regulating ESR1 by Cupid-inferred, CRINET-inferred, CRINET-eliminated, and negative control along with the mean difference of log fold-change of protein expression for the bottom 10% with respect to the top 10% samples. Each row is independently ranked by miRNA expression. (*Common miRNA regulators with Cupid).*

Figure 2 also illustrates nine regulators eliminated by CRINET following expression correlation and sufficient abundance filtering of miRNA-target selection. These interactions did not show strong anticorrelation between miRNA and protein expression with respect to fold-change for the majority of miRNAs, showing the strength of the CRINET filtering approach. As a negative control, I added the average of 100 random miRNAs which were not selected as ESR1's regulator by Cupid and CRINET, and they exhibited a low fold-change.

To assess the accuracy for ceRNA inference, I used the Library of Integrated Network-based Cellular Signature (LINCS) [22] L1000 shRNA-mediated gene knockdown experiment in breast cancer cell line as in [6] and checked whether ceRNA interactions can predict the effects of RNAi-mediated gene silencing perturbations in MCF7 cells. Since CRINET starts with a high number of genes, it was not computationally feasible to run many tools with the used dataset. However, Hermes [6] runs any given ceRNA pair independently, therefore I ran Hermes for the genes in the knockdown assays using the same expression datasets and CRINET-inferred miRNA-target interactions. LINCS database is a rich resource having an expression change of nearly 1000 genes as a response to a silenced gene. When a gene is silenced then its ceRNA partner will be affected since more miRNA regulators will be available to suppress the ceRNA partner. Thus, given a ceRNA pair, the expression level should be lower in response to the silenced ceRNA partners in comparison to the genes that are not ceRNA partners. Based on this assumption, I evaluated the CRINET- and Hermes-inferred networks. The accuracy of this assessment is shown in Table 5. Since Hermes was not selective in terms of the number of ceRNA interactions by inferring many significant interactions, I evaluated several networks from Hermes till having a similar number of genes with CRINET

in the knockdown assessment. Based on these results, CRINET outperformed Hermes at predicting gene expression change of ceRNA partners for each timepoint and overall accuracy.

| | 96h Timepoint | 144h Timepoint | Overall |
|---|---|---|---|
| Crinet | 48/77 ≈ 62% | 44/77 ≈ 57% | 92/154 ≈ 60% |
| Hermes 1.run | 412/897 ≈ 46% | 441/913 ≈ 48% | 853/1810 ≈ 47% |
| Hermes 2.run | 199/466 ≈ 43% | 217/481 ≈ 45% | 416/947 ≈ 44% |
| Hermes 3.run | 109/269 ≈ 41% | 121/273 ≈ 44% | 230/542 ≈ 42% |
| Hermes 4.run | 65/143 ≈ 45% | 68/144 ≈ 47% | 133/287 ≈ 46% |
| Hermes 5.run | 33/69 ≈ 48% | 39/71 ≈ 55% | 72/140 ≈ 51% |

Analysis to check the accuracy of inferred ceRNA interactions using LINCS-L1000 shRNA-mediated gene knockdown experiment in breast cancer cell line. Based on the ratios of gene expression fold-change following the knockdown of its ceRNA partners to following the genes that are not its ceRNA partners for each perturbagen ceRNA, the accuracy of a ceRNA network was accepted as the percentage of ceRNAs whose ratios were smaller than 1 with respect to all ceRNAs. We calculated the accuracy separately for each different timepoint (96h & 144h) and combined timepoints as the overall. Hermes's x.run had $10^{-(x+1)}$ for the significance of the common miRNA size and $10^{-(x+2)}$ for the significance of conditional regulation.

*Table 4 Evaluation of the accuracy of CRINET- and Hermes-inferred ceRNA interactions based on the shRNA-mediated gene knockdown experiment.*

To analyze the biological significance of the inferred ceRNA network, I applied enrichment analysis for the inferred ceRNAs. The inferred ceRNAs were significantly enriched in 398 GO terms from biological process ontology and 39 KEGG pathways. To associate enriched terms to broader categories, I analyzed GO Slim terms. Inferred ceRNAs were mostly involved in biological processes including immune system process, cell differentiation, cell death, cell cycle, response to stress, and cell-cell signaling. These suggest that ceRNA interactions could have an important role in biological processes in cancer.

To check if the inferred ceRNAs were associated with the cancer-related genes, I collected 3078 known cancer genes obtained from public databases. In the inferred ceRNAs, there was a significant overlap (hypergeometric test p-value 9.10−105) between the known cancer genes and the inferred ceRNAs having 789 out of 3078 known cancer genes in the inferred network. When I repeated the same analysis for non-inferred genes, they did not show a significant p-value (almost 1). I also had 54 breast cancer-related genes, 14 out of 54 breast cancer-related genes were inferred in my network. These results indicated that inferred ceRNAs were significantly associated with known cancer genes.

To evaluate the inferred ceRNA groups, I performed enrichment for four of 81 inferred ceRNA groups that have more than three members. CeRNA groups had a significant overlap with the cancer-related genes (hypergeometric p-value <0.0006). CeRNA groups were significantly overlapped with known cancer-related genes and significantly enriched in biological processes suggesting that ceRNA groups could have important roles as a group in cancer including the immune system and cell repair.

I applied more evaluation for inferred ceRNA interactions and groups, and they are in the published paper (See Appendix).

## Aim 2. Develop a node classification framework encoding multiomics features and convolutions on multiplex graph-structured data

I developed a node classification framework that integrates multiomics data with node associations utilizing graph convolutional neural networks [44]. I applied this framework specifically to the cancer subtype prediction problem deducing the signals from different biological datasets applying convolution on patient similarity networks, and utilizing the features and associations from multiple datatypes to predict subtypes of cancer patients.

### Aim 2. a. Background

Cancer is one of the deadliest diseases in the world [1], and associations in that complex disease are not fully demystified. Cancer patients show different characteristics in terms of the progression of disease and response to treatment [36]. To pave the road towards precision medicine, patients with highly similar biology ought to be grouped into cancer subtypes. Various biological datasets from cancer tissues have been generated to better characterize cancer biology. Utilizing the high dimensional biological datasets in public databases, computational approaches to discover subtypes of various cancers have been developed [2][3][4][5]. Several of the cancer subtype prediction studies rely only on one type of biological dataset [3][6][7][9]. However, each of these datatypes captures a different part of the underlying biology, and thus developing integrative computational methods has been an important problem in bioinformatics.

Breast cancer is one of the most common and the deadliest cancer among women in the world. There are three important receptors in breast cancer patients, and there are three basic therapeutic groups mainly based on the receptors: Estrogen receptor (ER) positive or progesterone receptor (PR) positive, HER2 amplified group, and triple-negative breast cancer (TNBC). These groups have different treatment options to block the growth in the cancer cells: endocrine therapy, therapeutic targeting of protein, and chemotherapy. Even receptors are very impactful on the cancer subtypes, they are not solely sufficient to classify a patient. In addition to the receptors, other features such as race, age, and some mutations play important roles in breast cancer. For example, 65% of tumors developing in women aged less than 50 are ER-positive, and this increases to 80% in women aged more than 50 [60]. TNBC has increasing events in patients with germline BRCA1 mutations of African ancestry [61].

