Multi-omics Data Integration by Generative Adversarial Network

Khandakar Tanvir Ahmed [1,2], Jiao Sun [1,2], Jeongsik Yong [3] and Wei Zhang [1,2]*

[1]Department of Computer Science, [2]Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA, and [3]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA

Accurate disease phenotype prediction plays an important role in the treatment of heterogeneous diseases like cancer in the era of precision medicine. With the advent of high throughput technologies, more comprehensive multi-omics data is now available that can effectively link the genotype to phenotype. However, the interactive relation of multi-omics datasets makes it particularly challenging to incorporate different biological layers to discover the coherent biological signatures and predict phenotypic outcomes. In this study, we introduce omicsGAN, a generative adversarial network (GAN) model to integrate two omics data and their interaction network. The model captures information from the interaction network as well as the two omics datasets and fuse them to generate synthetic data with better predictive signals. Large-scale experiments on The Cancer Genome Atlas (TCGA) breast cancer and ovarian cancer datasets validate that (1) the model can effectively integrate two omics data (i.e., mRNA and microRNA expression data) and their interaction network (i.e., microRNA-mRNA interaction network). The synthetic omics data generated by the proposed model has a better performance on cancer outcome classification and patients survival prediction compared to original omics datasets. (2) The integrity of the interaction network plays a vital role in the generation of synthetic data with higher predictive quality. Using a random interaction network does not allow the framework to learn meaningful information from the omics datasets; therefore, results in synthetic data with weaker predictive signals.

multiomics data| data integration | generative adversarial network | multi-omics interaction network

Correspondence: ***wzhang.cs@ucf.edu***

Introduction

Complex diseases such as cancer are highly heterogeneous with different subtypes leading to varying clinical outcomes including prognosis, response to treatment, and changes of recurrence and metastasis (1–3). Disease phenotype prediction has been the subject of interest to clinicians and patients for many decades. Disease phenotypes depend on molecular profiles and interplay at genomic, epigenomic, transcriptomic, proteomic, and metabolomic levels (4) which are connected with each other through a complex network (5). Therefore, phenotype prediction based on a single genotype data will fail to characterize the intricate mechanism of biological processes and result in an inaccurate prediction.

The recent development in high throughput sequencing technologies is capable of measuring molecular activities in cells and allows researchers to obtain multi-omics data with sufficient quality and yield (6). It has revolutionized medical and biological research by offering a more comprehensive view of the interplay of molecules and identify accurate molecular signatures for characterizing or predicting disease phenotypes. Analysis of multi-omics data along with clinical information of patients can help us to establish a link from genotype to phenotype by exploring the flow of information within different omics layers (4). For instance, (7) showed the importance of integrating proteomics data along with genomic and transcriptomic data to identify driver genes in colon and rectal cancers. They state the fact that the correlation between mRNA and protein abundance in two-thirds of the cases in The Cancer Genome Atlas (TCGA) studies (8) is not statistically significant. Moreover, copy number alteration (CNA) drive mRNA abundance changes but rarely translate into changes in protein abundance. Therefore, using proteomics data along with mRNA abundance, CNA, and other genomic and transcriptomic data will add crucial information for the investigation which otherwise would be ignored. Several other studies have inferred the importance of multi-omics data integration in different lines of research (9–12).

Multi-omics data generated for the same set of samples often have an interactive relation among them. Different profiling datasets provide non-redundant predictive signals for predicting therapeutic response. Removing one of them in a prediction model will lead to performance degradation (13). So it is possible that multi-omics data provide a complementary set of information to understand the molecular basis of diseases. Therefore, predicting the phenotype using the multi-omics data as independent sets of features will fall short in characterizing the prediction. For instance, microRNA (miRNA) regulates mRNA expression by complementarily binding to recognition sequences in the 3' untranslated region of their target mRNAs leading to mRNA degradation and/or mRNA translation inhibition. The abundance of a particular miRNA does not illustrate the full picture without knowing which mRNAs get inhibited by that miRNA; because miRNA does not directly influence the phenotype; rather, regulates the mRNA translation into protein that subsequently determines the phenotype. Hence, integrating the interaction network into analysis will capture the regulatory effect and establish a better correlation with the phenotype. Moreover, mRNA can be regulated by other modulators like RNA binding protein (RBP) (14). RBPs bind RNA through globular RNA-binding domains (RBDs) and alter the function of the bound RNAs (15). RNA-RBP interaction obtained from crosslinking and immunoprecipitation based CLIP-Seq can also be used to characterize the relation between omics

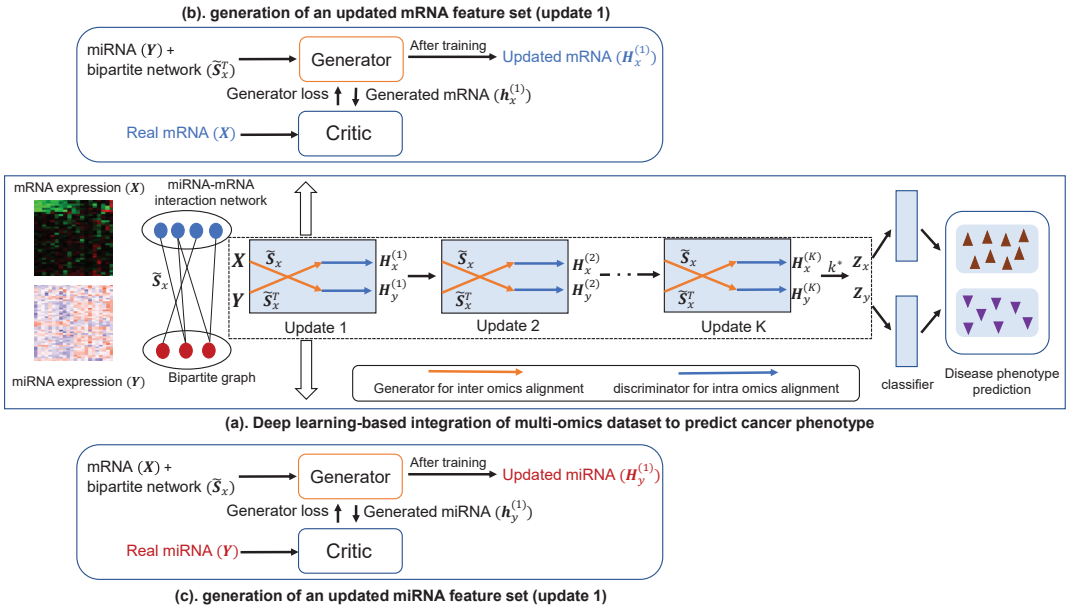

Fig. 1. (a) An illustration of the proposed generative adversarial framework (omicsGAN). Two omics datasets are updated once in each box through an adversarial game between the generator (marked by orange line) and critic (marked by blue line). Generator and critic are trained for each omics data independently and the updated datasets are applied for disease phenotype prediction. **(b) Update of mRNA feature set.** Generator uses miRNA expression data and miRNA-mRNA bipartite network to synthesize an mRNA expression data. Both synthetic and input mRNA expression data are passed through a critic that tries to differentiate the real and synthetic data. **(c) Update of miRNA feature set.** Generator uses mRNA expression data and miRNA-mRNA bipartite network to synthesize an miRNA expression data. Both synthetic and input miRNA expression data are passed through a critic that tries to differentiate the real and synthetic data.

