2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

HCNM: Heterogeneous Correlation Network Model for Multi-level Integrative Study of Multi-omics Data for Cancer Subtype Prediction

Reddy Rani Vangimalla[1] and Jaya Sreevalsan-Nair[1]

Abstract— **Integrative analysis of multi-omics data is important for biomedical applications, as it is required for a comprehensive understanding of biological function. Integrating multi-omics data serves multiple purposes, such as, an integrated data model, dimensionality reduction of omic features, patient clustering, etc. For oncological data, patient clustering is synonymous to cancer subtype prediction. However, there is a gap in combining some of the widely used integrative analyses to build more powerful tools. To bridge the gap, we propose a multi-level integration algorithm to identify representative integrative subspace and use it for cancer subtype prediction. The three integrative approaches we implement on multi-omics features are, (1) multivariate multiple (linear) regression of the features from a cohort of patients/samples, (2) network construction using different omics features, and (3) fusion of sample similarity networks across the features. We use a type of multilayer network, called heterogeneous network, as a data model to transition between a network-free (NF) regression model and a network-based (NB) model, which uses correlation networks. The heterogeneous networks consist of intra- and inter-layer graphs. Our proposed heterogeneous correlation network model, HCNM, is central to our algorithm for gene-ranking, integrative subspace identification, and tumor-specific subtypes prediction. The genes of our representative integrative subspace have been enriched with gene-ontology and found to exhibit significant gene-disease association (GDA) scores. The subspace in genes which is less than 5% of the total gene-set of each genomic feature is used with NB fusion integrative model to predict sample subtypes. As the identified integrative subspace data of multi-omics is less prone to noise, bias, and outliers, our experiments show that the subtypes in our results agree with previous benchmark studies and exhibit better classification between poor and good survival of patient cohorts.** ***Clinical relevance: Finding significant cancer-specific genes and subtypes of cancer is vital for early prognosis, and personalized treatment; therefore, improves survival probability of a patient.***

I. INTRODUCTION

There have been recent efforts in comprehensive studies of "multidimensional" omics data [1], which in oncology has been encouraged by the release of The Cancer Genomic Atlas (TCGA) dataset [2]. TCGA provides genomic, epigenomic, transcriptomic, and proteomic data of various cancer profiles, facilitating researchers to study significant cancer-causing genes and cancer subtypes using both single- and multi-omic features. These comprehensive studies are conducted by *integrating* either the data, its analytics, or both from these different omic features [1].

*This study has been supported by the Visvesvaraya PhD Scheme for Electronics and IT, the Ministry of Electronics and Information Technology, Government of India.

[1]Reddy Rani Vangimalla and Jaya Sreevalsan-Nair are with Graphics-Visualization-Computing Laboratory, and E-Health Research Center, International Institute of Information Technology Bangalore, Karnataka, India. reddyrani.vangimalla@iiitb.ac.in | jnair@iiitb.ac.in

For cancer studies pertaining to outcome prediction, multi-omics information has been routinely integrated at the data-level to obtain transformed data models, such as, regression and network models. For instance, multivariate multiple linear regression of multi-omics data has been used to construct gene-gene interaction (GGI) networks [3], and directed random walks with multi-omic information has been used on pathway information [4]. Recently, the multi-omics information has been integrated to form a discriminative dimensionality reduction tree [5], which is further used for outcome prediction. The available high-throughput omic data causes a "small n, large p" or "short-fat data" problem. The network topology-based algorithms can alleviate this problem through their gene ranking applications. Identifying these significant genes and using them as representative features creates "low-dimensional subspaces" [6]. Alternatively, networks used for single-omic studies can be fused, thus integrating analytics from different omic features [7]. Similarity network fusion (SNF) [7] and affinity network fusion (ANF) [8] are examples of methods where patient similarity networks from each omic data type are fused using affinity measurement.

Each of the state-of-the-art integrative studies has its own benefits and shortcomings, and are mostly used in isolation. We also observe that the integrative studies broadly fall under the category of data modeling or transformation. We hypothesize that the semantics of some of these data models allow them to be extendable, and also work with other integrative methods. We consider a specific example of extending the use of an integrative regression model for finding representative subspaces, followed by an appropriate network fusion method for predicting cancer subtypes. The integrative regression model captures the interdependence between two multi-omics features at the data-level [3], whereas the network fusion integrates the analytics performed separately from different omic features. Thus, we demonstrate that such integrative methods can be plugged into the same workflow or implementation to improve the overall understanding of the high-dimensional multi-omics data. In order to achieve a multi-level integration of the multi-omics data through existing integrative methods, namely regression and network fusion, we propose a data model that will transition one method to another, referred to as the Heterogeneous Correlation Network Model (HCNM). We propose a three-level integration algorithm driven by HCNM for gene-ranking, integrative subspace identification, and cancer subtype prediction (Figure 1). Finding integrative subspaces is equivalent to feature selection as well

as dimensionality reduction, and is a pertinent research problem in the face of increased dimensionality in integrative studies [1]. For finding subspace, some of the dimensions from the full space are ranked using appropriate methods and selected. The representative subspace is the subspace that best represents the full space for subtype classification in our work.