On the other side, genomic datatypes are informative for differentiating subgroups in cancer. In 2009, Parker et al. explored gene expression datatype for breast cancer and found a clear difference in the expression of 50 genes [13]. Leveraging from gene expression dataset and some clinical features, they introduced breast cancer molecular subtypes, called PAM50 subtypes. This study has important insights about breast cancer subtypes and shows the importance of genomic data for cancer subtype classification. TCGA Project is an important program that generates high-dimensional multiomics data for the same cohort. TCGA Project generated data for thousands of patients from 33 different cancer types having more than 2.5 petabytes of data. In 2012, the TCGA project published a study analyzing breast cancer subtypes and their associations with single datatypes [14]. Single datatypes were analyzed and included in a single analysis obtaining subtype-specific patterns in that datatype. Moreover, they did a comprehensive analysis using six different datatypes. Even they support the importance of gene expression as the PAM50 model does [13], there were differences between patient groups obtained from different datatypes. Some other cancer types such as brain cancer and kidney cancer have also defined subtypes that were generated mainly from one datatype such as microRNA expression, DNA methylation, and mutation or an integrative approach [3][4][13][14]. Even though we have important signals from clinical features and genomic

features, there is no clear separation from either side to decide the subtype of a patient. While we are getting more samples and datatypes to analyze, it is important to integrate all the available datatypes properly with advanced approaches to understand differences in the characteristics of cancer patients.

Recently several groups have developed computational tools to integrate different types of datasets to discover cancer subtypes such as iCluster [33], Similarity Network Fusion (SNF) [34], and PINSPLUS [35]. Those studies focus on unsupervised multiomics data integration without the additional patient labels. Especially going towards personalized medicine, phenotypes and traits of the patients are more widely collected, and those will help to develop supervised approaches with higher accuracy.

Graphs are utilized to store multiomics datasets. Real-world graphs (networks) are unstructured because of the different number of neighbors and lack of neighbor positions. Grid-like data allows us to utilize recent deep learning-based architectures such as Convolutional Neural Network (CNN). However, for the data that does not live on the grid, these architectures do not fit. Some methods are emerged encoding every node into a fixed low-dimensional vector, called *embedding,* representing the position and the local relationships in the network [41][42][43]. However, these embedding methods do not utilize the node features of the local neighborhood that we have plenty of, thus, current methods are started to be replaced with more advanced methods such as GNN. GNN models generate the node embeddings from a given network to represent its node features and local network structure. As compared to traditional embedding methods [40][41][42][43], GNN-based models consider the local neighborhood and integrate any additional features to the model. GNN models have the main difference in how the features are aggregated from the local structure. Graph Convolutional Networks (GCN) is one of the most popular GNN that uses a modified aggregation involving self-nodes with normalization across neighbors [44]. GCN models learn the data by performing convolution in graphs, considering one-hop local neighbors with equal contribution, and encoding the local structure of the graph. Stacked layers involve recursive neighborhood diffusion considering more than one-hop neighborhood. GNNs have recently been applied to biological problems such as cancer type/subtype prediction and drug response prediction [10][11][12][26][27]. However, GNN-based models are mostly applied on a single graph or had some limitations for integrative approaches.

There are some studies applying convolution approaches on graph-structured data. In [50], cancer type prediction of patients from 33 cancer types or normal types (collecting all normal samples from available cancer types as one group) was analyzed using GCNs. As the input network, they used gene coexpression or protein-protein interaction networks utilizing only gene expression datatype. Similarly, in [51], the authors propose a GCN model using only gene expression and utilizing association in gene coexpression network or protein-protein interaction network. They integrate global features from the data and predict cancer taxonomy labels obtained from TCGA (28 molecular subtypes from 33 cancer types). Those studies use only one datatype, thus, missing the information of multiple graphs. MOGONET is a supervised multiomics integration framework using GCNs with a patient similarity network for mRNA expression, DNA methylation, and microRNA expression separately [32]. The obtained labels are used as an input to View Correlation Discovery Network that uses latent correlations across different datatypes to get the final prediction. MOGONET gets the label independently from three different datatypes, then integrates them, but it does not consider the associations between datatypes, just considers the obtained prediction from separate GCN models. Also, MOGONET integrates three datatypes, thus could miss some signals from other available datatypes. Given networks are important to get the relations within datatypes, but integrating the network with node features from other datatypes is also crucial to get the additional signals between different datatypes. To utilize more, it is needed to get the important signals from a given network with associations from within and between multiomics datatypes.

To address the aforementioned challenges, I developed a computational tool named SUPREME, a subtype prediction methodology, integrating multiple types of biological datasets using GCNs. SUPREME convolves the multiomics features on each graph-structured data built using different datatypes and assumes that patients with a similar local neighborhood are likely to belong to the same subtype, where the similarity is based on a specific datatype's features. SUPREME encodes the relations from each available biological datatype and gets patient representations incorporating patient similarities and multiomics features. In addition, it integrates datatype-specific representations with raw multiomics features to capture the complementary signals between different datatypes along with local and global signals.

SUPREME was applied to the breast cancer dataset from TCGA to predict cancer subtypes of patients. Preliminary results showed that SUPREME had high prediction power and outperformed the baseline and state-of-the-art models. SUPREME outperformed multilayer perceptron (MLP) that does not consider node associations at all, showing the importance of GNN-based approaches. SUPREME results were consistent when excluding single graphs and single feature sets, thus, making sure that SUPREME will work robustly for the outsourced datasets that are missing some of the datatypes. I used the ground truth from PAM50 subtypes that mainly leverage gene expression. Unsurprisingly, gene expression-based features were observed as the most significant feature set. But, surprisingly, the copy number aberration (CNA) model was the most impactful on the SUPREME results. Even the CNA features do not show a significant impact on prediction results, exclusion of node embeddings learned from the CNA-based graph dropped the accuracy concretely. SUPREME is extendable to any number of datatypes to integrate. SUPREME integrates multiomics data with common and complementary signals. Biologists and clinicians can predict the subtype of any cancer patient by using SUPREME. SUPREME-generated models had consistently significant survival significance, supporting that SUPREME-predicted outcomes could be more meaningful as compared to the ground truth.

### Aim 2. b. Methodology

SUPREME is a computational tool for cancer subtype prediction integrating multiple biological datatypes using GCNs (Figure 3). Briefly, the first step is the data preparation step. In the second step, SUPREME extracted features from each seven datatypes, and using those feature sets, it generated the patient similarity networks for each datatype. In the third step, using the obtained networks and features, SUPREME ran the GCN model for each datatype separately and generated the embeddings from each. In the last step, SUPREME predicted the cancer subtype of patients integrating individual network representations and global multiomics features. In the following part, I will explain each step of the SUPREME pipeline as detailed.

1. Data Preparation: To prepare SUPREME data, I collected all available datatypes from the TCGA data portal (available at https://www.cancer.gov/tcga) only excluding protein expression, which has limited features for samples. Thus, I obtained the datasets from the TCGA project including gene expression, microRNA expression, DNA methylation, CNA, mutation, and clinical features for a total of 1022 breast tumor samples that have a PAM50 subtype label. I preprocessed each datatype separately obtaining normalized expression values for gene expression as FPKM, for miRNA expression as RPM, and obtaining gene-centric values from segmented CNA using the CNTools [47]. For DNA methylation, I converted the probe-level data to gene-centric separately for 450K-probed and 27K-probed patients then combined the common genes. To be able to obtain coexpression data, I ran the WGCNA tool [48] using gene expression datatype and obtained module eigengenes as coexpression features. For mutation data, I collected

masked somatic mutation from Simple Nucleotide Variation as gene-centric across samples. I collected age, menopause status, race, tumor stage, tumor status, and ER/PR/HER2 status of patients as clinical features.

