Graph-Based Semi-Supervised Learning for Multi-Omics Data Integration

Ramón Reszat
Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—Multi-omics analysis has become an increasingly important tool to help design personalized and precision medicine. This thesis compares graph-based data integration techniques for multi-omics data. In particular, affinity graphs constructed with SNF and ANF are analysed and the influence of the scaling factors on connectivity and modularity is determined. Afterwards, the two methods are evaluated on clustering cell lines from the CCLE dataset based on their multi-omics profile. The results show, that lung carcinoma subtypes can be clustered using ANF at 87% accuracy and using SNF with 91% accuracy. The predicted multi-omics cluster of small cell lung carcinoma (SCLC) shows signitures of drug resistance to dasatinib (p=6.686e-06). Additionally, a biomarker for doxorubicin resistance on gastric and breast cancer cells is examined (p=3.54e-02). Overall, analysis of affinity graphs reveals the potential for graph-based models on multi-omics data.**

***Index Terms*—graph-based, data integration, affinity graphs**

I. Introduction

Recent advances in high-throughput biochemical assays have made it possible to collect vast amounts of data on a single biological sample. Therefore, modern studies are increasingly being designed using measurements of multiple data types, i.e., gene, metabolite, microRNA expression, chromatin profiling, etc. to capture a single phenomenon of interest. The fields of study that collect data from these sources are commonly referred to as omics sciences. The progress in the development of molecular technology has helped us understand the structure and function of the genetic code and expanded our understanding of how biological processes are being regulated. For instance, decreasing costs for next-level sequencing techniques have made it possible to elucidate gene regulation by measuring the abundance of mRNAs [1] and therefore obtain transcriptomics data on a larger scale [2]. This has made it possible to further explore the non-coding parts of the genome such as microRNAs, that take a role in gene regulation after translation. At the same time, new instruments for mass-spectrometry (e.g. LC–MS, GC–MS) allowed the collection of metabolomics data [3] characterizing intermediary or final products that are produced on a single-cell level. Together, these technologies produce structured data, that can be used for detailed modelling and statistical analysis of biological processes.

Combining the datasets from these omics platforms can facilitate the discovery of disease subtypes that are characterized by features from multiple domains, as different data sources capture different aspects of the sample that do not tend to overlap [4]. Hence, the joint view of the samples can lead to a picture of the patient that considers all available diagnostic methods. Therefore, advances in multi-omics modelling have the potential to create great clinical value. Data Science and AI tools are required that can handle data from heterogeneous sources and assist clinicians to design personalized and precision medicine [5]. For example, genetic biomarkers are already used to assess effective treatment options in certain types of cancers. A prominent example would be the expression of *HER2* in breast cancer before trastuzumab can be used in a clinical setting [6] or *EGFR* expression before cetuximab can be used against pancreatic cancer [7]. Complementing these genetic markers with combined genetic, epigenetic and metabolic markers has the potential to shift the boundaries of current prognostic capabilities.

However, the use of multi-omics data for these purposes has been largely slowed down by the challenges of integrating the different data views into a comprehensive model [8]. Firstly, there is not always a direct association between gene expression, other omics measurements and the resulting phenotype [9]. Secondly, exact measurements depend strongly on the experimental conditions, as repeating a study at a different time and location will introduce biases in the form of batch effects [10]. Lastly, the underlying distributions of the data generated from different omics pipelines are vastly different, resulting in different feature scales and non-linear relationships between them.

Various methods have been discussed in literature to tackle these problems. The simplest way of combining measurements from different omics datasets is to concatenate them column-wise and to treat each expression value as a separate feature of the sample. Then traditional ML methods such as K-means [11] or random forests [12] can be used directly for classification and clustering of samples. Ensemble methods can be used to analyze the influence of individual features from each dataset, i.e. XGBoost [13] or Elastic Net (EN) [14] have been used to build models for each omics technology to then construct a combined prediction using a boosting approach [15]. In particular, ensemble methods have shown great performance due to their robustness to feature scales

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor: Dr. Rachel Cavill

and the possibility for feature selection. Even though these models are designed to minimize the generalization error rate [16] the computational cost and complexity is greatly increased. Correlation-based integration methods have the goal to identify relationships between variables and to reduce the overall dimensionality of the dataset. It is the type of statistical analysis that is most commonly used in practice to find correlations between different types of omics data. Another approach is pathway-based integration. It considers pre-existing biological models of interactions. This relies on predefined pathways from databases such as KEGG [17]. Those pathways can be altered or multiple unrelated pathways can cause the same phenotype. These methods have been used to explain causal effects and to elucidate molecular mechanisms showing promising results. Hence, tools such as IMPaLA [18] can be used successfully for integrated pathway analysis.

This paper will examine network-based data integration as a new approach of fusing omics datasets. Graph-based methods have been discussed for prediction and subtype classification [19], but more recently a survey [20] has suggested that Similarity Network Fusion (SNF) [21] still performs moderately compared to other dimensionality reduction techniques on a sample clustering task. However, Affinity Network Fusion (ANF) [22] has not been included in those surveys, although the authors argue that ANF achieves similar or better performance than SNF, while creating a representation that is suitable for semi-supervised learning tasks.

The interesting aspects of these algorithms remain, that they provide a sample based patient centered view, which can be used for instance-based learning on the graphs that they construct. Hence, we want to examine the difference between the two proposed methods. Firstly, we measure the influence of the hyperparameters of SNF and ANF on the affinity graphs. Secondly, we determine how the results of using data from a single source compare to those using multi-omics data fusion. Lastly, we discuss if these algorithms can reveal signatures of drug resistance that can be used to develop combined multi-omics biomarkers.

II. Methods

Given raw measurements X from omics pipelines of the size $(n \times m)$, where n is the number of samples and m is the total number of features, we construct a dataset $X^{(v)}$. $X^{(v)}$ is a collection of tables, where v is the index that corresponds to the view on the samples generated by a particular omics technology. The data is intrinsically high-dimensional and sparsely populated. Thus, we want to find a low-dimensional structure that the data points lie on [23]. For graph-based data integration, we construct an affinity graph $G = (V, E, W)$, where the vertices in V correspond to distinct samples $\{x_1, x_2, ..., x_n\}$ and $W^{(v)}$ is the matrix of edge weights for each view v that can represent i.e. gene expression, microRNA or metabolite levels. The edge weight is a measure for how similar two samples are in the omics-dimension that the graph is created for. Hence, this representation is suitable for clustering and instance-based learning techniques. The following sections introduces two methods for building and fusing graphs.

