Bioinformatics, 38(1), 2022, 179–186
doi: 10.1093/bioinformatics/btab608
Advance Access Publication Date: 20 August 2021
Original Paper

Systems biology

Multi-omics data integration by generative adversarial network

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

[1]Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA, [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

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on June 8, 2021; revised on July 27, 2021; editorial decision on August 13, 2021; accepted on August 18, 2021

Abstract

Motivation: 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 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.

Results: Large-scale experiments on The Cancer Genome Atlas breast cancer, lung cancer and ovarian cancer datasets validate that (i) the model can effectively integrate two omics data (e.g. mRNA and microRNA expression data) and their interaction network (e.g. 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. (ii) 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.

Availability and implementation: Source code is available at: https://github.com/CompbioLabUCF/omicsGAN.

Contact: wzhang.cs@ucf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Complex diseases such as cancer are highly heterogeneous with different subtypes leading to varying clinical outcomes including prognosis, response to treatment and chances of recurrence and metastasis (Ahmed *et al.*, 2020; Krzyszczyk *et al.*, 2018; Wang *et al.*, 2014b). Disease phenotype prediction has been the subject of interest to clinicians and patients for many decades. The recent developments in high throughput sequencing technologies are capable of measuring molecular activities in cells and allow researchers to obtain multi-omics data with sufficient quality and yield (Goodwin *et al.*, 2016). It has revolutionized medical and biological research by offering a more comprehensive view of the underlying biological process of disease and identify accurate molecular signatures for characterizing or predicting disease phenotypes. Analysis of multi-omics data along with clinical information of patients can help bridging the gap between genotype and phenotype by exploring the flow of information within different omics layers (Subramanian *et al.*, 2020). These omics layers provide non-redundant predictive signals for predicting therapeutic response. Removing one of them in a prediction system will lead to performance degradation (Wang *et al.*, 2014a). Therefore, multi-omics data may provide a complementary set of information to understand the molecular basis of diseases. Predicting the phenotype using the multi-omics data as independent sets of features will fall short in characterizing the prediction.

Disease phenotypes depend on molecular profiles and interplay at genomic, epigenomic, transcriptomic, proteomic and metabolomic levels (Subramanian *et al.*, 2020) which are interconnected with each other through complex networks (Ahmed *et al.*, 2021). 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 (Yeh *et al.*, 2017). 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. Moreover, mRNA can be regulated by other modulators like RNA binding protein (RBP) (Nussbacher and Yeo, 2018). RBPs bind RNA through globular RNA-binding domains (RBDs) and alter the expression of the bound RNAs (Hentze *et al.*, 2018). RNA–RBP interaction obtained from crosslinking and immunoprecipitation-based CLIP-Seq can also be applied to characterize the relation between omics data. Hence, integrating the interaction network into multi-omics data analysis will capture the regulatory effect and establish a better correlation with the phenotype.

Several advanced multi-omics data integration frameworks have been proposed in the last five years (Argelaguet *et al.*, 2018; Nguyen *et al.*, 2019; Rappoport and Shamir, 2019; Zhou *et al.*, 2019). However, few approaches link different omics profiles using molecular interaction (Koh *et al.*, 2019). Most of them ignore the relations across different biological layers in their analysis. The power of high throughput technologies cannot be fully utilized unless the multi-omics data with its intermodal relations are considered in studies.

In recent years, generative adversarial networks (GAN) (Goodfellow *et al.*, 2014) has gained popularity in solving problems within the scope of computational biology. GANs take random noise or predefined data as input and generate plausible synthetic data similar to a real dataset by imitating the distribution of the real data. There are several studies that use GAN-based algorithms to generate data from single or multiple omics datasets. Kim *et al.* (2018) used GAN for better biomarkers identification by generating a reconstructed functional interaction network from multi-omics datasets. Ghahramani *et al.* (2018) integrated diverse single-cell RNAseq (scRNA-seq) datasets from different labs and experimental protocols to simulate realistic scRNA-seq data that covers the full cell type diversity. Park *et al.* (2020) on the other hand used GAN to generate gene expression from bulk RNA-seq datasets. GANs can learn non-linear relationships between features of omics data during training that can be used later for additional insight (Ghahramani *et al.*, 2018). It can handle missing data and also promising for missing value imputation because of its capability of learning and imitating any distribution of data (Xu *et al.*, 2020). Based on its property of imitating distribution, we can design a GAN with one omics data from one distribution as input to the generator and another omics data with different distribution as real dataset in the discriminator to generate a synthetic data retaining information from both omics datasets.