data. There are several advanced multi-omics data integration frameworks, however, few approaches link different omics data sets using molecular interaction as a data feature (16) or consider the relation among features within an omics data (17). Most of them ignore the relations within and across different biological layers (e.g., gene co-expression network, miRNA-mRNA interaction network) in their analysis (18–21). The power of high throughput technologies cannot be fully utilized unless the multi-omics data with its intra and inter modal relations are considered in studies.

In recent years, generative adversarial networks (GAN) (22) has gained popularity in solving problems within the scope of computational biology. A traditional GAN has two parts: generator and discriminator; the generator learns to generate plausible data imitating a real dataset whereas the discriminator learns to distinguish the generator's synthetic data from the real data. After successful training in a competitive manner between the generator and discriminator, the generator will be able to create synthetic data with similar distribution as the real data that can fool the discriminator. There are several studies that use GAN based algorithms to generate data from single or multiple datasets. (23) used GAN for better biomarkers identification by generating a reconstructed functional interaction network from multi-omics datasets. (24) integrated diverse single-cell RNAseq (scRNA-seq) datasets from different labs and experimental protocols to simulate a realistic scRNA-seq data that covers the full cell type diversity. (25) on the other hand used GAN to generate gene expression from bulk RNA-seq datasets. A framework to eliminate batch effect from scRNA-seq data using GAN was proposed in (26). Recently, (27) designed an algorithm to formulate bipartite network propagation in terms of the generative adversarial network. Wasserstein GAN, a more stable version of GAN was proposed in (28) that uses Wasserstein distance to calculate the loss functions. It uses a critic in place of the discriminator which instead of classification, tries to make the output bigger for real instances than for generated ones and optimize the training in a similar competitive manner. GANs can learn non-linear relationships between features of omics data during training that can be used later for additional insight (24). It can handle missing data and also promising for missing value imputation because of its capability of learning and imitating any distribution of data (29).

In this study, we propose a biologically-motivated deep learning-based model, omicsGAN, to predict disease phenotype by integrating two omics data and the interaction between them (e.g., mRNA expression, miRNA expression, and miRNA-mRNA interaction network). The proposed model introduces a generative adversarial method to generate a new enriched feature set for each omics data combining information from the other omics dataset and the interaction network resulting in a better prediction. Experimental results verify that our proposed framework generates datasets with stronger molecular signatures to better understand the biological mechanism that leads to the disease state and improve disease outcome prediction compared to the biological features derived from only mRNA expression and miRNA expression.

Methods

In this section, we first introduce the mathematical notations employed in this study, followed by a two step framework, omicsGAN, for predicting disease outcomes using multi-

omics data. The framework can take any two omics data with biological relations between each other as input. In this study, we used miRNA, mRNA, and miRNA-mRNA interaction network in the data analysis. We then describe the first step of this framework, a generative adversarial network model that integrates two omics data and their interaction network to generate synthetic data with a new feature set for each omics profile. We also briefly outline the deep learning-based classifier that uses the synthetic data to predict the disease phenotype. Lastly, we discuss the evaluation metrics and algorithms used in this study to analyze the molecular signatures represented by the synthetic datasets.

Overview of the framework. For the multi-omics data analysis, using extra omics data as an independent feature set provides additional information for downstream analysis. However, different omics profiles are often not independent rather linked with each other through a complex biological interaction network. Our proposed framework, omicsGAN, can capture the information from this inter-omics network and integrate it with the omics datasets through a generative adversarial network to update them iteratively. After successful training of the network, it will generate new feature sets corresponding to each omics data that contain information from both modality and their interaction network. In this study, the framework is tested on mRNA and miRNA expression datasets; however, this framework can work with any two omics data that are related to each other, given that their interaction can be converted into an adjacency matrix. mRNA and miRNA expression are correlated to disease phenotype, although, the bipartite interaction network between them can be leveraged into increasing the correlation by incorporating miRNA regulation on mRNA translation. mRNAs directly influence phenotype by translating into proteins that control all physiological activities in a cell; however, miRNA binds to mRNA and regulates its translation into protein, thus indirectly controls the phenotype. From a biological point of view, knowing the expression of a miRNA does not provide enough information without knowing the mRNAs that it targets. For an accurate and realistic downstream analysis, realizing the interaction between omics data into calculation is crucial as well as challenging for the researchers.

The notations to define the proposed model, omicsGAN, are summarized in Table 1. Let $\boldsymbol{N}$ be the adjacency matrix of miRNA-mRNA interaction network and the dimension of the network is $p \times m$, where p is the number of miRNAs and m is the number of mRNAs. The dimensions of the mRNA ($\boldsymbol{X}$) and miRNA ($\boldsymbol{Y}$) expression data are $m \times n$ and $p \times n$ respectively, with n being the number of samples. Updated (synthetic) mRNA ($\boldsymbol{H}_x^{(k)}$) and miRNA ($\boldsymbol{H}_y^{(k)}$) where $k \in \{1, 2, 3, ..., K\}$, will correspond to the dimension of the input mRNA and miRNA expression datasets respectively and K is the total number of updates in omicsGAN.