A. Similarity Network Fusion

Similarity Network Fusion (SNF) [21] creates a graph from each omics source individually and then combines them to a consensus network that captures both shared and complementary information about a sample.

$$K(i,j) = exp(-\frac{\delta_{ij}^2}{\gamma \sigma_{ij}}) \tag{1}$$

The kernel function (1) provides a measure of similarity between two samples x_i and x_j based on the distance measure δ_{ij}. In the following experiments the euclidean distance is used. In principal, any valid distance metric can be applied here. However, locally euclidean space is preferred such that the graph is a discrete approximation of a manifold [24].

$$\sigma_{ij} = \frac{1}{3}(\mu_i + \mu_j) + \frac{1}{3}\delta_{ij} \tag{2}$$

Then, $\gamma\sigma_{ij}$ is the parameter determining the scale of the kernel, where σ_{ij}[1] is estimated using (2) and γ is a hyperparameter to tune. The kernel in SNF is a radial basis function (RBF) with a coefficient γ and fixed scale σ, that is determined by the mean distances μ_i and μ_j of the two samples to their k-nearest neighbors respectively.

$$W^{(v)}(i,j) = \frac{K^{(v)}(i,j)}{\Sigma_{k \in N_i} K^{(v)}(i,k)} \tag{3}$$

The final similarity networks $W^{(v)}$ are computed using the kernel function normalized over k-neighbors of sample x_i, such that clusters of nodes that resemble closely to their neighbors are driven apart and smaller similarity values that are nevertheless within the k-nearest are brought closer together.

B. Affinity Network Fusion

Affinity Network Fusion (ANF) [22] estimates the similarity measure between two data points based on their k-nearest neighbors as well. Compared to SNF, the underlying similarity metric has been adjusted by using a Gaussian RBF kernel.

$$\mu_i = \frac{\Sigma_{l \in N_k(i)} \delta_{il}}{k} \tag{4}$$

First, the local sample means μ_i (4) are calculated for the k nearest neighbors of each node. The kNN-network is constructed in order to encapsulate local geometric structure.

$$\sigma_{ij} = \alpha(\mu_i + \mu_j) + \beta\delta_{ij} \tag{5}$$

In contrast to SNF, the two hyperparameters α and β estimate the scale of the kernel.

[1] The notation, particularly in equation (2) and (5), was changed from the original publication to match and compare between SNF and ANF.

$$K(i,j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp(-\frac{\delta_{ij}^2}{2\sigma_{ij}^2}) \tag{6}$$

Then, the kernel function (6) can be tuned in α and β to adjust for over/underfitting. α is a coefficient for the local diameter of two nodes while β is a coefficient for the pairwise distance metric. When only the k-nearest neighbors are computed, it creates a kNN Gaussian kernel.

$$W^{(v)}(i,j) = \frac{K^{(v)}(i,j)}{\Sigma_{j=1}^{N} K^{(v)}(i,j)} \tag{7}$$

Normalization is performed over all columns of the similarity matrices $W^{(v)}$, which makes it a valid probability distribution over the outgoing edges of each node. The authors argue that this leads to an improved representation of patient samples. An argument in favour of computing a similarity matrix from a Gaussian kernel is, that it has been shown to improve community structure discovery in networks [25].

C. Cross-Diffusion Process

Both graph-based data integration methods use metric enhancement through cross-diffusion [26] to combine $W^{(v)}$ into a fused matrix W. This can be seen as a transition process in a Markov chain, because W is a stochastic matrix. Each view v provides easily observable local metric information, as i.e. genetic or metabolic distance can be computed directly from the samples.

$$S(i,j) = \begin{cases} W(i,j), & \text{if } x_j \in KNN(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Notice that constructing S is a nonlinear operation (8), as for each row and column only k values are passed on. As a result, we obtain the kNN graphs $S^{(v)}$.

$$S^{(-v)} = \sum_{i \neq v} \frac{w_i}{\sum_{j \neq v} w_j} S^{(i)} \tag{9}$$

The complementary view computed in (9) is the weighted average of all other omics measurements except of the view itself. $W^{(-v)}$ is constructed the same way. The diffusion process is equivalent to a t step random walk. The projection is robust to noise and scales of data points and it incorporates the intrinsic structure of similarity manifold of the whole data set [27].

$$\begin{aligned} S^{(v)} = {} & \alpha_1(S^{(v)} \times S^{(-v)}) + \alpha_2(S^{(-v)} \times S^{(v)}) + \\ & \alpha_3(S^{(v)} \times W^{(-v)}) + \alpha_4(W^{(-v)} \times S^{(v)}) + \\ & \alpha_5(W^{(v)} \times S^{(-v)}) + \alpha_6(S^{(-v)} \times W^{(v)}) + \\ & \alpha_7(W^{(v)} \times W^{(-v)}) + \alpha_8(W^{(-v)} \times W^{(v)}) \end{aligned} \tag{10}$$

Matrix multiplication of the complementary kNN graph with the state matrix results in a stochastic matrix again. Hence, this diffusion step in (10) will output a valid affinity graph. The complementary graph of other omics representations is propagated through the kNN graph. Therefore, the weighted average of the other views is included in the local neighborhoods of view v. This results in a smoother version of $W^{(v)}$. The fused matrix W can be defined as the weighted average of the diffused views. This is when the process doesn't converge. In practice, the process is only repeated $r = 1, 2$ steps, otherwise W converges to rank one. Then, all the similarities of a data point to all of its neighbors would become constant. The coefficients α_i are binary variables that indicate which terms are included. Selecting only the first two terms will ignore the pruned edges in $W^{(v)}$, when noise is expected.