Heterogeneous networks are a special class of multilayer networks, where the nodes in each layer are different [1]. Heterogeneous networks have intra-layer and inter-layer graphs, where the latter is a bipartite graph between nodes in different layers [9]. These networks have been used for multi-omics data for up to four different omic features [10] and are predisposed to embed the multi-omics data by design, and thus provide novel tools for integrative studies [1]. We use a heterogeneous network model here to transition between a network-free (NF) regression model and a network-based (NB) network fusion method, through the use of correlation networks. There exist several NF and NB integrative methods for multi-omics data [1]. Hence, our proposed data model is specifically a heterogeneous *correlation* network model. HCNM is similar to the heterogeneous network model, iHNMMO [10] in terms of the use of regression and correlation. The difference is that iHNMMO has normalized correlation network layers with regression coefficients as inter-layer graph edge weights, whereas HCNM has a partial correlation layer computed from the regression model, and cross-correlation coefficients as inter-layer graph edge weights.

Here, we propose a multi-step algorithm for the construction and use of HCNM to predict subtypes for a cancer phenotype. The steps are: (1) integration using multivariate multiple linear regression I_1, (2) construction of correlation network layers for intra-layer graphs, (3) community detection by consensus, (4) ranking genes to be integrated in an inter-layer graph I_2, (5) computing inter-layer graph edge weights, thus completing our HCNM, (6) finding integrative subspace by ranking edges in the inter-layer graph, (7) integration of sample similarity networks across different omic features by network fusion I_3, and (8) clustering of the fused similarity network to find subtypes. Steps (7)-(8), including I_3 can use several network fusion methods [6], [7], [8]. The novelty of HCNM lies in embedding the interdependence of different omic features in intra-layer edges, rather than inter-layer edges. Our contributions are in:

- Transforming a network-free multivariate multiple linear regression model to our proposed heterogeneous correlation network model, HCNM,
- Using consensus clustering in the intra-layer graphs of HCNM for ranking genes,
- Proposing an algorithm with multi-level integration of multi-omics data for gene-subspace identification, for an application of cancer subtype identification.

The scope of our current study is limited to undirected networks, which can be further improved using pathway information [4] or Bayesian networks [1].

II. MATERIALS AND METHODS

We consider two different omic features in our work, thus generating two layers in our proposed heterogeneous network model HCNM). The GGI networks in HCNM are of gene expressions (**Layer-1**) and methylation features (**Layer-2**). These layers, by design, are correlation networks, as given in the model name. By *multi-level integration* of multi-omics data, we imply three occurrences of multi-omics integration in our algorithm: I_1 when using a multivariate multiple (linear) regression (MMR) model, I_2 for selecting genes to compute the inter-layer graph using cross-correlations, and I_3 for network fusion of similarity networks of samples/patients.

Our multi-step algorithm can be broadly divided into two stages. Stage $\mathbb{S}_1$ is for finding representative integrative subspace of significant genes using our proposed model, HCNM, and Stage $\mathbb{S}_2$ is for predicting cancer subtypes using the subspace.

A. Data

Our case study is on breast invasive carcinoma of TCGA from TCGA-BRCA project, using gene expression and DNA methylation data. The downloaded[1] dataset is of 1098 samples. We use an R, Bioconductor package `TCGAbiolinks` [11] to download gene expression and methylation data from `Illumina HiSeq` and `Illumina Human Methylation 450` platforms, respectively, and data with `PAM50` labels of breast cancer subtypes. `PAM50` labels are amongst the most widely used breast cancer subtypes annotation [12], where the subtypes are luminal A, luminal B, HER2 positive, triple-negative or basal-like type, and normal categories.

Pre-processing:
The three-step pre-processing features in the multi-omics data includes outlier removal, imputing missing values, and standardization. We perform outlier removal for each omic dataset by removing the features that satisfy one of these three conditions: (1) its value across all samples is zero, (2) its missing values account for more than 25% of the overall sample size, (3) its variance is in the lower 25% of the overall variance of all features [7], [13]. For the retained omic features, we impute the missing values using the median value of all samples. For each omic feature, we then standardize the values using z-scores, such that $(\mu, \sigma) = (0, 1)$. Finally, methylation probes are mapped to genes; and if a probe is mapped to multiple genes, a least correlated feature with the gene expression trait is considered [14]. Suppose the gene is not available in expression data and multiple probes are associated with it. In that case, the methylation feature with the maximum variance is considered and mapped to that gene.