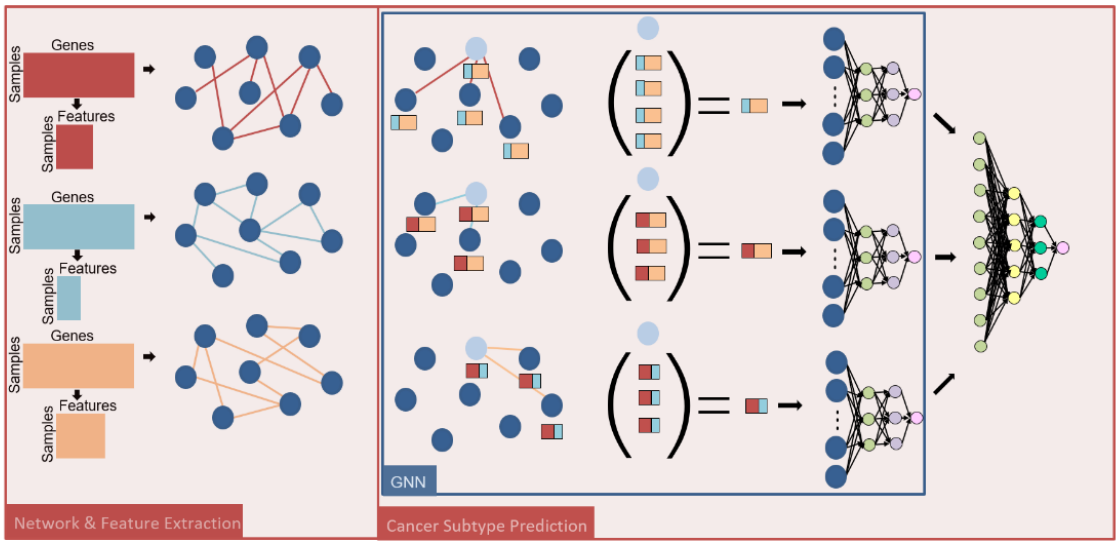


*Figure 3 SUPREME pipeline*

2. Network Generation & Feature Extraction: SUPREME incorporates seven different datatypes, so I generated corresponding feature sets and patient similarity networks for each seven datatypes, namely gene expression, microRNA expression, DNA methylation, CNA, mutation, coexpression, and clinical. I excluded features that are missing for most of the samples. For the datatypes with a small number of features (less than 500 features), namely miRNA expression, clinical and coexpression datatypes, I included all features. For the datatypes with a large number of features (more than 500 features), I used a feature selection algorithm [49] to have a manageable number of features (specifically 200, 500, 1000, and 1500). Also, I kept the top 2500 edges in the patient network using Pearson correlation for gene expression, CNA, DNA methylation, microRNA expression, and coexpression features; using the "Gower" metric from the daisy R package for clinical features; and using Jaccard distance for binary mutation features.

3. Single Model Generation: To integrate multiomics features with networks associations, I applied graph convolutions to the local neighborhood of patients based on each patient similarity network from seven different datatypes. In that way, I measured the performance of each patient similarity network and compared the results.

GNN-based convolution methods have different aggregation for local structure, and the popular GCN model of Kipf and Welling [11] involves self-edges to convolution and normalizes aggregated features across the number of local neighbors.

Let's call an undirected graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of $n$ nodes, i.e. $\mathcal{V} = \{v_1, v_2, \dots v_n\}$, and $\mathcal{E}$ is a set of edges between nodes where

$$(v_i, v_j) \in \mathcal{E} \qquad (4)$$

when $v_i \in \mathcal{V}$ and $v_j \in \mathcal{V}$ and $v_i$ and $v_j$ have an association based on the graph $\mathcal{G}$. Since the graph $\mathcal{G}$ is undirected,

$$(v_i, v_j) \in \mathcal{E} \leftrightarrow (v_j, v_i) \in \mathcal{E}. \quad (5)$$

Adjacency matrix $\mathcal{A} \in \mathbb{R}^{nxn}$ is:

$$\mathcal{A} = \begin{cases} 1 & if\ (v_i, v_j) \in \mathcal{E}\ or\ i = j \\ 0 & otherwise \end{cases} \quad (6)$$

The input for a GCN model is a feature matrix $\mathcal{X} \in \mathbb{R}^{nxk}$ and an adjacency matrix $\mathcal{A} \in \mathbb{R}^{nxn}$, where $n$ is the number of samples and $k$ is the feature size. The iteration process is defined as:

$$H^{(l+1)} = \sigma(\mathcal{D}^{-\frac{1}{2}} \mathcal{A}\, \mathcal{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (7)$$

where $\mathcal{D}$ is the degree matrix of nodes where the degrees appear in the diagonal, $H^{(l)}$ is the activation matrix in the l[th] layer, $W^{(l)}$ is the trainable weight matrix in the l[th] layer and $\sigma$ is the activation function. Learning neighborhood features is achieved by an aggregation function such as average or sum. I used the GCN model of Kipf and Welling [11] involving self-edges to convolution and scale the sum of aggregated features across the local neighbors.

SUPREME setup for the single model is as follows: there is one network (patient similarity network extracted in the second step) for each datatype modality, so totally there are 7 single networks. All networks have breast cancer patients as nodes, and edges are obtained based on the patient similarities based on the specific datatype. For instance, let us consider the gene expression datatype as $\mathcal{G}$ described above. There is a gene expression-based patient similarity network from the second step. This network has patient similarities based on the top high correlation from the gene expression data.

As node features, SUPREME combined the features extracted in the second step from all the seven datatypes. Features of $v_i$ is denoted as $x_i \in \mathbb{R}^k$ where $k$ is the total feature size. So, the stacked feature matrix $\mathcal{X} \in \mathbb{R}^{nxk}$ is:

$$\mathcal{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (8)$$

The local 1-hop neighborhood of a node $v_i$ is $\mathcal{N}_i = \{v_j : (v_i, v_j) \in \mathcal{E}\}$ that includes the set of nodes having an association with the node $v_i$. Feature aggregation on the local neighborhood of each node is done by multiplying $\mathcal{X}$ by the $nxn$-sized scaled adjacency matrix $\mathcal{A}'$ where

$$\mathcal{A}' = \mathcal{D}^{-\frac{1}{2}} \mathcal{A}\, \mathcal{D}^{-\frac{1}{2}} \quad (9)$$

and

$$\mathcal{D} = \sum_j \mathcal{A}_{ij} \quad (10)$$

representing the degree matrix of nodes as a diagonal. In SUPREME's setup, I used 2-layer GCN and the form of the forward model gives the output $\mathcal{Z}$ where

$$\mathcal{Z} = softmax(\mathcal{A}'\, ReLU(\mathcal{A}' \mathcal{X} \mathcal{W}^{(1)}) \mathcal{W}^{(2)}) \quad (11)$$

and $\mathcal{W}^{(1)} \in \mathbb{R}^{kx\hbar}$ and $\mathcal{W}^{(2)} \in \mathbb{R}^{\hbar xc}$ are the trainable weights for the first and second layers, respectively, where $\hbar$ is the hidden layer size and $c$ is the number of classes to predict (namely, Basal-like, Luminal A, Luminal B, Her2-enriched, and Normal-like, with $c$ = 5). The loss function was calculated by cross-entropy error. Adam optimization as the state-of-the-art for the stochastic gradient descent algorithm was used

[62], and dropout was added for the first GCN layer. Stopping criteria was used with the patience of 30 forced to have at least 200 epochs.

In that way, I generated seven different single models employing the features from all the datatypes using patient relations from one single datatype. I evaluated those single models separately. Moreover, I obtained the embedding from each datatype-specific model to use in the next step.