D. Graph Learning

After constructing the affinity graph $G = (V, E, W)$ the subsequent task is to make predictions about the samples or patients. This is based on the assumption of homophily as expressed by the weighted adjacency matrix W. Therefore, each node is assigned an embedding vector. It should capture the relationships between the nodes, the structure of the network, and task specific properties of the nodes. The new vector represents the position of node v relative to all other nodes in the network and can be used for instance-based learning. Computing the right representation is the goal for machine learning algorithms on graphs [28]. This suggests that there are a number of different approaches that can be used to compute an embedding for the affinity graph. The techniques that are presented here are based on matrix-factorization of the weighted adjacency matrix W of an undirected graph.

(1) Spectral clustering is a widely used methods to analyse similarity graphs [29]. The idea is to find a partition of the graph, such that the edges between the groups have low edge weights. The process can be used for graph embeddings in the following way.

$$L = D - W \tag{11}$$

Define D as the diagonal matrix of degrees, which is the sum of the edge weights for each node, as we are working with a weighted graph. Then, (11) is the unnormalized Laplacian of G. The Laplacian of a graph has interesting properties that are extensively studied in spectral graph theory [30].

$$L = Q\Lambda Q^{-1} \tag{12}$$

As the adjacency matrix and its Laplacian are real and symmetric, the corresponding eigendecomposition in (12) is a spectral decomposition. Thus, the eigenvectors of the graph Laplacian can then be used to construct an embedding [31].

(2) Another method to construct an embedding for G is to apply Generalized Singular Value Decomposition (GSVD) [32] to the weighted adjacency matrix W.

$$D_1^{-\alpha_1} W D_2^{-\alpha_2} = U\Sigma V \tag{13}$$

The node features are constructed from the left-singular vectors in $D_1^{-\alpha_1} U \Sigma^{1-\alpha}$. This represents the row embeddings by default. Note, that this embedding is not necessarily unique, because of the properties of SVD. The regularization factor α and the power factors α_i are not used by default. Afterwards, downstream prediction tasks can be applied.

III. Results

In the following, we compare the two tools Similarity Network Fusion (SNF) and Affinity Network Fusion (ANF) on a dataset containing mult-omics measurements from cancer cell lines. We need to estimate the density σ of the kernel functions $K(i, j)$. First, the influence of the scaling parameters γ for SNF as defined in (1) and α and β described for ANF in (5) is measured. These parameters change the measure of similarity within the two methods and their effects need to be determined. Next, the number of diffusion steps t, which is available as a free parameter in SNF, is increased within the data fusion process and its effect is measured. This allows us to reason about the implementational changes done in ANF and if they represent improvements to SNF. Thus, the fused networks can be constructed with the tuned hyperparameters. Then, Spectral analysis (11) of the final networks leads to embedding spaces that will be used for a subtype clustering task. Afterwards, the graph clusterings are applied to evaluate them for finding effective biomarkers that are predictive of drug sensitivity. These questions are being addressed with the following experimental setup.

A. Datasets

In this analysis, large-scale public datasets are used. The Cancer Cell Line Encyclopedia (CCLE) [33] contains data on cell line models that are used to study cancer biology. In that regard, the CCLE is similar to the NCI-60 [34] dataset for human cells. It provides a multi-omics view on the kind of cell lines that are often used to validate cancer targets or to identify drug efficacy in pre-clinical studies. The purity of the cell samples makes the dataset ideal for in-silico modelling. Therefore, information from the Cancer Therapeutic Response Portal (CTRPv2) [35] is added to complement the analysis. This repeated study on 860 CCLE cell lines recorded their sensitivity to 481 small-molecule compounds. The drugs are selected in order to affect a large number of cell functions.

TABLE I
CCLE Dataset

Cancer type	Disease type		Total
Breast	DC -*	24 23	**48**
Kidney	RCC ccRCC -*	7 8 8	**23**
Lung	SCLC NSCLC - AC	37 21 39	**172** 60
Pancreas	DC	22	**40**
Skin	MM -*	45 4	**48**
Stomach	AC -*	20 16	**37**

*Subtype not further specified by the CTRP.

Tab. I lists the types of cancer cells for which measurements from all omics technologies are available in the CCLE. The disease types annotated by the Broad Institute will serve as reference labels for subtype classification. However, the number of pre-defined clusters is limited, as labeling the data is time intensive and requires expert knowledge.

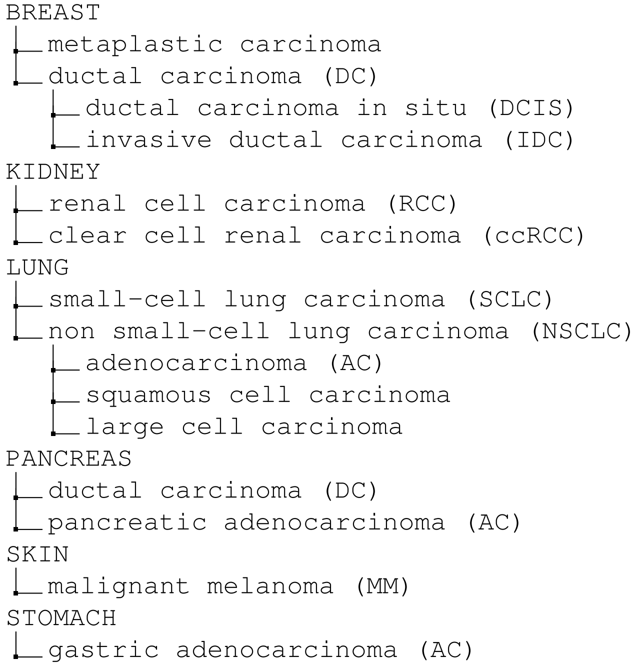
BREAST
metaplastic carcinoma
ductal carcinoma (DC)
ductal carcinoma in situ (DCIS)
invasive ductal carcinoma (IDC)
KIDNEY
renal cell carcinoma (RCC)
clear cell renal carcinoma (ccRCC)
LUNG
small-cell lung carcinoma (SCLC)
non small-cell lung carcinoma (NSCLC)
adenocarcinoma (AC)
squamous cell carcinoma
large cell carcinoma
PANCREAS
ductal carcinoma (DC)
pancreatic adenocarcinoma (AC)
SKIN
malignant melanoma (MM)
STOMACH
gastric adenocarcinoma (AC)

Fig. 1. Taxonomy of cancer types labelled in the CCLE dataset.