In this study, we predict cancer outcome using mRNA, miRNA, and the interaction network between them as illustrated in Figure 1(a). The framework takes mRNA ($\boldsymbol{X}$), miRNA ($\boldsymbol{Y}$), and normalized interaction network ($\tilde{\boldsymbol{S}}$) as input and iteratively updates them to find two new feature sets that incorporates information from both omics data and their biological interactions, where $\tilde{\boldsymbol{S}} = \boldsymbol{D}_Y^{-\frac{1}{2}} \boldsymbol{N} \boldsymbol{D}_X^{-\frac{1}{2}}$. $\boldsymbol{D}_X$ and $\boldsymbol{D}_Y$ are two diagonal matrices with $\boldsymbol{D}_X(i,i) = \sum_j |\boldsymbol{N}(j,i)|$ and $\boldsymbol{D}_Y(i,i) = \sum_j |\boldsymbol{N}(i,j)|$. A deep learning-based classifier is then applied on the new feature sets to predict the disease phenotype. Figures 1(b) and (c) illustrate the frameworks for the first update (k = 1) of the mRNA and miRNA datasets respectively. Each box in Figure 1(a) represents k^{th} update which contains two Wasserstein GANs (wGANs) for two omics data. After the wGANs are successfully trained, each generator generates a synthetic data which will be alike the input omics dataset and considered as the updated omics data from that box. For each update, an intermediate value for miRNA expression is first generated from the generator using mRNA expression and normalized adjacency matrix representing the interaction network. An intermediate value for the mRNA is also found in a similar procedure:

$$\boldsymbol{h}_x^{(k)} = G(\boldsymbol{H}_y^{(k-1)}, \tilde{\boldsymbol{S}}^{\boldsymbol{T}}) \quad \textbf{(1)}$$

$$\boldsymbol{h}_y^{(k)} = G(\boldsymbol{H}_x^{(k-1)}, \tilde{\boldsymbol{S}}). \quad \textbf{(2)}$$

This mRNA (or miRNA) intermediate value $\boldsymbol{h}_x^{(k)}$ contains information from miRNA (mRNA) in the last update $\boldsymbol{H}_y^{(k-1)}$ and interaction network $\tilde{\boldsymbol{S}}$ but has no relation with the mRNA (miRNA) expression value $\boldsymbol{H}_x^{(k-1)}$ in the last update. The intermediate mRNA (or miRNA) expression value $\boldsymbol{h}_x^{(k)}$ along with the input mRNA (miRNA) expression value $\boldsymbol{H}_x^{(k-1)}$ are then passed through a critic to ensure the similarity between them:

Table 1. Notations for omicsGAN.

Name	Definition
$\boldsymbol{X} \in \mathbb{R}^{m \times n}$	mRNA expression obtained from RNA-seq
$\boldsymbol{Y} \in \mathbb{R}^{p \times n}$	miRNA expression obtained from miRNA-seq
$\boldsymbol{h}_x^{(k)} \in \mathbb{R}^{m \times n}$	intermediate value of mRNA expression in the kth update
$\boldsymbol{h}_y^{(k)} \in \mathbb{R}^{p \times n}$	intermediate value of miRNA expression in the kth update
$\boldsymbol{H}_x^{(k)} \in \mathbb{R}^{m \times n}$	mRNA expression (synthetic) in the kth update
$\boldsymbol{H}_y^{(k)} \in \mathbb{R}^{p \times n}$	miRNA expression (synthetic) in the kth update
$\boldsymbol{Z}_x \in \mathbb{R}^{m \times n}$	final mRNA expression (synthetic), $\boldsymbol{Z}_x = \boldsymbol{H}_x^{(k^*)}$
$\boldsymbol{Z}_y \in \mathbb{R}^{p \times n}$	final miRNA expression (synthetic), $\boldsymbol{Z}_y = \boldsymbol{H}_y^{(k^*)}$
$\boldsymbol{N} \in \{-1, 1\}^{p \times m}$	adjacency matrix of miRNA-mRNA interaction network
$\boldsymbol{D}_X \in \mathbb{R}^{m \times m}$	diagonal matrix: $\boldsymbol{D}_X(i,i) = \sum_j \lvert\boldsymbol{N}(j,i)\rvert$
$\boldsymbol{D}_Y \in \mathbb{R}^{p \times p}$	diagonal matrix: $\boldsymbol{D}_Y(i,i) = \sum_j \lvert\boldsymbol{N}(i,j)\rvert$
$\tilde{\boldsymbol{S}} \in \mathbb{R}^{p \times m}$	normalized adjacency matrix, $\tilde{\boldsymbol{S}} = \boldsymbol{D}_Y^{-\frac{1}{2}} \boldsymbol{N} \boldsymbol{D}_X^{-\frac{1}{2}}$

$$loss_x = D_{loss}(\boldsymbol{h}_x^{(k)}, \boldsymbol{H}_x^{(k-1)}) \quad (3)$$

$$loss_y = D_{loss}(\boldsymbol{h}_y^{(k)}, \boldsymbol{H}_y^{(k-1)}) \quad (4)$$

D_{loss} is the critic loss between the intermediate value and the input value. After training by minimizing the critic loss, the updated mRNA and miRNA dataset $\boldsymbol{H}_x^{(k)}$ and $\boldsymbol{H}_y^{(k)}$ are learned respectively. This step force the distribution of $\boldsymbol{H}_x^{(k)}$ (or $\boldsymbol{H}_y^{(k)}$) towards the distribution of $\boldsymbol{H}_x^{(k-1)}$ ($\boldsymbol{H}_y^{(k-1)}$). The boxes are arranged in a cascaded structure where each box is trained separately. Once we have trained and got updates $\boldsymbol{H}_x^{(k)}$ and $\boldsymbol{H}_y^{(k)}$ from box k, it is used as input in the following $(k+1)^{th}$ box. $\boldsymbol{H}_x^{(0)} = \boldsymbol{X}$ and $\boldsymbol{H}_y^{(0)} = \boldsymbol{Y}$ are the input to the first layer (box) and after the Kth update. $\boldsymbol{Z}_x = \boldsymbol{H}_x^{(k^*)}$ and $\boldsymbol{Z}_y = \boldsymbol{H}_y^{(k^*)}$ are our final synthetic datasets which are used for the disease phenotype prediction, where k^* is the update that gives best prediction result on a separated validation set of samples.

Generative adversarial network model. Generative adversarial network (GAN) models are a class of unsupervised learning task that automatically discovers and learns patterns and distribution in input data in a way that the models can be used to generate new examples that plausibly could have been drawn from the original dataset. It has been widely used in image generation technologies (30, 31). With some appropriately placed conditions, it can also be used in computational biology to synthesize omics data. In general, GANs use random noise to generate synthetic dataset by requiring the distribution of the random noise towards the distribution of the original data. It does not have to retain information from the random noise; rather, try to make the noise as close to the original data as possible in terms of distribution. In multi-omics study, we can introduce a stream of information from one omics data in place of random noise and incentivize the GAN to retain information from this stream as well as forcing the distribution towards a second omics data. This will ensure the integration of information from both omics data in the generated samples. We can also fuse the interaction network in the model through the generator following the works of (32).