4. Integrated Model Generation: To investigate the integration of the single models, SUPREME concatenated embeddings from each single model and trained them in an integrated model with a fully connected Neural Network (FCNN). In that way, associations between different single models are captured, too. To analyze the impact of each single model, I ran the FCNN model with all the combinations of the single models ($2^{\#data\ type} - 1 = 2^7 - 1 = 127$ different models). To check the impact of the original node features (raw features), I also concatenated the features extracted in the second step to the embeddings of the single models. I repeated this for each combination of the single models. For the integration model, I used a 2-layered FCNN with cross-entropy loss. Adam optimization [62] was included, and dropout was added after the first layer.

I evaluated the prediction results using evaluation metrics such as accuracy and weighted F1 scores for the integrated models. To see the impact of the features from a single datatype, I ran the whole methodology excluding the features separately. I checked the impact of individual models on the integration by evaluating the integrative models that exclude each single model. I compared the results with some baseline methods, state-of-the-art machine learning algorithms, supervised and unsupervised tools for cancer subtype differentiation.

### Aim 2. c. Preliminary Results

I tested SUPREME on breast cancer, which is the deadliest cancer type among women. Gene expression, microRNA expression, DNA methylation, CNA, mutation, and clinical features of breast cancer were downloaded from the TCGA. 1022 breast tumor samples were having PAM50 labels were included in the analysis, and these labels were used as the ground truth labels. The sample distribution of each subtype is as follows:

- Basal-like (n=172)
- Luminal A (n=538)
- Luminal B (n=195)
- HER2-enriched (n=78)
- Normal-like (n=39)

I generated clinical features having 29 features including age, race, menopause status, tumor stage, tumor status, estrogen receptor status, progesterone receptor status, and HER2 status. Categorical clinical features are used as one-hot encoded. I generated coexpression features running the WGCNA tool [48], and obtained module eigengenes from gene expression datatype, and used 16 module eigengene values as coexpression features. I also generated microRNA expression features using 343 microRNAs that have $\geq 0.01\ RPM$(Reads per million mapped reads) value across $\geq 30\%$ of the entire cohort.

Since there is a large number of features for DNA methylation, CNA, gene expression, and mutation datatypes, SUPREME used a feature selection algorithm [49] that uses class labels to select important features and obtained the top 200, 500, 1000, and 1500 features. Based on the performance of the single models, SUPREME used the top 200 features for mutation, top 1000 features for methylation and gene expression, and top 500 features for CNA. SUPREME split 20% of the total samples as a test set and the

test set was never used except for the final evaluation of the tool. This splitting was done as stratified, i.e. keeping the same ratio of the subtype labels in the original data for each split. The remaining 80% of the samples were used for training (60%) and validation (20%), and those splits were randomly selected for each run as stratified. For any run, SUPREME repeated an evaluation metric (accuracy and weighted F1 score) 10 times for each hyperparameter combination (0.1 and 0.01 for learning rates; 64, 128, 256, and 512 for hidden layer sizes; and 0 for weight decay.) and used the hyperparameter combination with the best median weighted F1 score on the validation data to generate the final model. Finally, the final model was evaluated on the test data.

There are seven single models from each different datatype and 127 models as integrated. To compare their performance, I obtained the accuracy and weighted F1 scores in Table 5. Most of the single models give a weighted F1 score higher than MLP with all features. Some combination of integrated models outperformed MLP+FS with the best case having 91% as a weighted F1 score. Clinical- and mutation-based single models did not give results as good as others. The CNA-based model gave the best result.

| | F1 Score | Accuracy |
|---|---|---|
| SUPREME | 0.91±0.010 | 0.91±0.011 |
| SUPREME-CLI | 0.78±0.015 | 0.74±0.039 |
| SUPREME-CNA | 0.87±0.015 | 0.87±0.015 |
| SUPREME-COE | 0.84±0.015 | 0.82±0.014 |
| SUPREME-EXP | 0.85±0.018 | 0.84±0.017 |
| SUPREME-MET | 0.85±0.017 | 0.84±0.016 |
| SUPREME-MIR | 0.77±0.016 | 0.77±0.012 |
| SUPREME-MUT | 0.81±0.013 | 0.81±0.012 |
| MLP | 0.83±0.012 | 0.82±0.010 |
| MLP + FS | 0.89±0.015 | 0.89±0.015 |

*Table 5 Single datatype evaluation*

*Accuracies and weighted F1 scores from seven datatypes namely clinical (SUPREME-CLI), CNA (SUPREME-CNA), coexpression (SUPREME-COE), gene expression (SUPREME-EXP), DNA methylation (SUPREME-MET), microRNA expression (SUPREME-MIR), and mutation (SUPREME-MUT), and also MLP with all raw features, MLP with SUPREME-selected features (MLP+FS), and the best of integrated SUPREME models (SUPREME).*

To evaluate SUPREME models, I concatenated embeddings from all the combinations of the single models. For each of the models, SUPREME ran with and without raw multiomics features as concatenated to the embeddings. The model performance was similar for the overall model distribution when models were run with or without raw features (Figure 4). When we check the F1 scores for each different model, they give similar results. However, models when raw features were added did not have lower than 83.5%, while three cases were giving lower values for one-datatype models (specifically 0.748 when using embedding of clinical datatype only, 0.803 when using embedding of mutation datatype only, and 0.808 when using embedding of microRNA expression datatype only).

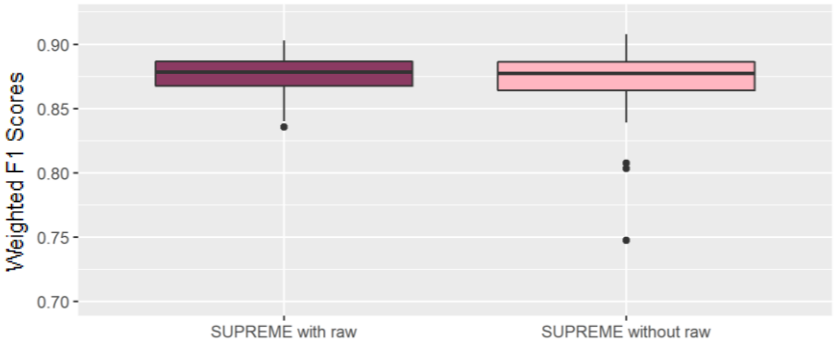


*Figure 4 Boxplot of weighted F1 scores SUPREME results with raw features integrated to embeddings, and without raw features integrated to embeddings.*

Figure 5 shows each weighted F1 score for different models with vertical lines separating one-datatype models, two-datatype models, and so on. Models using more datatypes gave consistently higher F1 scores especially from 1-datatype to 5-datatype models, showing that using limited models misses some important signals. After four datatypes, some datatypes seem not effective (Figure 6). As an overall comparison, single datatype integration was significantly lower than all integration models (p value = 0.0015) showing the importance of integration of single models (Figures 5 and 6).