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 single or concatenated omics data.

2 Materials and methods

In this section, we first introduce the mathematical notations employed in this study, followed by the proposed framework, omicsGAN, for generating synthetic omics data for disease outcome prediction using multi-omics data. The framework can take any two omics data with biological relations between each other as input. In this section, we used miRNA, mRNA and miRNA–mRNA interaction network for illustrative purposes. We then discuss the evaluation metrics and introduce two evaluation methods; a classification model and a penalized Cox regression model that use the synthetic data for disease phenotype prediction and patient survival prediction, respectively.

2.1 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 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 GAN 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 section, the framework is introduced 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 interactions are biologically meaningful. mRNA and miRNA expression are correlated to disease phenotype, although, the bipartite interaction network between them can be leveraged to increase 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 $k \in \{1, 2, 3, \ldots, K\}$, 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 disease outcome using two omics data and the interaction network between them as illustrated in Figure 1a. 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 $\tilde{\boldsymbol{S}} = \boldsymbol{D}_{\mathrm{Y}}^{-\frac{1}{2}}\boldsymbol{N}\boldsymbol{D}_{\mathrm{X}}^{-\frac{1}{2}}$. $\boldsymbol{D}_{\mathrm{X}}$ and $\boldsymbol{D}_{\mathrm{Y}}$ are two diagonal matrices with $\boldsymbol{D}_{\mathrm{X}}(i,i) = \sum_j |\boldsymbol{N}(j,i)|$ and $\boldsymbol{D}_{\mathrm{Y}}(i,i) = \sum_j |\boldsymbol{N}(i,j)|$. A classification model is then applied on the new feature sets to predict the disease phenotype. Figure 1b and c illustrate the frameworks for the first update ($k = 1$) of the mRNA and miRNA datasets, respectively. Each box in

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
$\boldsymbol{D}_{\mathrm{X}} \in \mathbb{R}^{m \times m}$	Diagonal matrix: $\boldsymbol{D}_{\mathrm{X}}(i,i) = \sum_j |\boldsymbol{N}(j,i)|$
$\boldsymbol{D}_{\mathrm{Y}} \in \mathbb{R}^{p \times p}$	Diagonal matrix: $\boldsymbol{D}_{\mathrm{Y}}(i,i) = \sum_j |\boldsymbol{N}(i,j)|$
$\tilde{\boldsymbol{S}} \in \mathbb{R}^{p \times m}$	Normalized adjacency matrix, $\tilde{\boldsymbol{S}} = \boldsymbol{D}_{\mathrm{Y}}^{-\frac{1}{2}}\boldsymbol{N}\boldsymbol{D}_{\mathrm{X}}^{-\frac{1}{2}}$