The hierarchy chart in Fig. 1 shows the most important disease types annotated in the dataset. Samples, for which there is only a single example are omitted. For some cancer types, a clear subdivision based on cancer histology is possible. For example, in lung carcinoma the aggressive small-cell carcinoma (SCLC) and the non small-cell carcinoma (NSCLC) type can be distinguished. In the kidney, clear cell renal cell carcinoma (ccRCC) are the most common and can be separated from renal cell carcinoma (RCC). This provides reliable labelling for network prediction tasks.

B. Affinity Graphs

To be able to find comprehensive clusters in the dataset, we need to evaluate how well affinity graphs can separate between potential clusters and how strongly related samples are connected within the clusters.

a) Connectivity: To measure the first quality, we can draw on results from spectral graph theory. We examine the second-smallest eigenvalue of the Laplacian (11) of the graphs constructed by SNF and ANF. The eigenvalue λ_1(L) of this matrix can be used as a measure for how tightly the graph clusters and is referred to as the algebraic connectivity of a graph as proposed by Fiedler in 1973 [36]. A first insight is, that if λ_1 is bigger than zero, then the graph is connected. The affinity graphs need to stay connected to provide useful information or else some samples cannot be assigned to a cluster. More specifically, we need to assure that all vertices have a degree above zero. Furthermore, λ_1 can be used as an approximation for the sparsest cut of the graph according to Cheeger's inequality [37].

$$\frac{1}{2}\lambda_1 \leq \Phi_G \leq \sqrt{2\lambda_1} \tag{14}$$

It constructs a bound on the conductance Φ_G of a graph (14), which is a constant that describes how easily one part of a network can be reached from another by a random walk. This property makes λ_1 a good measure for graph clustering, as it helps to minimize the edges between two potential partitions [38]. Hence, we observe how the changes in the hyperparameters for constructing the graphs in SNF and ANF affect this property.

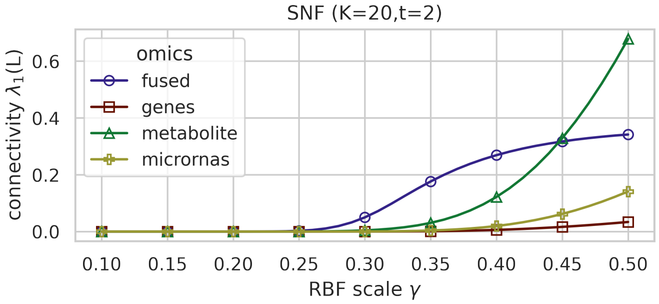

Fig. 2. Smallest non-zero eigenvalue λ_1(L) of the Laplacian for the scale γ of the radial basis function kernel (genes + metabolites + micrornas).

The scale parameter γ of the RBF kernel (1) is changed over a range of possible values to produce similarity networks for gene expression, metabolite levels and microRNAs of the cell lines. Fig. 2 shows, that γ changes the topology of the network drastically over the inspected range. At low values of γ, SNF produces affinity graphs that are not connected. On one hand, the genetic similarity network stays at a connectivity below 0.10 for higher values of γ. On the other hand, the metabolite similarity network shows a high connectivity of above 0.60 at the default value of $\gamma = 0.50$ for SNF. At the same point, fused network flattens to a value of $\lambda_1 = 0.34$, which bounds the conductance Φ_G at a larger range of 0.17 to 0.83 compared to $\gamma = 0.30$, where $\lambda_1 = 0.05$ and Φ_G can be bound between 0.025 and 0.32.

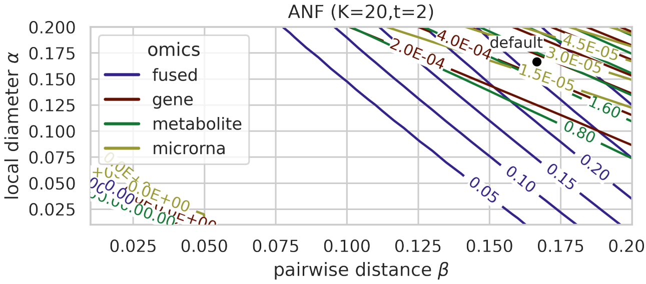

Fig. 3. Albebraic connectivity λ_1(L) of the affinity networks for scale parameter α and β of the Gaussian RBF as used in ANF.

The same measurements are repeated for the two hyperparameters α and β that determine the scale (5) in ANF. Therefore, we measure the algebraic connectivity of the resulting affinity graphs for combined values over a range of α and β. The two parameters α and β are clearly linearly related as shown in Fig. 3. For values within a region between 0.1 and 0.2 marked by the countours in the plot, the selection of the parameters has a large influence on the algebraic connectivity of the final affinity graphs. The default scaling parameters $\alpha = \frac{1}{6}$ and $\beta = \frac{1}{6}$ of ANF correspond to the lower end of this region. For the individual omics network the relationship between the hyperparameters is approximately the same, whereas the fused network is less sensitive to the scale on the pairwise distance β. Here, higher levels of connectivity can be obtained even with a sparser network.

b) Modularity: measures the internal density and external sparsity of a labelled partition. Hence, maximizing modularity can be used as an approach for community detection. However, this method can overfit on the noise within the dataset, especially if the network is sparse [39].

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right] \tag{15}$$

For this section, the definition of modularity in (15) is used, where L_c is the number of links within a cluster, m is the total number of edges and k_c is the sum of degrees for the nodes in that cluster [40]. The value of $Q(G)$ is measured for the partitioning of the graph into carcinoma cell lines, where 20% of the dataset that are labelled as lung cancer types are held out for further evaluation.

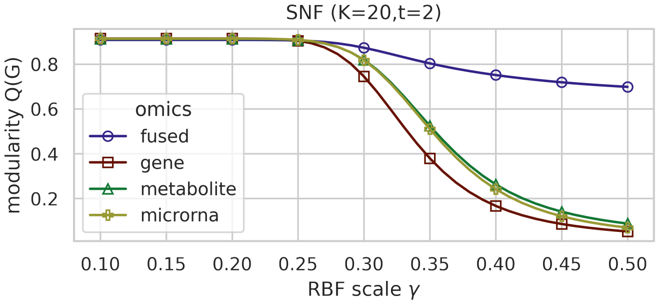

Fig. 4. Modularity Q(G) of the affinity graph constructed by SNF for the scale γ of the radial basis function kernel (gene + metabolite + microrna).