We additionally select specific samples based on the clinical information to meaningfully study a cohort. We have filtered the samples if a patient's vital status is 'alive,' yet the number of survival days is below the median of survival days

[1]The dataset has been downloaded in December 2020.

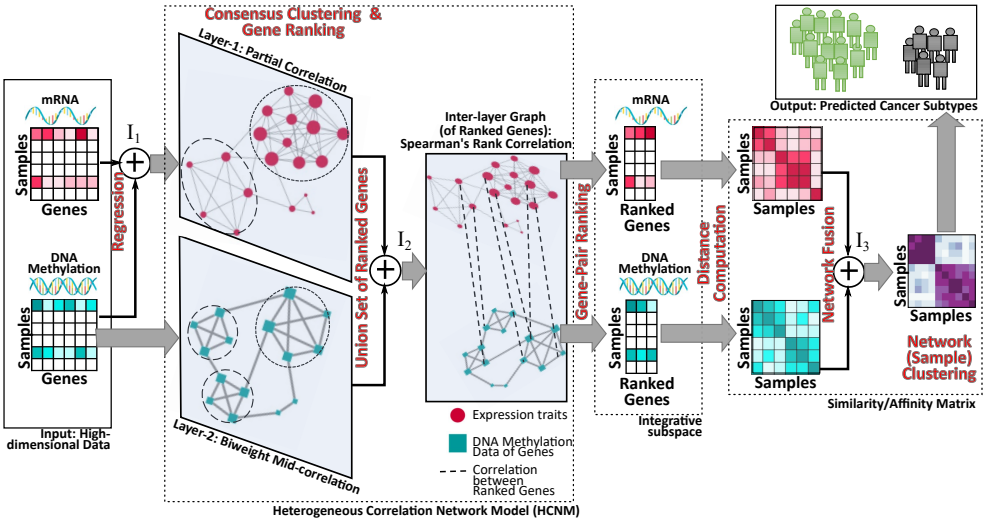

Fig. 1. Our proposed algorithm of multi-level integration of multi-omics data using HCNM, with I_1, I_2, and I_3 as the different integration steps.

across all samples. We finally take the intersection set of the samples retained for both expression traits and methylation features. The final data used further in our case study is of 486 samples, with 16,626 expression traits and 10,109 methylation features.

B. Stage $\mathbb{S}_1$: Finding Integrative Subspace Using HCNM

Here, we construct the intra-layer graphs in HCNM, detect communities by consensus in these layers, rank genes based on communities to construct the inter-layer graph, and finally construct the inter-layer graph in HCNM. Using the inter-layer graph, we perform a second iteration of ranking genes, to select highly ranked "significant" genes for finding representative subspace. We refer to this as an *integrative* subspace as it is the union-set of subspaces in all omic feature spaces. $\mathbb{S}_1$ includes two integration steps: I_1 using regression to construct one of the intra-layer graphs, and I_2 where genes are selected using consensus communities and ranking procedure.

Step-1: Multivariate Multiple Regression (I_1):
We use the MMR model by treating DNA methylation data of genes as features and expression traits as outputs, for integrating selected/filtered methylation features and expression traits [3].

For m gene expression levels and n methylation features, we have $Y \in \mathbb{R}^{k\times m}$ and $X \in \mathbb{R}^{k\times n}$, respectively, for k samples. The MMR model is written as

$Y = B \cdot X + E$, where $B \in \mathbb{R}^{m\times n}$ and $E \in \mathbb{R}^{m\times k}$,

for the regression coefficient matrix B and residual error matrix E. We now have $Y = [y_1, y_2, \ldots, y_m]$, corresponding to $E = [\epsilon_1, \epsilon_2, \ldots, \epsilon_m]$, with $\epsilon_i \sim N(0, \sigma^2)$, $\forall i \in [1, m]$, by the conventional linear regression model. Here, we implement MMR using Lasso (Least absolute shrinkage and selection operator) [15] regression model.

Step-2: Construction of Intra-layer Graphs in HCNM:
To transform a regression model into a network-based model, correlation networks are a natural choice. Regression models have been used for computing partial correlation coefficients [16], which quantifies the correlation between the dependent variables, when conditioning on the independent variables. The intra-layer graph for the independent variable is computed using conventional correlation values.