Our proposed pipeline has two separate wGANs for two omics data to update them into a new representation. Generators in each wGAN are four layers fully connected neural network that generates a dataset based on one omics data and the normalized adjacency matrix following the equations:

$$\boldsymbol{h}_x^{(k)} = (ReLU(ReLU(\tilde{\boldsymbol{S}}^{\boldsymbol{T}} \boldsymbol{H}_y^{(k-1)} \boldsymbol{W}^{(0)}))\boldsymbol{W}^{(1)})\boldsymbol{W}^{(2)} \quad (5)$$

$$\boldsymbol{h}_y^{(k)} = (ReLU(ReLU(\tilde{\boldsymbol{S}} \boldsymbol{H}_x^{(k-1)} \boldsymbol{W}^{(0)}))\boldsymbol{W}^{(1)})\boldsymbol{W}^{(2)} \quad (6)$$

where $\boldsymbol{W^l}$ is the weight matrix in l^{th} layer and rectified linear unit (ReLU) is the activation function. A fully connected neural network is then trained as a critic to assign values to the obtained intermediate representation $\boldsymbol{h}_x^{(k)}$ and input dataset $\boldsymbol{H}_x^{(k-1)}$. The critic is trained five times for one training of the generator. Objective function for training the critic is:

$$\mathcal{L}_C = C(\boldsymbol{h}_x^{(k)}) - C(\boldsymbol{H}_x^{(k-1)}) \quad (7)$$

where C stands for the critic. Critic assigns larger values to the real samples (i.e., $\boldsymbol{H}_x^{(k-1)}$) and smaller values to the synthetic ones (i.e., $\boldsymbol{h}_x^{(k)}$), thus trained by minimizing equation (7). On the other hand, generator tries to produce synthetic data that will fool the critic into thinking it as real. Objective function for training the generator is:

$$\mathcal{L}_G = -C(\boldsymbol{h}_x^{(k)}) + \alpha \|\boldsymbol{h}_x^{(k)} - \boldsymbol{X}\|_2 \quad (8)$$

where α is a coefficient to control the weight put on the two terms of the equation. For a successful training, generator has to produce data $\boldsymbol{h}_x^{(k)}$ realistic enough that will be assigned a larger value by the critic; therefore, it is trained by minimizing equation (8). A L_2-norm is added to further steer the updated dataset towards the original mRNA expression and preserve the feature characteristics. $\boldsymbol{h}_y^{(k)}$ and $\boldsymbol{H}_y^{(k)}$ for miRNA update is derived using analogous equations.

Evaluation methods. Classification model: We designed cancer outcome classification tasks with the assumption that better quality of the synthetic datasets will lead to better signatures for disease phenotype prediction compared to mRNA and miRNA expression. A feed forward fully connected deep neural network is implemented as a classifier with a binary cross-entropy loss function,

$$\mathcal{L} = -\boldsymbol{h} log(\boldsymbol{p}) - (1 - \boldsymbol{h}) log(1 - \boldsymbol{p}) \quad (9)$$

where $\boldsymbol{h}$ is the truth label of the cancer patients and $\boldsymbol{p}$ is the predicted label. Adam optimizer is used with learning rate 0.01 and a ratio of 60%, 20%, 20% as numbers of training, validation, and test samples respectively. This is a two hidden layer neural network with 250 and 100 neurons in each layer respectively. Both hidden layers use ReLU as activation function and dropout with probability 0.2. The output layer uses Sigmoid as activation function. Area Under the Receiver Operating Characteristic Curve (AUC) is applied to evaluate the performance of the classifier and the quality of the input omics features.

Survival prediction model: A Cox proportional hazards model with Elastic Net penalty (33) is applied to study the correlation between patients overall survival and omics profiles. The Elastic Net penalty uses a weighted combination of the L_1-norm and L_2-norm penalties by maximizing the following log-likelihood function,

$$\log L(\boldsymbol{\beta}) - \alpha(r \sum_{i=1}^{m} |\beta_i| + \frac{1-r}{2} \sum_{i=1}^{m} \beta_i^2) \quad (10)$$

where $L(\boldsymbol{\beta})$ is the partial likelihood of the Cox model, $\alpha \geq 0$ is a hyper-parameter that controls the amount of shrinkage, $r \in [0, 1]$ is the relative weight of the L_1-norm and L_2-norm

penalties, and $\beta_i (i \in [1, m])$ represents the coefficient for the ith genomic feature in the omics data. The omics data is randomly splitted into training (80%) and test (20%) sets. Five-fold cross validation is performed on training data to tune the hyper-parameter α. The high risk group and low risk group are determined by the prognostic index ($\boldsymbol{PI}$) on the independent test set. The $\boldsymbol{PI}$ is the linear component of the Cox model, $\boldsymbol{PI} = \beta^T \boldsymbol{X}_{test}$, where $\boldsymbol{X}_{test}$ is the omics profile of the test set, and its risk coefficient was estimated from the Cox model fitted on the training set. The high risk and low risk groups are generated for Kaplan-Meier survival plot by splitting the ordered $\boldsymbol{PI}$ with equal number of samples in each group in the test set. Python package *scikit survival* (34) is applied to implement Cox proportional hazards model with elastic net, and *lifelines* (35) is used for Kaplan-Meier plotting.

Experiments

In the experiments, we performed three experiments on The Cancer Genome Atlas (TCGA) datasets to evaluate the performance of omicsGAN and compared it to the original mRNA and miRNA expression datasets. The first experiment was to assess the cancer outcome prediction power of the synthetic datasets. The assessment was conducted in two ways, the first of which was classifying clinical variables of breast cancer and ovarian cancer patients. Secondly, we compared the number of significant features identified in each dataset. The second experiment was to show the impact of an accurate interaction network on the prediction power of the synthetic datasets. The last experiment was to evaluate the quality of the synthetic mRNA data on cancer patients overall survival prediction.