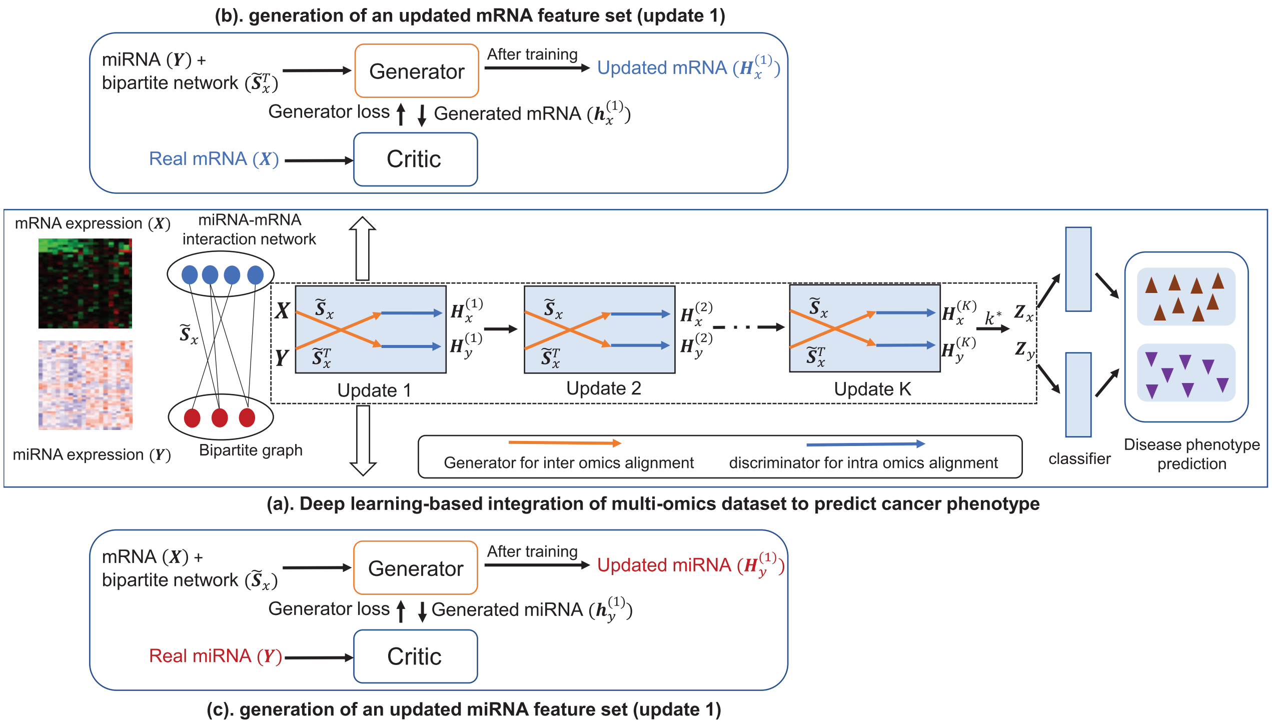
(b). generation of an updated mRNA feature set (update 1)
miRNA (Y) +
bipartite network ($\tilde{S}_x^T$)
Generator
After training
Updated mRNA ($H_x^{(1)}$)
Generator loss
Generated mRNA ($h_x^{(1)}$)
Real mRNA (X)
Critic
mRNA expression (X)
miRNA-mRNA interaction network
$\tilde{S}_x$
miRNA expression (Y)
Bipartite graph
X
Y
$\tilde{S}_x$
$\tilde{S}_x^T$
$H_x^{(1)}$
$H_y^{(1)}$
Update 1
$H_x^{(2)}$
$H_y^{(2)}$
Update 2
$H_x^{(K)}$
$H_y^{(K)}$
Update K
k^*
Z_x
Z_y
classifier
Disease phenotype prediction
Generator for inter omics alignment
discriminator for intra omics alignment
(a). Deep learning-based integration of multi-omics dataset to predict cancer phenotype
mRNA (X) +
bipartite network ($\tilde{S}_x$)
Generator
After training
Updated miRNA ($H_y^{(1)}$)
Generator loss
Generated miRNA ($h_y^{(1)}$)
Real miRNA (Y)
Critic
(c). generation of an updated miRNA feature set (update 1)

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

Figure 1a represents kth update which contains two Wasserstein GANs (wGANs) (Arjovsky *et al.*, 2017) 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:

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

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

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 they are similar to each other:

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

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

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 $\boldsymbol{H}_y^{(k)}$) toward the distribution of $\boldsymbol{H}_x^{(k-1)}$ ($\boldsymbol{H}_y^{(k-1)}$). The boxes (updates) in Figure 1a 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.

2.2 GAN model

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 (Zhang *et al.*, 2017). 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 toward 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 by using appropriate hyperparameters as well as forcing the distribution toward 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 Kipf and Welling (2016).

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

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

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

where $\boldsymbol{W}^l$ is the weight matrix in lth 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:

$$\ell_C = C(\boldsymbol{h}_x^{(k)}) - C(\boldsymbol{H}_x^{(k-1)}), \tag{7}$$

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:

$$\ell_G = -C(\boldsymbol{h}_x^{(k)}) + \alpha||\boldsymbol{h}_x^{(k)} - \boldsymbol{X}||_2, \tag{8}$$

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). An L_2-norm is added to further steer the updated dataset toward 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.

2.3 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 the original omics data. Support vector machine (SVM) with linear kernel is implemented as a classifier for all experiments. The datasets are divided into a ratio of 60%, 20%, 20% as numbers of training, validation and test samples, respectively. This model was implemented via Python package sklearn.svm (SVC) (Pedregosa *et al.*, 2011).