Again, the scale parameter γ is changed over the inspected range and the modularity is recorded. Fig. 4 shows a drop in modularity as the algebraic connectivity of the fused network increases. While the modularity of the individual omics networks decreases to below 0.2 as γ approaches 0.5, the fused network stays more modular. At $\gamma = 0.3$ the modularity is the greatest. Hence, SNF can be tuned to produce a fused network with high modularity and connectedness.

Similarly, in Fig. 5 the drop in modularity is at the same range of value for α and β as in Fig.3 when the algebraic connectivity increases. The effect of α and β on the modularity

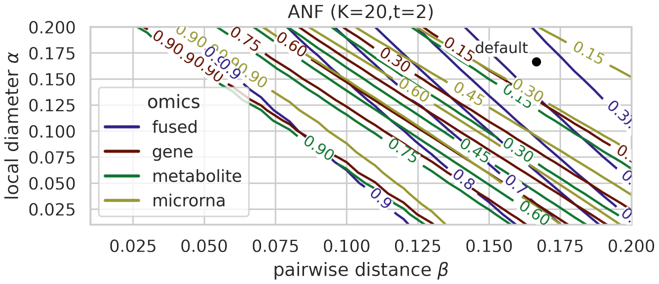

Fig. 5. Modularity Q(G) of the affinity graph for scale parameter α and β of the Gaussian RBF used in ANF.

is approximately equal for the individual and the multi-omics affinity graphs. With small values of the scale parameters in both SNF and ANF modularity increases to around 90%. However, λ_1 approaching zero indicates that the graph is disconnected. Hence, we choose α, β and γ such that the fused affinity graphs have high modularity while preserving connectivity.

C. Data Fusion

The data fusion process is designed to keep similarities that are shared across the different data sources and to add strong similarities that can only be found in one of the views. To test the convergence of SNF the parameters are set to the default values ($K = 20$ and $\gamma = 0.5$) and the process is stopped after t diffusion steps.

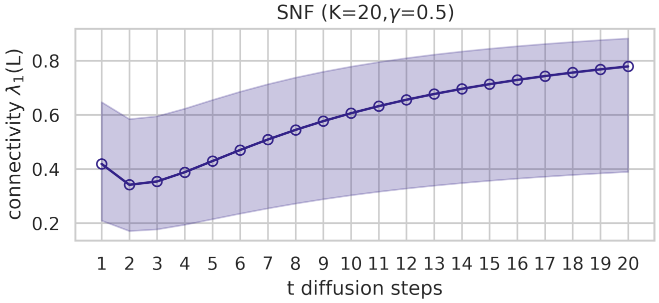

Fig. 6. Algebraic connectivity λ_1(L) of the fused adjacency matrix W after applying SNF for t diffusion steps (genes + metabolites + micrornas).

For $t = 2$, the algebraic connectivity λ_1 decreases compared to the baseline of taking the weighted average of the omics views at $t = 1$. At larger values for t more edges are reinforced, which consequentially increases conductance between more loosely related parts of the graph. Hence, larger values for t can lead to high similarity value to a sample's k-nearest neighvors even if the samples are only loosely related. In particular, when the data fusion process converges, intra-cluster differences vanish. A decrease in connectivity albeit more edges are introduced speaks for increased clustering quality of the network. In the ANF implementation only the values $t = 1$ and $t = 2$ are available, which corresponds to the terms for a two-step diffusion process (10).

After tuning the parameters of the fusion process the resulting networks can be visualized using a force-directed layout. Therefore, the Fruchterman-Reingold algorithm [41] is used, where the edge weights represent an attracting force that pulls nodes with high similarities together.

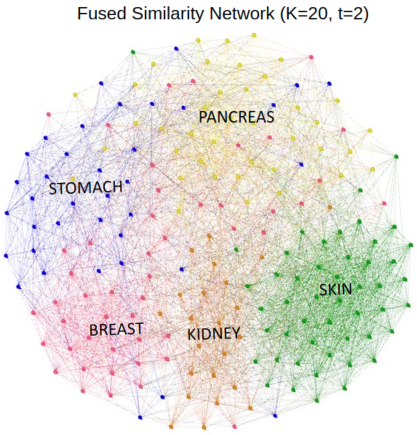

Fig. 7. Fused network from SNF after $t = 2$ diffusion steps with $k = 20$ nearest neighbors and scaling factor $\gamma = 0.5$ (genes + metabolites + mirnas).

Fig. 7 shows the network constructed by SNF. The similarity measure for the samples from kidney cells clearly benefits from the combined view. The cluster of pancreatic cancer cells is more diffused in the combined multi-omics network and those cells have a high similarity to other cancer types.

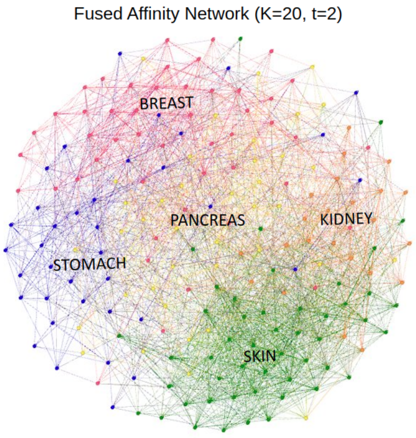

Fig. 8. Fused network from ANF after $t = 2$ diffusion steps with $k = 20$ nearest neighbors (genes + metabolites + mirnas).

Fig. 8 presents the fused affinity matrix. ANF produces a network with sparser edge weights within the clusters. The clusters for skin and kidney cancer types remain clear, whereas pancreas cancer cells share high similarity with the samples from the rest of the network.

D. Prediction Tasks

Subsequently, the information captured in the affinity graph needs to be extracted. Therefore, the samples in the CCLE dataset are clustered based on their similarity and the resulting groups are evaluated in three different ways. First of all, the histology and subtype annotations of the CCLE in Fig.1 are used as external information to calculate quality metrics on the test data. This is done for clusters created from SNF and ANF in order to compare the two methods as well as for the different embedding algorithms. Additionally, a randomization test is used for cluster validation. Finally, we evaluate the clusters based on additional data about drug sensitivities of the cell lines from the CTRPv2 to compare between single omics and multi-omics clustering. Drug compounds are screened for differences between the clusters and the partitions are evaluated as potential multi-omics biomarkers.