Layer-1 *(Gene expression levels)*:- When a linear regression model is used, the n$^{\text{th}}$-order partial correlation, *i.e.,* conditioned to n independent variables, can be computed as the total linear (Pearson) correlation between the residual errors [16]. When Y is regressed on X, the residual error $e^{(Y)}$ represents the parts of Y that are uncorrelated with X.

$$e^{(Y)} = Y - \left(\hat{\beta}_0^{(Y)} + X\hat{\beta}_1^{(Y)}\right).$$

Thus, the partial correlation coefficient z of Y, when conditioning on X, is:

$\mathrm{z}\{Y\} = \{\rho(\epsilon_i, \epsilon_j)\} = \rho\left\{e^{(Y)}\right\}$, where $\rho_{\epsilon_i,\epsilon_j} = \frac{\mathrm{cov}(\epsilon_i,\epsilon_j)}{\sigma_{\epsilon_i}\sigma_{\epsilon_j}}$, and cov and σ refer to covariance and standard deviation, respectively. These computed partial correlation coefficients are now weights of edges between m expression traits, in **Layer-1** of HCNM.

Layer-2 *(DNA methylation features)*:- Since we are computing the linear correlation amongst the methylation features, we determine the biweight midcorrelation (bicor) coefficients [17]. Bicor is widely used for computing correlation between genomic features, as it is a median-based measure, making it less prone to outliers. Despite their similarities, bicor is preferred over Pearson correlation in genomic applications, where it is also widely used as a similarity measure. These computed bicor coefficients are now weights of edges between n methylation features, in **Layer-2** of HCNM.

Step-3: Community Detection by Consensus:
Clusters of genes in GGI networks, identified using their coexpression values, are often enriched with similar functional annotations [18]. Communities identified in these networks give such gene clusters. Both **Layer-1** and **Layer-**

2 are completely connected networks, which requires them to be sparsified for performing community detection using popularly used methods, such as, walktrap [19], fast greedy optimization [20], and Louvain community detection [21].

Graph Sparsification:- Wolfe *et al.* [22] have explained how the guilt-by-association (GBA) heuristic implies that the weaker co-expressions or edges in a network are frequently connected to the dissimilar functional clusters. Thus, these edges have to be filtered out of the network to enable us to identify significant connected components in the network. These connected components are also *"locally dense, globally sparse"* communities with strong inter- and weak intra-community links. Several techniques have been successfully used for determining the threshold for edge-filtering [23], which include p-value based methods, and percolation analysis (PA). PA involves observing the connected components of the network while progressively increasing the threshold for edge weights [24]. The threshold at which the giant connected component begins to fragment is considered optimal to filter out edges that retain the network as a single connected component. We use this threshold value τ as an absolute value cutoff, implying filtering out edges with weights in the interval $[-\tau, \tau]$. Hence, we first filter the edges based on p-value, and then based on τ from PA. When using correlation networks, we retain only statistically significant edges, which represent correlations with p-value$<$ 0.05. Here, τ for **Layer-1** and **Layer-2** are 0.36 and 0.32, respectively. We now have 15,756 gene expression levels with 3,367,038 edges in **Layer-1**, and 9,826 methylation features with 7,545,734 edges in **Layer-2**.

Communities:- "Locally dense, globally sparse" communities regularly occur in biological networks, where hubs distributed in the dense subnetworks play specific biological roles [25]. Depending on the semantics of community formation, intra-community genes have a higher likelihood of similar roles in a specific disease [26]. In the absence of ground truth for communities in the intra-layer graphs in our case study, we find communities by consensus from selected widely used community detection techniques, namely:

- **Walktrap Method:** identifies denser neighborhoods as communities, based on the assumption that a random walker tends to visit denser neighborhoods for extended periods of time compared to sparser ones. The distance between the two nodes is the likelihood of reaching from one to another in n steps. Communities are clusters that are merged using Ward's hierarchical clustering [27], while minimizing intra-cluster distances. We use an optimal value of $n = 4$, here.
- **Fast Greedy Optimization Method:** uses a hierarchical, agglomerative model, à la Walktrap method. It is also similar to Newman's modularity maximization method [28], but is more suitable for large networks attributed to its linear running time. The greedy optimization identifies communities with high intra- and low inter-cluster edge densities.
- **Louvain Community Detection:** is another greedy optimization method that uses modularity maximization to partition a network. This is widely used in genomic analysis.

Consensus Voting:- We select these methods based on the similarity of the semantics of their outputs by design. Thus, we expect to get similar results from these methods, which can be aggregated for a final outcome by consensus. We arrive at a consensus by voting if pairwise nodes, *i.e.,* genes, are likely to be in a community. Thus, the *co-association* votes are equivalent to the likelihood of genes i and j are in the same community across the results from the selected methods. Modeled as a network, the co-association votes $D_{ij}^{(k)}$ are computed between nodes i and j, belonging to communities C_i and C_j, respectively, using different community detection methods, for $k = 1, \ldots, N_M$, and are averaged to get the aggregated co-association vote, D_{ij}.

$$D_{ij}^{(k)} = \begin{cases} 1 & \text{, if } C_i = C_j, \\ 0 & \text{, otherwise} \end{cases}, \text{ and } D_{ij} = \frac{1}{N_M} \cdot \sum_{k=1}^{N_M} D_{ij}^{(k)}.$$