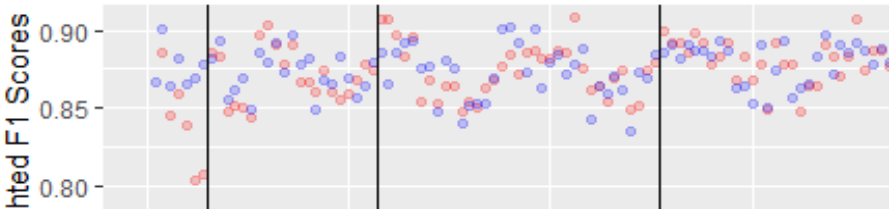


*Figure 5 Scatterplot of weighted F1 scores for each different combination of models (separately for with and without integrating raw feature to embeddings). Vertical lines separate models with 1-, 2-, 3-, 4-, 5-, 6-, and 7-datatypes. [Pink: without raw features, Purple: with raw features]*

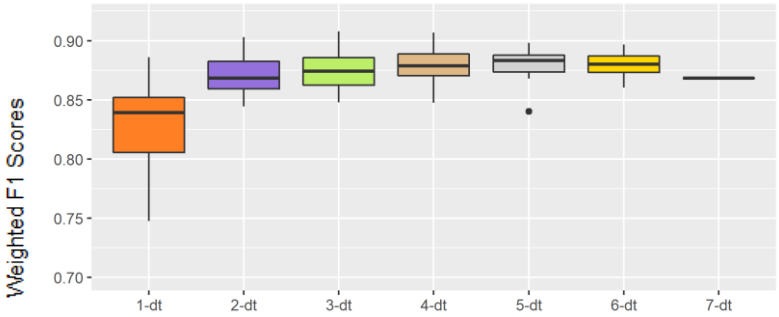


*Figure 6 Boxplot of weighted F1 scores for models integrating different number of datatypes when no raw features are integrated to embeddings (Significance is with respect to SUPREME all models). X-axis shows the number of datatypes used for the integration.*

Since simple MLP does not incorporate graph structure at all, I ran an MLP model using all the multiomics features to see the improvement of the GNN-leveraging model. I repeated the MLP model run for all the combinations similar to SUPREME using both the SUPREME-selected features (MLP+FS in Figure 7) and all multiomics features (MLP in Figure 7). Based on those results, our model outperforms the MLP model with more consistent results.

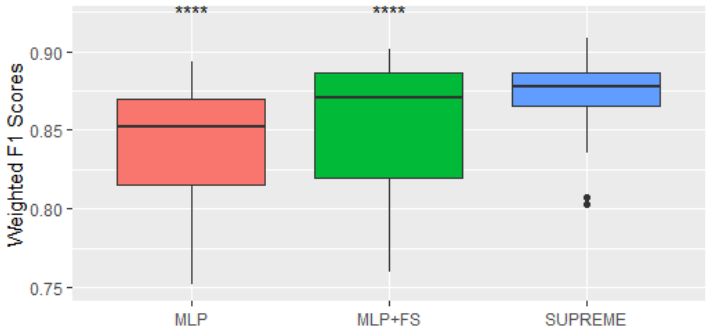


*Figure 7 Boxplot of weighted F1 scores in MLP as compared to SUPREME (MLP: all multiomics features, MLP+FS: SUPREME-selected multiomics features) (Significance is with respect to SUPREME and denoted by *, and more * means more significant).*

I compared SUPREME results with supervised baseline models, namely MOGONET, Random Forest (RF), Support Vector Machine (SVM: SVMl for the linear kernel, SVMr for the radial kernel), and XGBoost. I ran the algorithms for all the integration combinations and also tried all multiomics features and SUPREME-selected features separately. SUPREME outperformed all the methods with significant differences in each (Figure 8). It was obvious that SUPREME gives more consistent results for different combinations of integrated models.

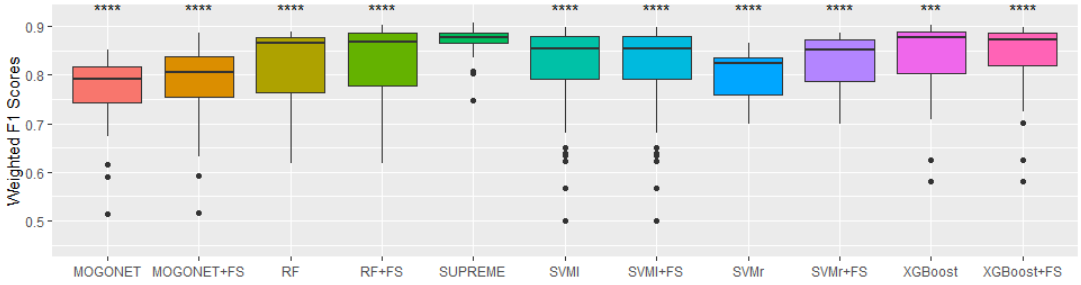


*Figure 8 Boxplot of weighted F1 scores in supervised baseline models as compared to SUPREME (RF: Random Forest, SVMl: Support Vector Machine with linear kernel, SVMr: Support Vector Machine with radial kernel) (x: any algorithm with all multiomics features, x+FS: x algorithm with SUPREME-selected multiomics features) (Significance is with respect to SUPREME and denoted by *, and more * means more significant).*

Since there is embedding from different datatypes, I ran separate models to check the effect of each datatype by excluding features and patient networks from the individual datatype. As in Figure 9, we did not see a big difference when any datatype was excluded except the CNA-based model. CNA-based single model gave the higher accuracies, and excluding it from integration dropped the results significantly. SUPREME also excluded the features from node features of the single models and finally-integrated raw

features and ran all the models from the beginning to see the effect of the datatype-specific features on the final results. Among them, gene expression features seem to be very effective on the results since the exclusion of those features dropped the evaluation accuracy drastically, while clinical features have the least effect (Figure 10). I also checked the model behavior when I excluded node features from one specific graph and excluded the single model of that graph from the integration, and, we got very similar results to the previous results where only node features are excluded (Figure 10 and 11). This result suggests that with the exclusion of both features and networks from one datatype (like excluding the datatype from the entire analysis), the results are not affected much except the gene expression dataset. This is expected where the ground truth mainly depends on the gene expression features.

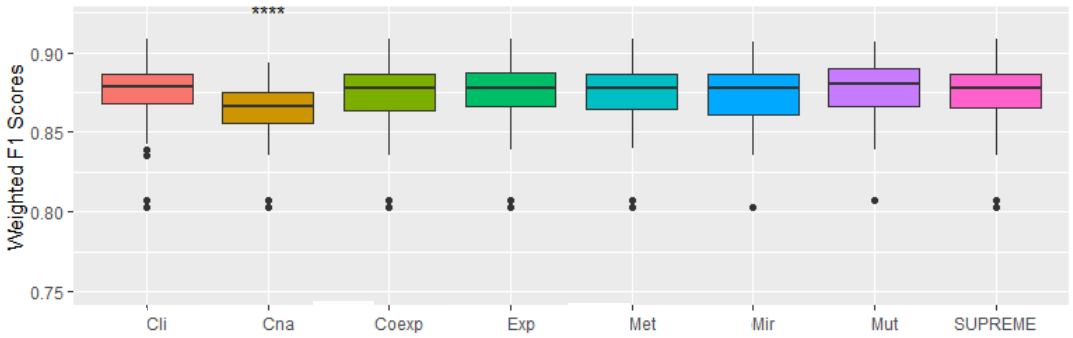


*Figure 9 Boxplot of weighted F1 scores in SUPREME when excluding one single model from integration. X-axis keeps the excluded datatype (Significance is with respect to SUPREME) [Cli: Clinical, Cna: Copy Number Aberration, Coexp: C , Exp: Expression, Met: DNA Methylation, Mir: MicroRNA expression, Mut: Mutation].*