E. Clustering

The goal of the first task is to correctly cluster subtypes for lung and kidney cells that are annotated and listed in Tab. I. The lung cancer type represents 20% of the CCLE dataset and is held out as a test set. Quality is measured by purity, normalized mutual information and adjusted rand score. As subtypes within the cancer types are balanced, purity can be used equivalently to accuracy as a suitable performance metric. To validate the clusterings, edge weights are randomly permutated to construct fused affinity graphs with the same distribution for the values in the adjacency matrix as SNF and ANF. Then, p-values can be calculated for each metric to describe how likely the cluster is produced by chance. For both methods, if $t = 1$ is chosen, the weighted average of the individual networks is used in place of the two-step diffusion process (10). Hence, this can be used to compare to data fusion step to a simple combination of the omics-graphs.

TABLE II
GRAPH CLUSTERING

Methods		**Cancer type**	
		Lung	***Kidney***
ANF $t = 2$ ($\alpha = 0.166$, $\beta = 0.166$)	Purity	0.87 *(p≤1.0e-05)*	0.73 *(p=5.1e-02)*
	ARI	0.54 *(p≤1.0e-05)*	0.05 *(p=0.177)*
	NMI	0.46 *(p≤1.0e-05)*	0.16 *(p=9.4e-02)*
ANF $t = 1$ ($\alpha = 0.166$, $\beta = 0.166$)	Purity	0.84 *(p≤1.0e-04)*	0.80 *(p=7.1e-03)*
	ARI	0.50 *(p≤1.0e-04)*	0.31 *(p=2.9e-02)*
	NMI	0.46 *(p≤1.0e-04)*	0.29 *(p=7.0e-03)*
SNF $t = 2$ ($\gamma = 0.3$)	Purity	0.91 *(p≤1.0e-03)*	0.60 *(p=0.29)*
	ARI	0.66 *(p≤1.0e-03)*	-0.03 *(p=0.61)*
	NMI	0.59 *(p≤1.0e-03)*	0.03 *(p=0.28)*
SNF $t = 1$ ($\gamma = 0.3$)	Purity	0.92 *(p≤1.0e-03)*	0.53 *(p=0.59)*
	ARI	0.71 *(p≤1.0e-03)*	-0.06 *(p=0.60)*
	NMI	0.67 *(p≤1.0e-03)*	0.01 *(p=0.61)*
KMeans *(gene, metabolite, miRNA, fused)*	Purity	0.74,0.72,0.74,0.74	0.8,0.73,0.93,0.93
	ARI	0.21,0.19,0.19,0.19	-0.04,-0.06,0.02,0.02
	NMI	0.17,0.18,0.21,0.21	0.02,0.001,0.11,0.11

The comparison between those methods can be seen in Tab. II. Adenocarcinoma and small-cell carcinoma in the lung can be clustered with a purity of 87% from the network of lung cancer cells constructed by ANF. SNF achieves 92% accuracy on the lung cancer cells, where the two-steps and one step fusion precess performs equally well. SNF cannot separate the relatively small number of renal cell carcinoma and clear cell renal cell carcinoma. Both can't establish statistical significant partitions for the kidney cell type.

F. Biomarkers

Afterwards, the graph clustering methods are applied to the networks to compare potential separations of the cell lines. As samples are grouped into clusters based on their phenotypic similarity, cells with related traits should lie closer to each other on the graph. Therefore, similar drug response properties speak for a high quality of the clustering results. Again, small cell carcinoma (SCLC) and adenocarcinoma (AC) in the lung are picked as groups to be distinguished. Furthermore, other carcinoma types from Tab. I are added to the analysis to construct a comparable baseline. In particular, the drug sensitivity of gastric adenocarcinoma (AC) relative to ductal breast carcinoma (DC) is considered. Thus, a drug screening process is performed for these cell lines to narrow down the list of potential drug candidates.

TABLE III
DRUG CANDIDATES

Cancer type	***compound***	***p-value adjusted***	***logfoldchange***
Lung	dasatinib	$5.4\mathrm{e}{-5}$	0.42
	vincristine	$5.6\mathrm{e}{-5}$	0.65
Breast	doxorubicin	$7.5\mathrm{e}{-140}$	0.11
	leptomycin B	$2.5\mathrm{e}{-84}$	0.19

The 481 available compounds are scanned for a fold change in sensitivity and its statistical significance across the cancer types to produce a list of candidates shown in Tab. III. The Benjamini-Hochberg procedure [42] was applied to correct for multiple hypothesis testing. Then, the drug resistance for different clusters can be compared using a Welch's t-test.

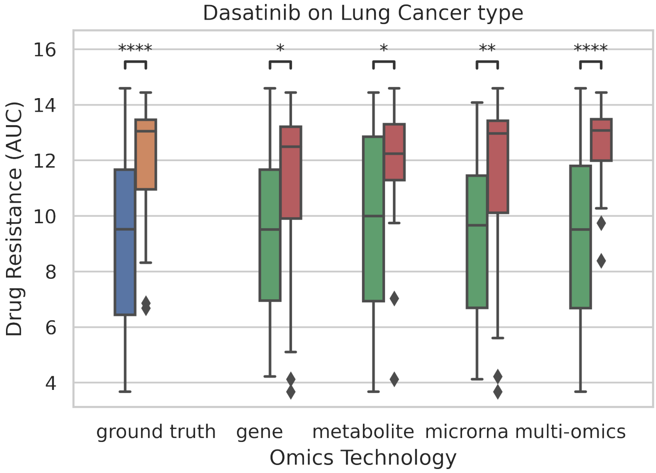

Fig. 9. Combined multi-omics clusters from ANF with K=20, local diameter α=0.166, pairwise distance β=0.166 and t=2 (gene + metabolite + mirna).