Thus, the network represented by the (aggregated) co-association matrix, D, is the sparsified version of the correlation network, for each intra-layer graph. We now use the genLouvain community detection algorithm [29] on the transformed network. GenLouvain is a variant of the Louvain community detection algorithm that additionally uses a resolution parameter γ and a modified modularity score:

$$Q(\vec{g}, \gamma) = \sum_{j=1}^{n} \sum_{i=1}^{n} (A_{ij} - \gamma P_{ij}) \delta(g_i, g_j),$$

where A is adjacency matrix (*i.e.,* D, here) of size n, and P_{ij} is an expected matrix under null model = $\frac{k_i k_j}{2m}$, where $k_i = \sum_j A_{ij}$ and $2m = \sum_i k_i$.

Using Walktrap, fast greedy optimization, and Louvain methods, we get 2902, 626, and 430 communities in **Layer-1**, respectively. Similarly, we get 757, 458, and 310 communities in **Layer-2**, respectively. GenLouvain implemented on co-association networks, for 25 iterations, gives 130 and 770 communities in **Layer-1** and **Layer-2**, respectively.

Step-4: Ranking Genes for Inter-layer Graphs (I_2*)*:
In each intra-layer graph, the genes are ranked based on the significance in the network, quantified by the measures, such as, node degree, node betweenness centrality, and eccentricity. The highest betweenness centrality score implies that the node is on the shortest path of most other nodes. The nodes with eccentricity equal to the network radius are considered to be *central* in the network. We thus identify three sets of nodes in each community in each intra-layer graph, (1) all central nodes, (2) top 10% of genes, ranked based on their node degree, and (3) top 10% of genes, ranked based on betweenness centrality. The union-set of these sets gives us the significant genes in the layer. This process finds 3,895 of 16,626 expression traits, and 1,882 of 10,109 DNA methylation data of genes to be significant. The union-set of the selected features from both layers gives the inter-layer graph, which is now an integration, albeit a weaker one compared to I_1 and I_3.

Step-5: Construction of Inter-layer Graph in HCNM:
The inter-layer graph is computed using Spearman's correlation matrix between the selected genes from **Layer-1** and **Layer-2**. We sparsify this graph using the p-value and threshold from PA, as done for intra-layer graphs. The sparsi-

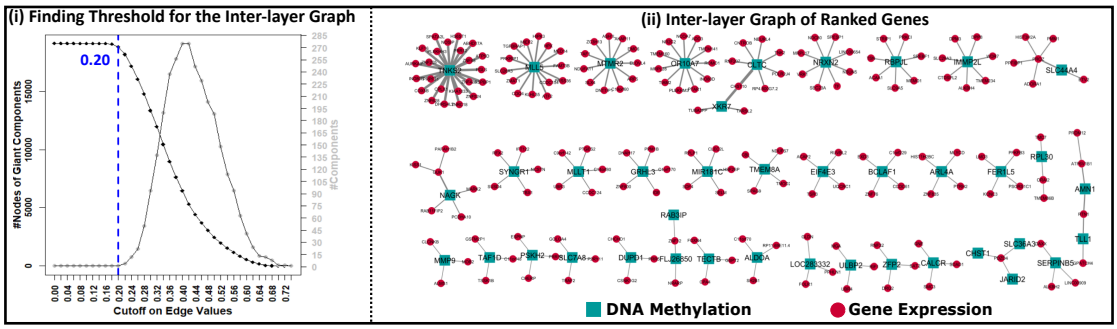

Fig. 2. Inter-layer graph of HCNM. (i) A plot of the edge cutoff value against the network components using percolation analysis, giving selected threshold $\tau = 0.20$, for the graph. (ii) The connected components of the graph, with 3+ nodes and of genes retained after edge (gene-pair) ranking, with edges between gene expression levels (red circular glyphs) and methylation features (turquoise square glyphs), and the edge width indicates the edge betweenness.

fication could also be done along with the intra-layer graphs with the genes present in step-3. Using this sparsification, we get $\tau = 0.2$ (Figure 2(i)). This gives us 974,812 edges in the inter-layer graphs between 3895 expression genes and 1882 DNA methylation data of genes. This completes the construction of our proposed HCNM.

Step-6: Representative Integrative Subspace:
To further reduce the number of significant genes in the representative set, we rank the gene-pairs in the inter-layer graph based on their edge-betweenness centrality measure. The top 10% of these pairs found in most of the shortest paths of the network are finally selected. The connected components with 3+ nodes of the bipartite inter-layer graph are shown in Figure 2(ii). This representative subspace of these selected genes is thus integrative and feature-rich, making it adequate to study the subtypes instead of the entire gene space across both layers. In our case study, the resultant subspace now has 531 gene expression traits and 339 methylation features, from 486 samples.