Dataset. The proposed framework, omicsGAN, was tested on The Cancer Genome Atlas (TCGA) breast cancer (BRCA) and ovarian cancer (OV) datasets (36, 37). The RNA-seq mRNA expression and miRNA expression datasets were downloaded from UCSC Xena Hub (38). For the mRNA expression, the $log2(x+1)$ transformed RSEM normalized count with 20,531 genes was used and for the miRNA expression, the $log2(x+1)$ transformed RPM value with 2,166 miRNAs was used in this study. The clinical information of the two cancer studies was downloaded from cBioPortal (39). There are 185 Estrogen Receptor positive (ER+) and 54 ER negative (ER-) samples, 46 triple negative positive (TN+) and 193 TN negative (TN-) samples in the breast cancer dataset as well as 51 cancer patients in the early stage ($\leq$ IIIA) and 359 in the late stage ($>$ IIIA) in the ovarian cancer dataset. The miRNA-mRNA interaction network was obtained from TargetScanHuman (40) which predict effective miRNA target sites in mRNAs. A modified adjacency matrix represented the interaction network, where each interaction was valued as -1 to imitate that miRNA negatively regulates the expression of the targeted mRNA and no interaction was valued as 1. The miRNA-mRNA bipartite network contained 163,568 interactions in total.

Table 2. Hyperparameters in omicsGAN used in the study.

Hyperparameter	Breast cancer	Ovarian cancer
mRNA generator learning rate	5e-5	1e-4
mRNA critic learning rate	5e-5	1e-7
mRNA L_2-norm coefficient (α)	0.0001	1
miRNA generator learning rate	5e-3	1e-4
miRNA critic learning rate	1e-4	1e-7
miRNA L_2-norm coefficient (α)	0.001	0.001

Running omicsGAN on the TCGA datasets. To evaluate the proposed generative model on the TCGA cancer omics datasets, we first updated the mRNA and miRNA expression profiles 10 times ($K = 10$). The generator and critic are fully connected neural networks with two hidden layers. The generator hidden layers have 512 and 768 neurons whereas the critic hidden layers have 256 and 128 neurons respectively. In both generator and critic, the activation function of the hidden layers is LeakyReLU and the output layer is linear. Moreover, hidden layers in critic have dropout with probability 0.2. RMSprop optimizer was applied to train both generator and critic. Hyperparameters were selected through grid search and details of the hyperparameters used in this study are listed in Table 2. In particular, the learning rate was chosen from {1e-8, 1e-7, 5e-7, 1e-6, 5e-6, 1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2} and the candidates of the coefficient α were {1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1, 10}. For batch size, we selected among the options {16, 32, 64, 128, 256}, and no mini batch. We used no mini batch as there was no significant improvement with its use. The classifier and validation set described in the Method section were employed for tuning all parameters. All updated mRNA and miRNA datasets ($k = 1, 2, .., 10$) are sequentially fed into the classifier. A fully connected deep neural network described in the Method Section was applied as the classifier for all datasets. In the classifier, the dataset was divided into five folds with three folds for training, one fold for validation (parameter tuning), and one fold for test. We repeated the five-fold splitting 50 times on each dataset. The update (k^*) with the highest AUC score for mRNA validation samples was selected as the final output from the model.

Input data	Breast cancer		Ovarian cancer
	ER	TN	cancer stage
mRNA	0.924	0.954	0.74
synthetic mRNA (omicsGAN)	**0.958***	**0.961**	**0.751**
miRNA	0.926	0.923	0.676
synthetic miRNA (omicsGAN)	0.913	0.905	0.721*

Table 3. The classification performance on TCGA breast cancer and ovarian cancer datasets. Average AUC scores of classification performance between synthetic mRNA, miRNA generated from omicsGAN and the original mRNA, miRNA expression on breast cancer and ovarian cancer datasets. The best results across all the input data are bolded. *The difference between the results on the original expression data and the synthetic data is statistically significant (p-value < 0.001).

omicsGAN improved cancer outcome prediction. To evaluate the quality of the synthetic datasets generated by omicsGAN, we designed cancer outcome prediction and signature identification tasks on the TCGA breast cancer and ovarian cancer datasets under the assumptions: (1) The synthetic datasets learned in omicsGAN considers the expressions in both mRNA and miRNA profiles and the biological interactions between them. So they will provide more signatures compared to mRNA and miRNA expressions. (2) The better molecular signatures will improve the disease phenotype prediction. In the experiment, the complete miRNA-mRNA interaction network from TargeScanHuman was applied to generate the synthetic datasets in omicsGAN. For phenotype prediction on breast cancer patients, the synthetic mRNA from the 4th update had the highest validation AUC score, thus the synthetic mRNA and miRNA after the 4th update (i.e., $k^* = 4$) were considered as the output from omicsGAN for the downstream analysis. Similarly, for phenotype prediction on ovarian cancer patients, the synthetic mRNA from the 3rd update was applied in the data analysis based on the highest validation AUC.

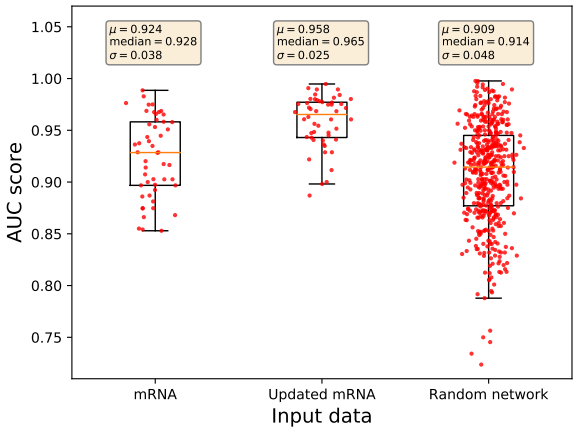

Fig. 2. Prediction results of the ER status on TCGA breast cancer patients. Each dot represents the AUC score from one splitting. Statistics (mean, median,and standard deviation) of the prediction performance of the 50 splittings are shown above each boxplot.