Fig. 9 shows the drug resistance to dasatinib for adenocarcinoma (AC) and small cell lung carcinoma (SCLC) clusters found in ANF. For 10 out of the 37 SCLC cell lines measurements for dasatinib are missing, which creates an unequal samples size. Neither the gene ($p = 1.526\mathrm{e}{-02}$) nor the metabolite ($p = 3.949\mathrm{e}{-02}$) affinity graph can be used to identify drug resistant small cell lung carcinoma (SCLC) at a significance level of 1%. Clustering based on microRNA expression at $p = 1.152\mathrm{e}{-02}$ doesn't separate the samples at the level of the ground truth either. Only the partition of the fused multi-omics network leads to the most significant division between drug sensitive and resistant cell lines with a p-value of $p = 6.686\mathrm{e}{-06}$.

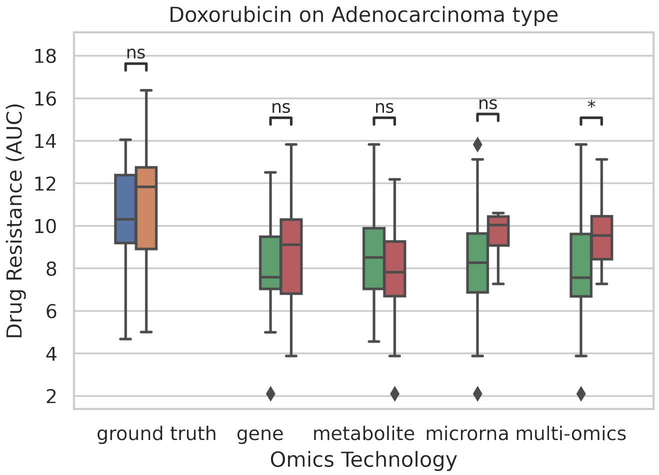

Fig. 10. Combined multi-omics clusters from ANF with K=20, local diameter α=0.166, pairwise distance β=0.166 and t=2 (gene + metabolite + mirna).

In Fig. 10 the drug response to doxorubicin is shown for gastric adenocarcinoma (AC) and ductal carcinoma (DC). While none of the clusterings on the individual omics networks creates a significant partition between drug resistant and drug sensitive cell lines, genomics at $p = 0.29$, metabolomics at $p = 0.20$ and microRNA with $p = 0.30$, the cluster separation in the multi-omics network reaches a p-value of $p = 3.54\mathrm{e}{-02}$.

IV. Discussion

To compare graph-based data integration methods, we have examined the influence of the hyperparameters in SNF and ANF on the clustering tendency of affinity graphs. When constructing the multi-omics networks, SNF and ANF are both sensitive to the choice of the scale α and β or γ in their RBF kernels and the parameters can be tuned within the regions shown in Fig. 2 and Fig. 3. With a smaller scaling factor γ the absolute differences in the distance metric δ_{ij} are weighted more strongly, as the same range in the distance is mapped to a larger range in similarity. The same is true for α and β in ANF. Both methods can produce affinity networks with an algebraic connectivity around $\lambda_1 = 0.2$, which bounds the conductance Φ_G of the fused network to a range of approximately $[0.1, 0.63)$. On one hand, constructing networks with larger λ_1 increases the distance between the bounds on Φ_G defined by Cheeger's inequality (14), which decreases clustering quality by fusing parts of unrelated clusters. On the other hand, decreasing the algebraic connectivity λ_1 further will reduce the representation from a complete to a disconnected graph, which splits up known clusters in the graph. This represents a new approach of estimating the parameters for the two methods by maximizing modularity and connectivity of the affinity graphs. Fig. 4 shows that for the fused network modularity is preserved even as the algebraic connectivity in Fig. 2 increases. In the subsequent analysis both methods perform better than the baseline of concatenating the raw omics measurements on the sample clustering task, which indicates that shared and complementary information is indeed preserved in the multi-omics networks. For this purpose, SNF has been successfully used to i.e. prepare clusters of samples to identify clinically relevant *KEAP1/NFE2L2*-mutant lung adenocarcinoma [43]. Doxorubicin is an active drug used in the treatment of breast cancer, where low levels of expression of its target gene topo II-alpha correlate with drug resistance in cell lines [44]. Hence, the drug appears in Tab. III as a potential compound that has different effects across the clusters. Dasatinib is a *LIMK1* inhibitor that has been shown to supress the proliferation of lung cancers with high expression of that gene [45]. Fig. 9 shows that the multi-omics cluster can identify resistant cells with high significance. Compared to other data integration methods, neither SNF nor ANF allow to inspect the influence of individual features on the final metrics used in the network, which makes it harder to elucidate molecular mechanisms behind the cellular differences. Further research can address the potential of affinity graphs in a semi-supervised learning context [46]. In contrast to the sample clustering tasks, that SNF and ANF have been used for extensively, the similarity matrix can also be used as a constraint, such that similar samples are promoted to be assigned close values by a predictor function [47]. This presents the opportunity to combine graph-based models, with their ability to fuse data from different omics technologies, with more complex estimators such as ensembles methods that work on the original features. Hence, using SNF and ANF for regularization.

V. Conclusion

Multi-omics analysis is becoming an increasingly important tool to discover clusters with clinically relevant features. When implemented in a comprehensive toolset, researchers who need to analyse and prepare multi-omics samples could benefit greatly from the graph-based methods presented in this paper. The main differences between SNF and ANF lie in the similarity measure that is used and in how it is normalized. This results into sparser graphs for ANF and lower intra-cluster differences of the similarity measure for SNF. Additionally, the results from using SNF and ANF show potential for these methods to be used for regularization by graphs, opening up opportunities to use graph-based methods in combination with other models in a semi-supervised manner.

References

[1] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[2] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, and F. Stahl, "Transcriptome analysis using next-generation sequencing," *Current opinion in biotechnology*, vol. 24, no. 1, pp. 22–30, 2013.

[3] B. Zhou, J. F. Xiao, L. Tuli, and H. W. Ressom, "Lc-ms-based metabolomics," *Molecular BioSystems*, vol. 8, no. 2, pp. 470–481, 2012.

[4] M. Olivier, R. Asmis, G. A. Hawkins, T. D. Howard, and L. A. Cox, "The need for multi-omics biomarker signatures in precision medicine," *International journal of molecular sciences*, vol. 20, no. 19, p. 4781, 2019.