C. Stage $\mathbb{S}_2$: Subtype Prediction Using Integrative Subspace

We first identify the similarity or affinity networks of the samples for each omic feature in the integrative subspace, and then fuse them as an integrative step (I_3). We then find clusters of samples, representative of cancer subtypes.

Step-7: Network Fusion (I_3):
Most multi-omics integrative algorithms, such as, similarity network fusion (SNF) [7], and affinity network fusion (ANF) [8] integrate data of both omic features after computing similarity or affinity matrices internally. ANF is an improvised integrative procedure on SNF; both the methods first compute the distance between patients. Using the distance matrix, affinity measure of each genomic feature is computed separately. Implementing network fusion procedure on the affinity measures, a final multi-omics integrated network is generated. We use SNF and ANF by tuning the hyperparameters K (number of neighbours), σ (variance for affinity measurement), α (measure for local diameter) and β (measure for pair-wise distance) using the correlation measure (*ref:* Equation 7 [3]). We have used $K = 15$, $\sigma = 0.3$, $\alpha = 0.17$, $\beta = 0.2$, in 20 iterations, for running both SNF and ANF in our case study. The number of clusters N_C is estimated using eigen gap and rotation cost methods.

We also use iCluster [30] as an integrative method where similarities between the samples and clusters are computed simultaneously by minimizing the intra-cluster variance. We use the 'tune.iClusterplus' method to find the optimal number of clusters and Lasso penalties. iCluster takes longer computation time than SNF and ANF, as iCluster directly outputs the clusters, thus combining step-7 and step-8.

We have implemented these methods using R packages, namely, SNFtool [7], ANF [31], and iClusterPlus [32]. In our case study, the rotation cost method has estimated $N_C = 3$ and $N_C = 4$ for SNF and ANF, respectively, and the iCluster tune procedure, $N_C = 4$.

Step-8: Sample Clustering for Subtype Prediction:
We extract clusters of samples in the fused similarity or affinity networks using spectral clustering, with N_C (from step-7) as an input parameter. We then compare these clusters or subtypes with the popularly known breast-cancer subtype annotation data of TCGA, namely `PAM50` [12].

III. RESULTS AND DISCUSSION

For 486 samples, our HCNM has reduced the gene space of 16,626 gene expression traits and 10,109 DNA methylation data of genes to representative subspace, comprising of 531 genes and 339 methylation features, i.e., less than 5% of the total gene space. Our model also demonstrates novel characteristics of the selected genes, *e.g.,* the connected components in the bipartite inter-layer graph exhibit star structures. These star-graphs of 3+ nodes, with several gene expression traits around a single methylation feature, imply a many-to-one association between gene expression traits and methylation features (Figure 2(ii)).

Feeding the 870 genes of our integrative subspace into 'Database for Annotation, Visualization and Integrated Discovery (DAVID)' tool [33], [34], we get the enriched gene ontology (GO) terms, with N_G genes belonging to each term. The top 10 terms are sorted based on their false discovery p-value in Table I. We have verified the gene-disease association (GDA) score of the shared genes found in the GO terms with $N_G > 25\%$ of total genes (highlighted

TABLE I

THE TOP 10 ENRICHED GO TERMS OF OUR INTEGRATIVE SUBSPACE

GO	**Term**	N_G	**PValue**
UP_KEYWORDS	Phosphoprotein	**382**	5.33E-08
UP_KEYWORDS	Alternative splicing	**451**	4.91E-05
GOTERM_MF_DIRECT	GO:0005515 ~protein binding	**394**	1.13E-04
INTERPRO	IPR013164:Cadherin, N-terminal	11	1.93E-04
UP_KEYWORDS	Transit peptide	38	2.78E-04
GOTERM_CC_DIRECT	GO:0005737 ~cytoplasm	**245**	6.07E-04
UP_SEQ_FEATURE	splice variant	**339**	7.29E-04
GOTERM_CC_DIRECT	GO:0005829~cytosol	**164**	8.24E-04
UP_KEYWORDS	Cytoplasm	**218**	0.001086
UP_SEQ_FEATURE	domain:Cadherin 6	11	0.001183

TABLE II

A FEW TOP GENES ASSOCIATED TO BREAST CANCER, RANKED BY THEIR GENE-DISEASE ASSOCIATION SCORES (GDA-SC.) FROM DISGENET

Disease: Breast Carcinoma		Disease: Malignant Neoplasm of Breast		Disease: Triple Negative Breast Neoplasms	
Disease ID: C0678222		Disease ID: C0006142		Disease ID: C3539878	
Gene	GDA-Sc.	Gene	GDA-Sc.	Gene	GDA-Sc.
STAT3	0.4	STAT3	0.4	STAT3	0.1
MAPT	0.2	ATF2	0.37	SPAG9	0.02
ATF2	0.08	PPHLN1	0.3	DAXX	0.01
TRAF2	0.07	UBR4	0.3	CRTC1	0.01
NUMA1	0.03	RFX2	0.3		