We ran the classifier with above mentioned five-fold splitting 50 times for predicting the ER status and triple negative (TN) status of the breast cancer patients and cancer stage of the ovarian cancer patients with the original mRNA, miRNA expression datasets as well as the synthetic mRNA, miRNA datasets generated by omicsGAN. The results for the prediction are tabulated in Table 3. We report the average AUC of the 50 repeats. The synthetic mRNA and miRNA expression generated by omicsGAN achieved better average classification results than the ones using original mRNA and miRNA expression for ovarian cancer stage prediction. Synthetic mRNA expression also performs better in breast cancer outcome prediction on both ER and TN status; however, synthetic miRNA expression did not perform better than the original miRNA expression. To update the miRNA expression, we use the mRNA expression and adjacency matrix as described in equation (6) creating a flow of information from

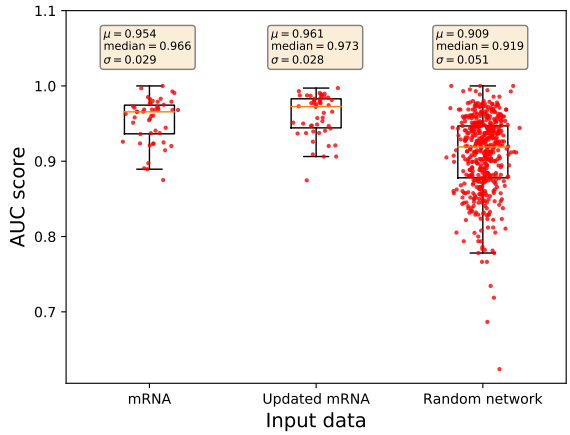

Fig. 3. Prediction results of the TN status on TCGA breast cancer patients. Each dot represents the AUC score from one splitting. Statistics (mean, median,and standard deviation) of the prediction performance of the 50 splittings are shown above each boxplot.

mRNA to miRNA. From a biological point of view, this flow happens from miRNA to mRNA as miRNA binding to the 3' untranslated region of mRNA regulates its translation. On the contrary, mRNA expression has no direct interaction with the miRNA that can impact its properties. This reverse flow of information causes the model for synthetic miRNA generation to be unstable resulting in better performance in ovarian cancer but worse in breast cancer phenotype prediction. Synthetic mRNA expression generated following the equation (5) on the other hand mimics the actual flow of information in the post-transcriptional regulation (i.e., miRNA regulation), thus, shows better performance in both breast cancer and ovarian cancer phenotype prediction. The prediction results for original and synthetic mRNAs are also shown as boxplot in Figures 2, 3, and 4. Each dot on the boxplot represents the result from one random splitting. Statistics (mean, median, and standard deviation) of the prediction performance of the splittings are shown above each boxplot. The third boxplot for random network in Figures 2 and 3 will be described later.

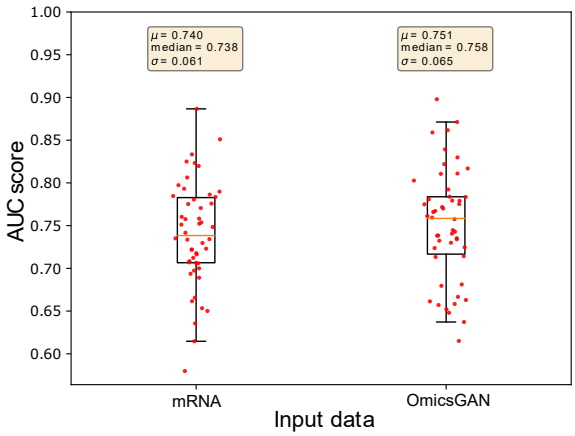

Fig. 4. Prediction results of the cancer stage on TCGA ovarian cancer patients. Each dot represents the AUC score from one splitting. Statistics (mean, median,and standard deviation) of the prediction performance of the 50 splittings are shown above each boxplot.

We also evaluated the quality of the original and synthetic

datasets by comparing the number of significant features identified in each of them. We performed t-test on the expression datasets with different clinical variables. The number of features in each dataset with adjusted p-value smaller than 0.01 is presented in Table 4. We can see an increased number of significant features in synthetic mRNA compared to the original one for both breast cancer and ovarian cancer datasets on all clinical variables. Synthetic miRNA on the other hand has more significant features for breast cancer and less for ovarian cancer compared to the original miRNA expression datasets. This can also be attributed to the unstable nature of the miRNA update model in omicsGAN due to the reverse flow of information from the miRNA-mRNA interaction network.

Input data	Breast cancer		Ovarian cancer
	ER	TN	cancer stage
mRNA	113	102	21
Synthetic mRNA (omicsGAN)	159	155	33
miRNA	117	102	51
Synthetic miRNA (omicsGAN)	146	140	48

Table 4. Number of significant features. Number of significant features between synthetic mRNA, miRNA generated by omicsGAN and the original mRNA, miRNA expression on breast cancer and ovarian cancer datasets.

Impact of interaction network on cancer outcome prediction. miRNA expression provides additional predictive signals for cancer outcome prediction on top of the mRNA expression; therefore, integrating them into a new feature set will contain more information compared to mRNA and miRNA expression individually. However, we have seen a difference in the performance of synthetic mRNA and miRNA expression in phenotype prediction as tabulated in Table 3 and 4. Synthetic mRNA expression consistently does better in all evaluations whereas synthetic miRNA fails to do so even though both of them contains integrated information from original mRNA and miRNA. The difference arises from the interaction network and setting a flow of information from one omics data to another. Using the bipartite interaction network is an integral part of our framework that introduces the information about biological interaction between two omics data. So, we design an experiment to explore the effects of the interaction network on synthetic omics data and their predictive performance. We ran the framework 10 times with same parameters and input $\boldsymbol{X}$ (mRNA expression) and $\boldsymbol{Y}$ (miRNA expression) as before except the interaction network on TCGA breast cancer datasets. We replaced the true network with 10 different randomized network with same density as the true one. We consider the output from update 4 as the synthetic data to make it comparable to the previous results using true network. Classification results for ER and TN status for breast cancer using original mRNA, miRNA expression datasets as well as synthetic mRNA, miRNA datasets are tabulated in Table 5. We report the average AUC score of the 10 randomized network results each with 50 repeats. The prediction results for synthetic mRNA expression using random networks are also shown as boxplot in Figures 2 and 3. Each dot on the boxplot represents the result from one random splitting. Statistics (mean, median, and standard deviation) of the prediction performance of the splittings are shown above each boxplot.

Input data	ER		TN	
	true	random	true	random
Synthetic mRNA (omicsGAN)	0.958	0.909	0.961	0.910
Synthetic miRNA (omicsGAN)	0.913	0.918	0.905	0.912

Table 5. The classification performance on TCGA breast cancer dataset using random interaction network. Average AUC scores of classification performance between synthetic mRNA, miRNA generated from omicsGAN using original (true) and random interaction network on breast cancer study.