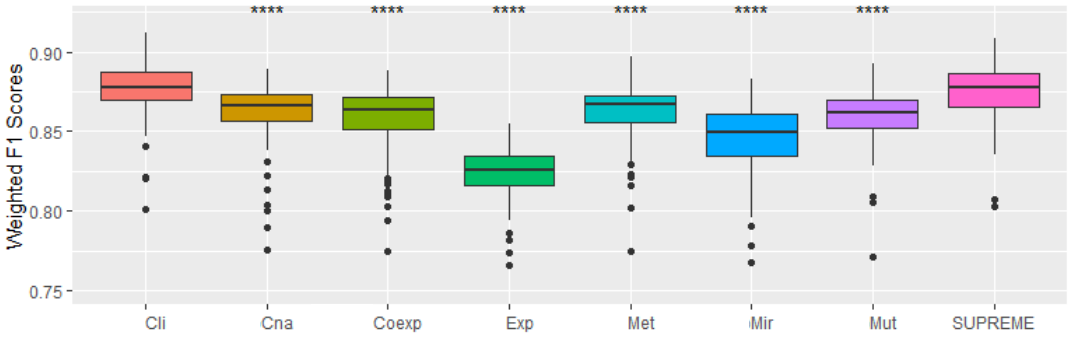


*Figure 10 Boxplot of weighted F1 scores in SUPREME when excluding node features from one datatype for single models. X-axis keeps the excluded datatype [Cli: Clinical, Cna: Copy Number Aberration, Coexp: Coexpression, Exp: Expression, Met: DNA Methylation, Mir: MicroRNA expression, Mut: Mutation].*

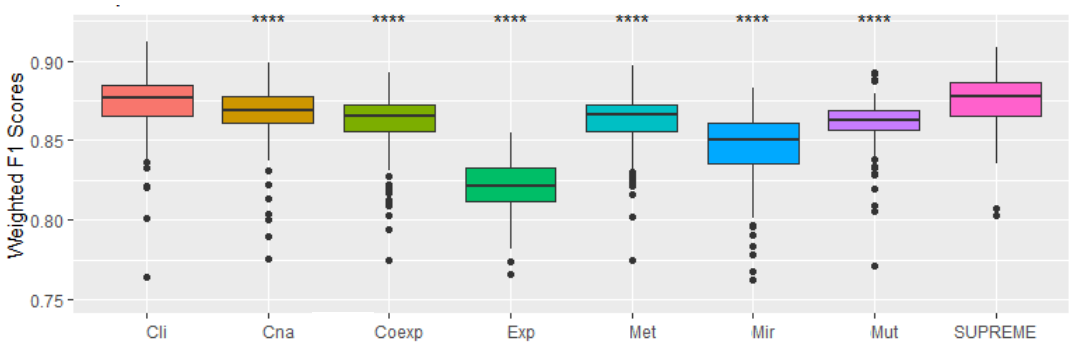


*Figure 11 Boxplot of weighted F1 scores in SUPREME when excluding node features and embeddings from one datatype through all workflow at the same time. X-axis keeps the excluded datatype (Significance is with respect to SUPREME) [Cli: Clinical, Cna: Copy Number Aberration, Coexp: Coexpression, Exp: Expression, Met: DNA Methylation, Mir: MicroRNA Expression, Mut: Mutation].*

I also obtained the clusters with the number of classes as five from popular unsupervised methods of cancer subtype prediction, and compare the survival differences for obtained clusters (using the log-rank test to obtain a p-value to measure the difference between survival curves). According to the results in Figure 12, SUPREME-inferred classes have significant differences in survival rates, while other unsupervised algorithms except PINSPLUS give a wider range. PINSPLUS gives consistent results for different combinations like SUPREME, but the consistent part has no significant survival differences. SUPREME consistently gives the significantly different groups in terms of survival, and this could suggest that SUPREME inferred the cancer subtype classes utilizing multiple datatypes and node associations, and in that way, SUPREME could demystify the undiscovered characteristics in cancer subtypes that cause significant survival differences.

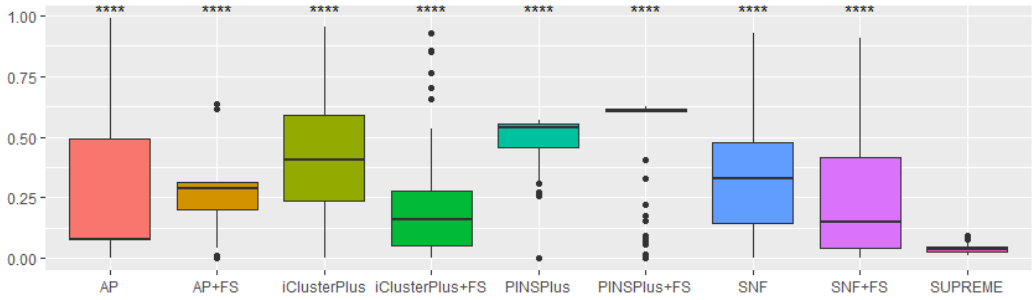


*Figure 12 The log rank p-values obtained from survival analysis for predicted subtype classes for popular unsupervised methods of cancer subtype differentiation. (Significance is with respect to SUPREME) [SNF: Similarity Network Fusion, AP: Affinity Propagation clustering].*

In future work, I will apply SUPREME to other cancer types.

## Aim 3.Develop a GNN-based architecture with convolutions on multiplex heterogeneous graphs with attention mechanisms

I will develop a novel GNN-based architecture integrating multiple graphs having adjustable node relations where the given graphs could be more than one node type. This tool will be applicable to problems of node classification and link prediction problems.

### Aim 3. a. Background

CNN has been successfully applied to grid-like data such as images. However, they were not applicable for the unstructured data such as graphs, where the number of neighbors and their positions are not fixed. There are some recent attempts to extend neural networks to operate on graph-structured data such as embedding methods that are already mentioned in the *Background section of Aim 2*. GNN considers the association between nodes integrating with all possible node features, thus, GNN started to take place of current methods. GNN generates low-dimensional graph representation, called *embedding*, of nodes from a given network and represents its node features and local network structure. Considering the limitations of the current GNN-based architectures as mentioned in *Aim 2's Background section*, it is worthy to develop a novel GNN architecture that directly operates on multiplex graphs.

The attention mechanism is a popular approach using a more advanced aggregator considering the relation weights between nodes, and even providing the possibility for multiple heads for attention [46]. Attention-based models achieved impressive performance on sequence-based tasks like in Natural Language Processing. Attention-based mechanisms are recently used on GNN-based architectures and have good improvement. Velickovic et al. [45] proposed Graph Attention Network (GAT), which gives different importance to local neighbors. Adjusting the associated neighbor's impact and improving the given neighborhood, inferred weighted associations could be utilized more. Thus, false-positive interactions in the input networks will be also handled by assigning low weights. However, attention mechanisms need to be integrated into GNN-based architectures properly. In [52], authors use attention mechanisms with GCN models, aggregating node features from neighbors using multiple graphs with an attention mechanism. In [53], since GAT cannot assign different weights for neighbors depending on the edge type, they do clustering using GCN with attention mechanism, reconstructing node features and network structure. In [54], with attention-based aggregation, they learn different representations for each datatype. Then, using those representations, they obtain global node representation and use that final representation for node classification or link prediction problems. Since embedding is regularized, and regularization between different network embedding encourages them to be similar, global node representation could miss the complementary signals from the raw features and associations. Thus, attention-based models are still under-explored, and applying attention-based mechanisms on GNN-based models will be important especially considering different relation types separately.

To have all these properties on one architecture, I will develop a multiplex heterogeneous GNN-based architecture with attention mechanisms. Being *multiplex*, it will integrate multiple graphs. Instead of fusing different representations into one, it will enable to have *different impacts* for different relation types between the same nodes, so the local neighbor's impact will be adjusted with *attention* mechanisms for each relation type separately. Being *heterogeneous*, it will improve the utilization from node associations simultaneously considering multiple node types. This novel multiplex heterogeneous GNN architecture will utilize multiomics features and multiple types of associations, where the associations might come from multiple graphs of multiple node types.