[5] Z. Ahmed, "Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis," *Human genomics*, vol. 14, no. 1, pp. 1–5, 2020.

[6] E. H. Romond, E. A. Perez, J. Bryant, V. J. Suman, C. E. Geyer Jr, N. E. Davidson, E. Tan-Chiu, S. Martino, S. Paik, P. A. Kaufman *et al.*, "Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer," *New England journal of medicine*, vol. 353, no. 16, pp. 1673–1684, 2005.

[7] S. Cascinu, R. Berardi, R. Labianca, S. Siena, A. Falcone, E. Aitini, S. Barni, F. Di Costanzo, E. Dapretto, G. Tonini *et al.*, "Cetuximab plus gemcitabine and cisplatin compared with gemcitabine and cisplatin alone in patients with advanced pancreatic cancer: a randomised, multicentre, phase ii trial," *The lancet oncology*, vol. 9, no. 1, pp. 39–44, 2008.

[8] B. Palsson and K. Zengler, "The challenges of integrating multi-omic data sets," *Nature chemical biology*, vol. 6, no. 11, pp. 787–789, 2010.

[9] R. Cavill, D. Jennen, J. Kleinjans, and J. J. Briedé, "Transcriptomic and metabolomic data integration," *Briefings in bioinformatics*, vol. 17, no. 5, pp. 891–901, 2016.

[10] G. W. Tillinghast, "Microarrays in the clinic," *Nature biotechnology*, vol. 28, no. 8, pp. 810–812, 2010.

[11] S. Jozefczuk, S. Klie, G. Catchpole, J. Szymanski, A. Cuadros-Inostroza, D. Steinhauser, J. Selbig, and L. Willmitzer, "Metabolomic and transcriptomic stress response of escherichia coli," *Molecular systems biology*, vol. 6, no. 1, p. 364, 2010.

[12] A. Acharjee, B. Kloosterman, R. C. de Vos, J. S. Werij, C. W. Bachem, R. G. Visser, and C. Maliepaard, "Data integration and network reconstruction with omics data using random forest regression in potato," *Analytica chimica acta*, vol. 705, no. 1-2, pp. 56–63, 2011.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[15] M. S. Ghaemi, D. B. DiGiulio, K. Contrepois, B. Callahan, T. T. Ngo, B. Lee-McMullen, B. Lehallier, A. Robaczewska, D. Mcilwain, Y. Rosenberg-Hasson *et al.*, "Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy," *Bioinformatics*, vol. 35, no. 1, pp. 95–103, 2019.

[16] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[17] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes. 599 nucleic acids res 28: 27-30," 2000.

[18] A. Kamburov, R. Cavill, T. M. Ebbels, R. Herwig, and H. C. Keun, "Integrated pathway-level analysis of transcriptomics and metabolomics data with impala," *Bioinformatics*, vol. 27, no. 20, pp. 2917–2918, 2011.

[19] 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.

[20] M. Lovino, V. Randazzo, G. Ciravegna, P. Barbiero, E. Ficarra, and G. Cirrincione, "A survey on data integration for multi-omics sample clustering," *Neurocomputing*, vol. 488, pp. 494–508, 2022.

[21] 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, pp. 333–337, 2014.

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

[23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[24] F. Waldhausen, "On irreducible 3-manifolds which are sufficiently large," *Annals of Mathematics*, pp. 56–88, 1968.

[25] C. Guo and H. Zhao, "Community structure discovery method based on the gaussian kernel similarity matrix," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 6, pp. 2268–2278, 2012.

[26] Y. Wang, W. Zhang, L. Wu, X. Lin, and X. Zhao, "Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 1, pp. 57–70, 2015.

[27] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[28] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[29] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[30] F. R. Chung and F. C. Graham, *Spectral graph theory*. American Mathematical Soc., 1997, no. 92.

[31] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[32] C. F. Van Loan, "Generalizing the singular value decomposition," *SIAM Journal on numerical Analysis*, vol. 13, no. 1, pp. 76–83, 1976.

[33] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[34] R. H. Shoemaker, "The nci60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer*, vol. 6, no. 10, pp. 813–823, 2006.

[35] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould *et al.*, "Harnessing connectivity in a large-scale small-molecule sensitivity dataset," *Cancer discovery*, vol. 5, no. 11, pp. 1210–1223, 2015.

[36] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak mathematical journal*, vol. 23, no. 2, pp. 298–305, 1973.

[37] N. Alon, "Eigenvalues and expanders," *Combinatorica*, vol. 6, no. 2, pp. 83–96, 1986.

[38] N. M. M. De Abreu, "Old and new results on algebraic connectivity of graphs," *Linear algebra and its applications*, vol. 423, no. 1, pp. 53–73, 2007.

[39] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, "Modularity from fluctuations in random graphs and complex networks," *Physical Review E*, vol. 70, no. 2, p. 025101, 2004.

[40] 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.

[41] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[42] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[43] X. Yang, M. Li, Z. Chen, X. Fan, L. Guo, B. Jin, Y. Huang, Q. Wang, L. Wu, and C. Zhan, "Multi-omics analysis identifies distinct subtypes with clinical relevance in lung adenocarcinoma harboring keap1/nfe2l2." *Journal of Cancer*, vol. 13, no. 5, pp. 1512–1522, 2022.

[44] B. J. Lynch, D. G. Guinee Jr, and J. A. Holden, "Human dna topoisomerase ii-alpha: a new marker of cell proliferation in invasive breast cancer," *Human pathology*, vol. 28, no. 10, pp. 1180–1188, 1997.

[45] M. Zhang, J. Tian, R. Wang, M. Song, R. Zhao, H. Chen, K. Liu, J.-H. Shim, F. Zhu, Z. Dong *et al.*, "Dasatinib inhibits lung cancer cell growth and patient derived tumor growth in mice by targeting limk1," *Frontiers in Cell and Developmental Biology*, p. 1361, 2020.

[46] X. J. Zhu, "Semi-supervised learning literature survey," *Technical report 1530*, 2005.

[47] D. Zhou and B. Schölkopf, "A regularization framework for learning from graph data," in *ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields (SRL 2004)*, 2004, pp. 132–137.