In Table 5, we see a reduction in performance of synthetic mRNA expression generated using a random interaction network compared to the one generated using the true interaction network. This signifies the importance of the interaction network in phenotype prediction and the capability of our framework to capture the information within the network. However, the performance of synthetic miRNA expression does not differ much with the use of a random network. Synthetic miRNA can not integrate meaningful network information as efficiently as synthetic mRNA even from the true network as mentioned in section .

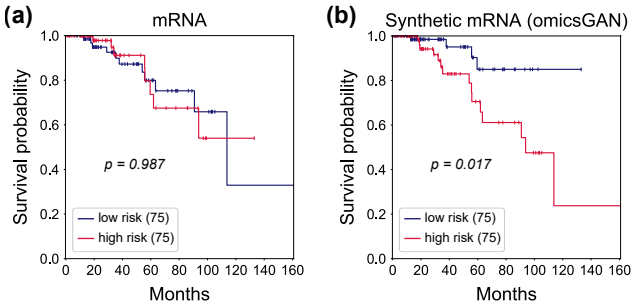

Fig. 5. Survival prediction on breast cancer patients. Kaplan-Meier survival plots for high (red line) and low (blue line) risk groups generated by (a) original mRNA, (b) synthetic mRNA expression data on breast cancer patients. The number in the parenthesis indicates the number of samples in low or high risk group. The p-value is calculated by the log-rank test to compare the overall survival of two groups of cancer patients.

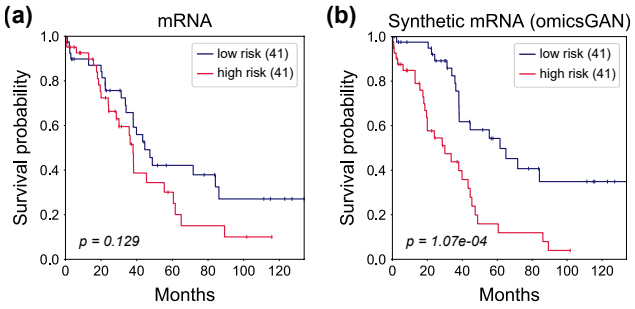

Fig. 6. Survival prediction on ovarian cancer patients. Kaplan-Meier survival plots for high (red line) and low (blue line) risk groups generated by (a) original mRNA, (b) synthetic mRNA expression data on ovarian cancer patients. The number in the parenthesis indicates the number of samples in low or high risk group. The p-value is calculated by the log-rank test to compare the overall survival of two groups of cancer patients.

omicsGAN improved survival prediction. To further investigate the quality of the synthetic mRNA expression data produced by omicsGAN, the patients overall survival was predicted on the breast cancer and ovarian cancer datasets. The Cox proportional hazards model with elastic net penalty as described in section to evaluate the correlation between patients overall survival and genomic features, i.e., the original mRNA expressions and the synthetic mRNA expressions in this study. The relative weight r in equation 10 was set to be 0.5 to combine the subset selection property of the L_1-norm with the regularization strength of the L_2-norm. 80% of the patient samples were applied to train the model and the performance was tested on 20% test samples. The low and high risk groups on the independent test set were generated based on the prognostic index ($\boldsymbol{PI}$) as mentioned in section . The survival predictions were visualized by Kaplan-Meier plots and compared by the log-rank test p-values. The Kaplan-Meier plots in Figure 5 and 6 exemplify the improved patient survival predictions on breast and ovarian cancer using the synthetic mRNA expression generated by omicsGAN compared to the original mRNA expression. The log-rank test p-values clearly demonstrate a strong additional prognostic power of synthetic omics profiles beyond the original signatures.

Discussion

Disease phenotype prediction plays a key role in the fight against heterogeneous diseases like cancer. Multi-omics data powered by next generation sequencing technologies has transformed the field of phenotype prediction by providing a broader view of the molecular profiles. Non-redundant predictive signals from multi-omics datasets make it crucial to develop an efficient and effective framework for multi-omics data integration. However, integrating them as an independent set of features is not enough as multi-omics data generated for the same set of samples often have an interactive relation among them. Incorporating the interaction network into the analysis will set a flow of information from one omics data to another like the flow within different omics layers in a cell. However, in most studies, these intra-omics and inter-omics relations are neglected and it is inefficient to predict phenotype using only the omics data. Therefore, integrating the bipartite interaction network with multi-omics data can improve disease phenotype prediction.

Synthetic data generated from our proposed framework, omicsGAN, shows improvement in prediction performance which illustrates the capability of the model to successfully retain information from multiple omics data and establish a link between them. However, there is a performance difference between synthetic mRNA and miRNA datasets. Synthetic mRNA performs better in all evaluation metrics whereas synthetic miRNA is comparatively unstable, with better performance in some cases and worse in others. The main source of the difference is the miRNA-mRNA interaction network. This network is not an independent feature set with a significant predictive ability that can be fused directly to the omics datasets, it rather has a biological interpretation. miRNA-mRNA interaction network identifies the target mRNA for each miRNA that it will bind and regulate during translation. Therefore, to successfully integrate the information from this network, we need to create this regulatory effect representing its biological functionality in our model. Updating mRNA expression using miRNA and adjacency matrix in equation (5) resembles the effect of miRNA regulation on mRNA translation. This enables an efficient propagation of information from miRNA to mRNA expression dataset. mRNA on the other hand, having no such effect on miRNA properties, can not propagate biologically meaningful information during miRNA update in equation (6) as efficiently as equation (5). It also signifies that omicsGAN does not fuse information from the two omics data directly; rather functionally incorporate the interaction network into the integration. A L_2-norm is added in equation (8) to ensure the similarity between the updated and original omics data expression; thus allowing the synthetic data to retain feature space properties of the original omics data. It also helps to control the instability in the generation of synthetic miRNA and conserve useful information from both omics data in it even if its prediction performance remains similar.

The framework presents an innovative way for multi-omics data integration incorporating their biological interaction. However, the framework for synthetic miRNA generation is not as efficient as synthetic mRNA generation, and improvement for some phenotype prediction is limited. A larger comprehensive study involving more cancer types can draw a better picture of the improvements in phenotype prediction. Although our study was focused on miRNA-mRNA interaction, the same technique can be extrapolated to any two omics data. However, to integrate two omics data with different range, distribution, and format (e.g., mutation and gene expression), an extra pre-processing step is necessary to make them compatible. There are existing studies with the preprocessing framework that works with such multi-omics data (41).