## Aim 3. b. Methodology

I will develop a GNN-based architecture operating on multiplex heterogeneous networks with attention mechanisms. This approach will be applicable for node classification and edge prediction problems. The overall structure is briefly as follows: The first step is the data preparation step. In the second step, I will extract features and generate networks from each different datatype, that could belong to multiple node types. In the last step, using the obtained networks and features, I will run my GNN-based novel method and predict the outcome of interest integrating all weighted multi-relational networks.

**Application 1: A node classification framework**

Specifically, the cancer subtype prediction problem as a node classification problem will be applied. This problem and its background information are previously mentioned in Aim 2's Background. In this application, I will generate a cancer subtype prediction methodology integrating multi-relational networks from multiomics data using attention-based methods. I will utilize multiple patient similarity networks of different datatypes such as gene expression and CNA. In addition, I will consider the weight of each relation from different datatypes, so that I will capture the more/less significant relation and influence the prediction results considering those weights. I will get gene-based networks from multiple graphs such as coexpression network, protein-protein interaction network, and ceRNA regulation network (inferred by CRINET [56], as a novel gene interaction network). I will obtain node embeddings from the heterogeneous network. In that way, it is probable to have some additional signals to differentiate cancer subtypes using multiple node associations. Since the PAM50 assignment of TCGA is used as ground truth, and these labels are dependent on the very limited features from the patients, we may have a better subtype assignment using all the available multiomics data and utilizing local neighborhood associations from multiple types of nodes.

The methodology for cancer subtype prediction is as follows:

1. Data Preparation: I will obtain the datasets from breast cancer patients of the TCGA project including gene expression, microRNA expression, DNA methylation, CNA, mutation, and clinical features that have a PAM50 subtype label. I will preprocess each datatype separately. To obtain the coexpression network, I will run the WGCNA tool [48], and also I will use module eigengenes from gene expression datatype as node features. Also, I will run CRINET [56] to infer competing endogenous RNA networks to process for datatype generation.

2. Feature Extraction: Since I will integrate different datatypes, I will generate features and corresponding networks for each. For patient similarity networks, I will use seven datatypes namely gene expression, microRNA expression, DNA methylation, CNA, mutation, coexpression, and clinical datatypes. This tool will generate corresponding feature sets and networks similar to SUPREME. I will obtain module eigengenes from the WGCNA tool [48] as coexpression features. I will generate patient networks as weighted keeping the strength in local neighbors. In addition, I will keep multiple relations between patient networks reflecting each datatype's impact, but the impact from each datatype will be learnable.

In addition to the patient similarity network, I will generate gene-related networks such as protein-protein interaction networks, coexpression networks, and ceRNA regulation networks inferred by CRINET. I will use gene expression, DNA methylation, CNA, and some processed features from the ceRNA network as node features. Using multiple networks from genes and patients, this model will allow us to utilize directly from multiple graphs of different nodes.

3. Outcome Prediction: Using all datatypes along with their features and networks, I will run this novel GNN-based model to predict the cancer subtypes of patients considering the local neighborhood and

important features from all available datatypes. Having heterogeneous networks with genes connected to the patient networks, patient embeddings will be more informative. I will evaluate the prediction results using evaluation metrics such as accuracy and F1 scores. I will compare MLP, state-of-the-art methods, machine learning methods, and GNN-based architectures.

**Application 2: An edge prediction framework**

This study will be applicable for edge prediction problems such as drug-side effect prediction, drug-disease associations, drug-target associations, and drug repurposing. Specifically, I will work on the drug-side effect prediction problem.

Polypharmacy is a therapy where multiple drugs are given to the patients to improve the efficacy of diseases. However, some drug combinations have side effects when used together. It is not practical to check all drug-drug combinations to test side effects occurrence. Existing studies check drug-drug interactions and decide side effects existence as compared to the single drug usage. However, they do not say the type of side effect, that could potentially lead to proper treatment. To the best of our knowledge, there is only one study, called Decagon [26], that studies the drug-side effect link prediction problem. Decagon is a graph convolutional neural network model using Hamilton et al.'s approach [10] and uses protein-protein interactions, drug-protein interactions, and drug-drug interactions, labeling the edge types with the side effects. However, this study does not utilize attention mechanisms and does not utilize other multiomics data such as expression profiles.

The methodology for drug-side effect prediction will be very similar to Application 1 except for the loss function of prediction and data leveraged. For data preparation, I will collect drug-related data from Drug Bank[54], and drug similarities based on the PubChem database [63] with SMILES notation [29] to build a drug-drug network. I will get side effect data from the SIDER and OFFSIDES databases [64][65]. I will generate protein-protein interaction, drug-protein target interaction, and drug-drug interaction networks. Using drug features and networks obtained from databases, I will apply this GNN-based architecture for drug-side effect prediction and evaluate the predictions.

# 6. Timeline

| | | | 2022 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Project** | **Task** | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov |
| SUPREME | Evaluation | | | | | | | | | | | | |
| | Writing Manuscript | | | | | | | | | | | | |
| Project 3 | Literature Search | | | | | | | | | | | | |
| | Data Preparation | | | | | | | | | | | | |
| | Network & Feature Extraction | | | | | | | | | | | | |
| | ceRNA Integration | | | | | | | | | | | | |
| | Subtype Prediction | | | | | | | | | | | | |
| | Drug-based Network & Feature Extraction | | | | | | | | | | | | |
| | Drug-Side Effect Prediction | | | | | | | | | | | | |
| | Evaluation | | | | | | | | | | | | |
| | Writing Manuscript | | | | | | | | | | | | |

# References


[1] R. L. Siegel, K. D. Miller, and A. Jemal. (2016). Cancer statistics, CA Cancer J Clin, vol. 66, no. 1, pp. 7–30.

[2] R. G. W. Verhaak et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1, Cancer Cell, vol. 17, no. 1, pp. 98–110.

[3] H. Noushmehr et al. (2010). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma, Cancer Cell, vol. 17, no. 5, pp. 510–522.

[4] M. Baysan et al. (2012). G-cimp status prediction of glioblastoma samples using mRNA expression data, PLoS One, vol. 7, no. 11, p. e47839.

[5] T. Xu et al.,CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification,validation and visualization,Bioinformatics,vol.33,no.June,pp.3131–3133,2017.

[6] Vural, S., Wang, X., & Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches.BMC systems biology,10(3), 62.

[7] Youssef, Y. M., White, N. M. A., Grigull, J., Krizova, A., Samy, C., Mejia-Guerrero, S.,1,... Yousef, G. M. (2011). Accurate molecular classification of kidney cancer subtypes using MicroRNA signature. European Urology, 59(5), 721-730. doi:10.1016/j.eururo.2011.01.004

[8] TCGA GDC Portal, 2021. [Online]. https://portal.gdc.cancer.gov/

[9] H. Noushmehr et al., Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma, Cancer Cell, vol. 17, no. 5, pp. 510–522, 2010.

[10] Hamilton W, Ying Z, and Leskovec J, Inductive representation learning on large graphs, Advances in Neural Information Processing Systems, 2017, pp. 1024–1034.

[11] Kipf TN, and Welling M, Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016).

[12] Rhee S, Seo S, and Kim S, Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. arXiv preprint arXiv:1711.05859 (2017).