TABLE III

GROUND TRUTH ANALYSIS OF BREAST-CANCER SUBTYPES PREDICTION

Scores	SNF		iCluster	ANF	
	HCNM	Full	HCNM	HCNM	Full
NMI	**0.43**	0.42	0.33	0.39	**0.43**
S_j	**0.46**	0.44	0.42	**0.51**	**0.51**

in column N_G) using the DisGeNET database[2]. A total of 38 genes have been found in common across seven top enriched GO terms. 35 out of these 38 genes have positive GDA scores, implying published evidence of the association of a majority of genes with the disease. A few of the top-ranked genes associated with the three different breast cancer types are listed in Table II. Overall, our HCNM has successfully identified the significant feature-rich sampled data for each genomic feature, which is representative of the application of subtype identification.

Using our representative subspace in multi-omics integrative procedures such as SNF, ANF, and iCluster, we have identified the subtypes in samples/patients. The subtypes found using subspace in genes are comparable to the subtypes found with complete multi-omics feature space. The reduced dimensionality data is beneficial, as the subspace is less prone to noise, bias, and outliers. The comparison of subtypes using Sankey plot (Figure 3) shows that those identified by SNF used with both full gene space and our representative subspace (HCNM-based), and by ANF with the full gene space are almost identical, but we observe slight differences with subtypes from ANF with our subspace. We

[2]https://www.disgenet.org/

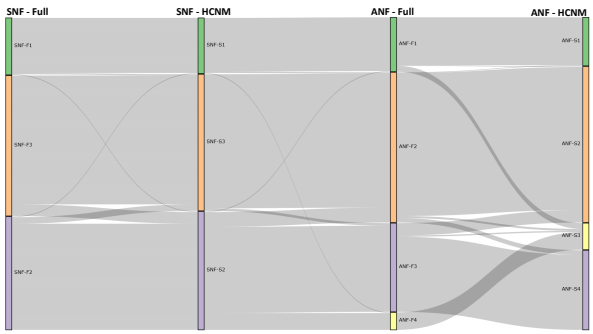

Fig. 3. Sankey plot of the patient subtypes from using our algorithm using network fusion methods (SNF, ANF) with data from the complete (Full) gene space and our representative subspace (HCNM), shows that subtypes found using SNF-HCNM data agree with SNF/ANF-Full data, more than ANF-HCNM.

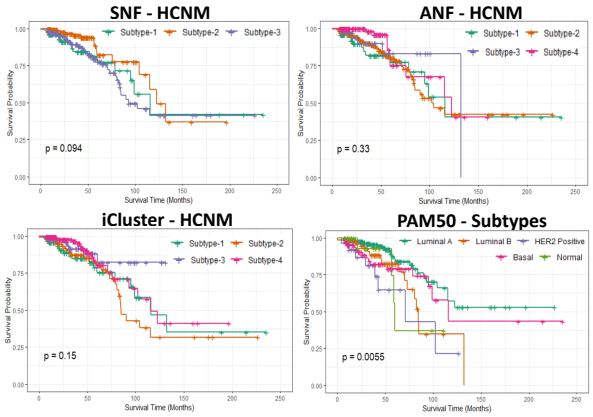

Fig. 4. The good and poor survival times for subtypes were predicted using different methods. Subtypes are significant at median survival probability in all methods. We see clear survival probability separation for subtypes identified using SNF with subspace in genes and benchmark study using annotated subtypes, *i.e.,* `PAM50`, than using ANF and iCluster.

have also observed similar behaviour when comparing the patients subtypes with popularly known subtypes annotation `PAM50,(BRCA_Subtype_PAM50)` [12].

We use Jaccard similarity S_j and Normalized Mutual Information *NMI* to compare the clustering results with the ground truth. Table III shows that (i) ANF with full gene space is comparable with SNF with our subspace, and (ii) iCluster on our subspace has the least S_j and *NMI* values.

We have studied the effect of subtypes on survival probabilities using Kaplan-Meier survival curves, comparing results of our representative subspace used in different multi-omics integrative procedures and the reference `PAM50` subtypes (Figure 4). SNF with our representative subspace displays a clear separation between good survival and poor survival subtypes based on their survival probabilities when observed at 50% of survival probability (Figure 4). In all three multi-omics integrative procedures, most subjects of the basal-like class are classified under subtype-1; the subjects of the classes, luminal A, and luminal B, are spread across two subtypes. Overall, we have observed that the network fusion based methods are more favorable over iCluster, when used with our representative subspace owing to the computational

time, and the results in Table III and Figure 4.