Conclusion

Thanks to the rapid evolution of high-throughput technologies, abundant genotype data is accruing, which is expected to grow continuously in the era of precision medicine. Because of the complex interactive nature of omics layers, integration of multi-omics data to extract biologically meaningful information of clinical relevance is a challenging task. The promise of multi-omics analysis will remain unfulfilled unless we can functionally incorporate the inter-omics interaction network into the analysis. In this study, we introduced omicsGAN, a generative adversarial network model to effectively integrate the interaction network and the omics datasets into new synthetic data with better predictive signals. We observed that the synthetic data generated from omicsGAN has better discriminative power on cancer outcome classification and cancer patients survival prediction compared to the original omics datasets. Synthetic datasets also contain more significant features that result in better predictive performance. Additionally, we analyzed the effect of interaction

network on the quality of synthetic data. Our results show that omicsGAN does not only gather information from two omics datasets; rather functionally incorporate their biological interaction into the integration. Using a random interaction network does not create a flow of information from one omics data to another as efficiently as the true network.

Acknowledgements

The results are based upon data generated by The Cancer Genome Atlas established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov. The dbGaP accession number to the specific version of the TCGA dataset is phs000178.v8.p7.

Funding

This work has been supported by grants from the National Science Foundation (NSF) [NSF-III1755761]; National Institutes of Health (NIH) [2R01GM113952-06].

Bibliography

1. Paulina Krzyszczyk, Alison Acevedo, Erika J Davidoff, Lauren M Timmins, Ileana Marrero-Berrios, Misaal Patel, Corina White, Christopher Lowe, Joseph J Sherba, Clara Hartmanshenn, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology*, 6(03n04):79–100, 2018.
2. Hugo Fitipaldi, Mark I McCarthy, Jose C Florez, and Paul W Franks. A global overview of precision medicine in type 2 diabetes. *Diabetes*, 67(10):1911–1922, 2018.
3. Chao Wang, Raghu Machiraju, and Kun Huang. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods*, 67(3):304–312, 2014.
4. Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:11779322198990 51, 2020.
5. Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.
6. Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.
7. Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C Chambers, Lisa J Zimmerman, Kent F Shaddox, Sangtae Kim, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, 2014.
8. John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.
9. Emilie EL Muller, Nicolás Pinel, Cédric C Laczny, Michael R Hoopmann, Shaman Narayanasamy, Laura A Lebrun, Hugo Roume, Jake Lin, Patrick May, Nathan D Hicks, et al. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature communications*, 5(1):1–10, 2014.
10. Marco Fondi and Pietro Liò. Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiological research*, 171:52–64, 2015.
11. Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.
12. Biswapriya B Misra, Carl Langefeld, Michael Olivier, and Laura A Cox. Integrated omics: tools, advances and future approaches. *Journal of molecular endocrinology*, 62(1):R21–R45, 2019.
13. Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.
14. Julia K Nussbacher and Gene W Yeo. Systematic discovery of rna binding proteins that regulate microrna levels. *Molecular cell*, 69(6):1005–1016, 2018.
15. Matthias W Hentze, Alfredo Castello, Thomas Schwarzl, and Thomas Preiss. A brave new world of rna-binding proteins. *Nature Reviews Molecular Cell Biology*, 19(5):327, 2018.
16. Hiromi WL Koh, Damian Fermin, Christine Vogel, Kwok Pui Choi, Rob M Ewing, and Hyungwon Choi. iomicspass: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ systems biology and applications*, 5(1):1–10, 2019.
17. Khandakar Tanvir Ahmed, Sunho Park, Qibing Jiang, Yunku Yeu, TaeHyun Hwang, and Wei Zhang. Network-based drug sensitivity prediction. *BMC medical genomics*, 13(11):1–10, 2020.
18. Nimrod Rappoport and Ron Shamir. Nemo: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356, 2019.
19. Hung Nguyen, Sangam Shrestha, Sorin Draghici, and Tin Nguyen. Pinsplus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, 35(16):2843–2846, 2019.
20. Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.
21. Tao Zhou, Kim-Han Thung, Xiaofeng Zhu, and Dinggang Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human brain mapping*, 40(3):1001–1016, 2019.
22. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.
23. Minseon Kim, Ilhwan Oh, and Jaegyoon Ahn. An improved method for prediction of cancer prognosis by network learning. *Genes*, 9(10):478, 2018.
24. Arsham Ghahramani, Fiona M Watt, and Nicholas M Luscombe. Generative adversarial networks simulate gene expression and predict perturbations in single cells. *BioRxiv*, page 262501, 2018.
25. Jinhee Park, Hyerin Kim, Jaekwang Kim, and Mookyung Cheon. A practical application of generative adversarial networks for rna-seq analysis to predict the molecular progress of alzheimer's disease. *PLoS computational biology*, 16(7):e1008099, 2020.
26. Mojtaba Bahrami, Malosree Maitra, Corina Nagy, Gustavo Turecki, Hamid Rabiee, and Yue Li. Deep feature extraction of single-cell transcriptomes by generative adversarial network. *bioRxiv*, 2020.
27. Chaoyang He, Tian Xie, Yu Rong, Wenbing Huang, Yanfang Li, Junzhou Huang, Xiang Ren, and Cyrus Shahabi. Bipartite graph neural networks for efficient node representation learning. *arXiv preprint arXiv:1906.11994*, 2019.
28. Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
29. Yungang Xu, Zhigang Zhang, Lei You, Jiajia Liu, Zhiwei Fan, and Xiaobo Zhou. scigans: single-cell rna-seq imputation using generative adversarial networks. *Nucleic acids research*, 48(15):e85–e85, 2020.
30. Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.
31. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.
32. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
33. Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39 (5):1, 2011.
34. Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.
35. Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Sean Reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel WIlson, Gabriel, Luis Moneda, Arturo Moncada-Torres, Kyle Stark, Harsh Gadgil, Jona, Karthikeyan Singaravelan, Lilian Besson, Miguel Sancho Peña, Steven Anton, Andreas Klintberg, GrowthJeff, Javad Noorbakhsh, Matthew Begun, Ravin Kumar, Sean Hussey, Skipper Seabold, and Dave Golland. Camdavidsonpilon/lifelines: v0.25.8, January 2021.
36. Cancer Genome Atlas Network TCGA et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.
37. Cancer Genome Atlas Research Network TCGA et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609, 2011.
38. Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature Biotechnology*, pages 1–4, 2020.
39. Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269):pl1–pl1, 2013.
40. Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015.
41. Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. Moli: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, 2019.