[13] Parker, Joel S., et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes." Journal of clinical oncology 27.8 (2009): 1160. Accurate Molecular Classification of Kidney Cancer Subtypes Using MicroRNA Signature [Youssef, 2011]

[14] Koboldt, D. C. F. R., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., ... & Iglesia, M. (2012). Comprehensive molecular portraits of human breast tumours. Nature, 490(7418), 61-70.

[15] Dhingra, P., Martinez-Fundichely, A., Berger, A., Huang, F. W., Forbes, A. N., Liu, E. M., ... & Khurana, E. (2017). Identification of novel prostate cancer drivers using RegNetDriver: a framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network. Genome biology, 18(1), 1-23.

[16] Chiu HS, Martínez MR, Bansal M, Subramanian A, Golub TR, Yang X, et al. High-throughput validation of ceRNA regulatory networks. BMC genomics. 2017;18(1):418. pmid:28558729

[17] Do D, Bozdag S. Cancerin: A computational pipeline to infer cancer-associated ceRNA interaction networks. PLoS computational biology. 2018;14(7):e1006318. pmid:30011266

[18] Wen X, Gao L, Hu Y. LAceModule: Identification of Competing Endogenous RNA Modules by Integrating Dynamic Correlation. Frontiers in genetics. 2020;11:235. pmid:32256525

[19] Chiu HS, Llobet-Navas D, Yang X, Chung WJ, Ambesi-Impiombato A, Iyer A, et al. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. Genome research. 2015;25(2):257–267. pmid:25378249

[20] Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011;146(3):353–358. pmid:21802130

[21] Yang J, Li T, Gao C, Lv X, Liu K, Song H, et al. FOXO1 3' UTR functions as a ceRNA in repressing the metastases of breast cancer cells via regulating miRNA activity. FEBS letters. 2014;588(17):3218–3224. pmid:25017439
[22] Zhou M, Wang X, Shi H, Cheng L, Wang Z, Zhao H, et al. Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer. Oncotarget. 2016;7(11):12598. pmid:26863568
[23] Kumar MS, Armenteros-Monterroso E, East P, Chakravorty P, Matthews N, Winslow MM, et al. HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. Nature. 2014;505(7482):212–217. pmid:24305048
[24] Qi X, Zhang DH, Wu N, Xiao JH, Wang X, Ma W. ceRNA in cancer: possible functions and clinical implications. Journal of Medical Genetics. 2015;52(10):710–718. pmid:26358722
[25] Huang M, Zhong Z, Lv M, Shu J, Tian Q, Chen J. Comprehensive analysis of differentially expressed profiles of lncRNAs and circRNAs with associated co-expression and ceRNA networks in bladder carcinoma. Oncotarget. 2016;7(30):47186. pmid:27363013
[26] Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457-i466.
[27] Mohamed, S. K., Nováček, V., & Nounu, A. (2020). Discovering protein drug targets using knowledge graph embeddings. Bioinformatics, 36(2), 603-610.----Classification of Cancer Types Using Graph Convolutional Neural Networks [Ramirez, 2020]
[28] Liu, Q., Hu, Z., Jiang, R., & Zhou, M. (2020). DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics, 36(Supplement_2), i911-i918.
[29] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1), 31-36.
[30] Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, Zhang J, Salama P, Rizkalla M, Han Z, Huang K. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. Front Genet. 2019 Mar 8;10:166. doi: 10.3389/fgene.2019.00166. PMID: 30906311; PMCID: PMC6419526.
[31] Xie, G. et al. Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. Genes 10, 240 (2019).
[32] Wang, T., Shao, W., Huang, Z. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun 12, 3445 (2021). https://doi.org/10.1038/s41467-021-23774-w
[33] Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics, 25(22), 2906-2912.
[34] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3), 333-337.
[35] Hung Nguyen, Sangam Shrestha, Sorin Draghici, Tin Nguyen, PINSPlus: a tool for tumor subtype discovery in integrated genomic data, Bioinformatics, Volume 35, Issue 16, 15 August 2019, Pages 2843–2846, https://doi.org/10.1093/bioinformatics/bty1049
[36] V. Asati, D. K. Mahapatra, and S. K. Bharti, "K-Ras and its inhibitors towards personalized cancer treatment: Pharmacological and structural perspectives," Eur. J. Med. Chem., vol. 125, pp. 299–314, 2017.
[37] Teran Hidalgo, S., Ma, S. Clustering multilayer omics data using MuNCut. BMC Genomics 19, 198 (2018). https://doi.org/10.1186/s12864-018-4580-6
[38] Cheang, Maggie CU, et al. "Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer." JNCI: Journal of the National Cancer Institute 101.10 (2009): 736-750.
[39] Prat, Aleix, et al. "Clinical implications of the intrinsic molecular subtypes of breast cancer." The Breast

24 (2015): S26-S35.
[40] Page, Lawrence, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.
[41] Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. "Latent space approaches to social network analysis." Journal of the american Statistical association 97.460 (2002): 1090-1098.
[42] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
[43] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.
[44] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).
[45] Veličković, Petar, et al. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017).
[46] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
[47] Zhang, J., Zhang, M. J., & biocViews Microarray, C. (2013). Package 'CNTools'.
[48] Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 1-13.
[49] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. J Stat Softw, 36(11), 1-13.
[50] Ramirez, R., Chiu, Y. C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., ... & Jin, Y. F. (2020). Classification of cancer types using graph convolutional neural networks. Frontiers in physics, 8, 203.
[51] Li, B., Wang, T., & Nabavi, S. (2021, August). Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 1-9).
[52] Multi-View GCNs with Attention Mechanism (MAGCN). [Online]. https://sxu-yaokx.github.io/MAGCN/
[53] Cheng, J., Wang, Q., Tao, Z., Xie, D. Y., & Gao, Q. (2020). Multi-View Attribute Graph Convolution Networks for Clustering. In IJCAI (pp. 2973-2979).
[54] Xie, Y., Zhang, Y., Gong, M., Tang, Z., & Han, C. (2020). Mgat: Multi-view graph attention networks. Neural Networks, 132, 180-189.
[55] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 46(D1), D1074-D1082.
[56] Kesimoglu, Ziynet Nesibe, and Serdar Bozdag. "Crinet: A computational tool to infer genome-wide competing endogenous RNA (ceRNA) interactions." Plos one 16.5 (2021): e0251399.
[57] Sumazin, Pavel, et al. "An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma." Cell 147.2 (2011): 370-381.
[58] Tay, Yvonne, et al. "Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs." Cell 147.2 (2011): 344-357.
[59] Sarver, Aaron L., and Subbaya Subramanian. "Competing endogenous RNA database." Bioinformation 8.15 (2012): 731.
[60] Anderson WF, Chatterjee N, Ershler WB, Brawley OW. Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database. *Breast cancer research and treatment.* 2002;76:27–36.
[61] Dietze, Eric C., et al. "Triple-negative breast cancer in African-American women: disparities versus biology." Nature Reviews Cancer 15.4 (2015): 248-254.
[62] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint

arXiv:1412.6980 (2014).
[63] Kim S, Chen J, Cheng T, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021;49(D1):D1388–D1395. doi:10.1093/nar/gkaa971
[64] Kuhn,M. et al. (2016) The SIDER database of drugs and side effects. Nucleic Acids Res., 44, D1075–D1079.
[65] Tatonetti,N.P. et al. (2012) Data-driven prediction of drug effects and interactions. Sci. Transl. Med., 4, 125ra31.

# Appendix

- CRINET paper:
  Kesimoglu, Ziynet Nesibe, and Serdar Bozdag. "Crinet: A computational tool to infer genome-wide competing endogenous RNA (ceRNA) interactions." Plos one 16.5 (2021): e0251399.
- Supplementary for CRINET