IV. CONCLUSIONS

In our case study of gene expression traits and methylation features in the TCGA-BRCA project, our proposed HCNM has successfully been used for finding the representative integrative subspace of genes associated with breast cancer. This multilayer network model uses correlations within and across the different omic features. The HCNM has significantly decreased the data dimensionality using ranking based on gene communities and network topology. We have used our subspace of biologically significant genes, and appropriate integrative fusion procedures to predict cancer subtypes. We have found that subtypes predicted using integrative network fusion methods, SNF and ANF, are comparable with the state-of-the-art benchmark studies, more than with iCluster, a Bayesian method. For evaluating the overall performance of the method, we need to perform an ablation study, which is in the future scope of this work. We intend to integrate Bayesian methods more extensively into multi-level integration algorithms to improve combining network-based and Bayesian methods as future work.

ACKNOWLEDGMENT

The dataset for the case study has been generated by the TCGA Research Network (https://www.cancer.gov/tcga). The authors would like to thank Hyun-hwan Jeong at the University of Texas Health Science Center at Houston, and Kyung-Ah Sohn at Ajou University for their support.

REFERENCES

[1] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi, "Methods for the integration of multi-omics data: mathematical aspects," *BMC bioinformatics*, vol. 17, no. 2, pp. 167–177, 2016.

[2] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary Oncology*, vol. 19, no. 1A, p. A68, 2015. [Online]. Available: http://cancergenome.nih.gov/

[3] R. R. Vangimalla, H.-h. Jeong, and K.-A. Sohn, "Integrative regression network for genomic association study," *BMC Medical Genomics*, vol. 9, no. 1, p. 31, 2016.

[4] S. Y. Kim, H.-H. Jeong, J. Kim, J.-H. Moon, and K.-A. Sohn, "Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies," *Biology Direct*, vol. 14, no. 1, pp. 1–13, 2019.

[5] M. Shi, J. Wang, and C. Zhang, "Integration of Cancer Genomics Data for Tree-based Dimensionality Reduction and Cancer Outcome Prediction," *Molecular Informatics*, vol. 39, no. 3, p. 1900028, 2020.

[6] S.-G. Ge, J. Xia, W. Sha, and C.-H. Zheng, "Cancer subtype discovery based on integrative model of multigenomic data," *IEEE/ACM trans. on comput. bio. & bioin. (TCBB)*, vol. 14, no. 5, pp. 1115–1121, 2016.

[7] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, p. 333, 2014.

[8] T. Ma and A. Zhang, "Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 398–403.

[9] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[10] C. Peng, A. Li, and M. Wang, "Discovery of Bladder Cancer-related Genes Using Integrative Heterogeneous Network Modeling of Multi-omics Data," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[11] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni *et al.*, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic acids research*, vol. 44, no. 8, pp. e71–e71, 2016.

[12] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, p. 61, 2012.

[13] K.-A. Sohn, D. Kim, J. Lim, and J. H. Kim, "Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors," *BMC Systems Biology*, vol. 7, no. S6, p. S9, 2013.

[14] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer," *Journal of biomedical informatics*, vol. 56, pp. 220–228, 2015.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[16] A. De La Fuente, N. Bing, I. Hoeschele, and P. Mendes, "Discovery of meaningful associations in genomic data using partial correlation coefficients," *Bioinformatics*, vol. 20, no. 18, pp. 3565–3574, 2004.

[17] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[18] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. of the Nat. Acad. of Sc. (PNAS)*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[19] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *International symposium on computer and information sciences*. Springer, 2005, pp. 284–293.

[20] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[22] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks," *BMC bioinformatics*, vol. 6, no. 1, p. 227, 2005.

[23] B. R. Borate, E. J. Chesler, M. A. Langston, A. M. Saxton, and B. H. Voy, "Comparison of threshold selection methods for microarray gene co-expression matrices," *BMC research notes*, vol. 2, no. 1, p. 240, 2009.

[24] C. Bordier, C. Nicolini, and A. Bifone, "Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold," *Frontiers in neuroscience*, vol. 11, p. 441, 2017.

[25] A. D. Perkins and M. A. Langston, "Threshold selection in gene co-expression networks using spectral graph theory techniques," in *BMC bioinformatics*, vol. 10, no. 11. BioMed Central, 2009, p. S4.

[26] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics*, vol. 12, no. 1, p. 56, 2011.

[27] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[28] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[29] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010.

[30] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi, and C. Sander, "Integrative subtype discovery in glioblastoma using iCluster," *PloS one*, vol. 7, no. 4, p. e35236, 2012.

[31] T. Ma and A. Zhang, "Affinity network fusion and semi-supervised learning for cancer patient clustering," *Methods*, vol. 145, pp. 16–24, 2018.

[32] Q. Mo and R. Shen, "Package 'iClusterPlus'," 2018.

[33] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2009.

[34] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protoc.*, vol. 4, no. 1, p. 44, 2009.