Survival prediction model: A Cox proportional hazards model with Elastic Net penalty (Simon *et al.*, 2011) is applied to study the correlation between patient's 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,

$$\log L(\boldsymbol{\beta}) - \alpha(\mathrm{r}\sum_{\mathrm{i}=1}^{\mathrm{m}}|\beta_\mathrm{i}| + \frac{1-\mathrm{r}}{2}\sum_{\mathrm{i}=1}^{\mathrm{m}}\beta_\mathrm{i}^2), \tag{9}$$

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 (***PI***) on the independent test set. The ***PI*** is the linear component of the Cox model, $\boldsymbol{PI} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{test}}$, where $\boldsymbol{X}_{\mathrm{test}}$ 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 ***PI*** with equal number of samples in each group in the test set. Python package *scikit survival* (Pölsterl, 2020) is applied to implement Cox proportional hazards model with elastic net, and *lifelines* (Davidson-Pilon, 2019) is used for Kaplan–Meier plotting.

3 Experiments

We performed experiments on The Cancer Genome Atlas (TCGA) datasets to evaluate the performance of omicsGAN with two different interaction networks [e.g. miRNA–mRNA interaction network and transcription factor (TF)-gene interaction network]. In this section, we first describe the datasets and two interaction networks used in experiments. Next we introduce the experimental setup where we explain how to run our proposed model on TCGA data and generate synthetic omics datasets. Lastly, we performed three experiments to evaluate the performance of omicsGAN and the quality of its generated synthetic data: (i) comparing cancer outcome prediction power of the real and synthetic datasets. The comparison was conducted in two ways: classifying clinical variables of cancer patients and number of significant features identified in each dataset; (ii) exploring the impact of an accurate interaction network on the prediction power of synthetic datasets; (iii) comparing the cancer patient's overall survival prediction using real and synthetic datasets.

3.1 Dataset and networks

The proposed framework, omicsGAN, was tested on TCGA breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD) and ovarian serous cystadenocarcinoma (OV) datasets (The Cancer Genome Atlas Network *et al.*, 2012, The Cancer Genome Atlas Research Network *et al.*, 2011, 2014). The RNA-seq mRNA expression and miRNA expression datasets of each cancer type were downloaded from UCSC Xena Hub (Goldman *et al.*, 2020). For the mRNA expression, the $\log 2(x+1)$ transformed RSEM normalized count with 20 531 genes was used and for the miRNA expression, the $\log 2(x+1)$ transformed RPM value with 2166 miRNAs was used in this study. The clinical information of the three cancer studies was downloaded from cBioPortal (Gao *et al.*, 2013). In breast cancer study, we classify the cancer patients based on estrogen receptor (ER+ versus ER-) and triple negative (TN+ versus TN-) status. Triple negative breast cancer patients test negative for all three receptors that are commonly found in breast cancer: estrogen receptors, progesterone receptors and excess HER2 protein. For lung cancer and ovarian cancer studies, we classify the patients based on their survival time.

The miRNA–mRNA interaction network was obtained from TargetScanHuman (Agarwal *et al.*, 2015). TargetScanHuman reports effective miRNA–mRNA interactions with context++ model, thereby providing valuable gene-regulatory networks with the miRNA involved. miRNA can bind to mRNA to cause more rapid degradation of the mRNA molecule, therefore reducing the amount of protein translated from that mRNA. 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. The TF–gene interaction network was downloaded from RegNetwork (Liu *et al.*, 2015). The genes present in both lists of TFs and target genes were removed from the list of target genes. The modified bipartite interaction network contained sets of 1053 and 2859 non-overlapping genes representing TFs and their target genes, respectively, with 8170 total interactions between them.

3.2 Running omicsGAN on the TCGA datasets

To evaluate the proposed generative model on the TCGA omics datasets, we first updated the mRNA and miRNA (or TF and their target gene) expression profiles 5 times ($K = 5$). The generator and critic are fully connected neural networks with two hidden layers for the generator and one for the critic. The generator hidden layers have 512 and 768 neurons, respectively, whereas the critic hidden layers have 256 and 128 neurons, respectively. In both generator and critic, the activation function of the hidden layers is ReLU and the output layer is linear. Moreover, hidden layers in critic have dropout with a probability of 0.3. RMSprop optimizer was applied to train both the generator and the critic. Hyperparameters were

Table 2. Hyperparameters in omicsGAN used in the study

Hyperparameter	miRNA–mRNA			TF–gene
	BRCA	LUAD	OV	LUAD
Omics 1 generator learning rate	5e-6	5e-6	5e-6	5e-6
Omics 1 critic learning rate	5e-5	5e-5	5e-5	5e-5
Omics 1 L_2-norm coefficient (α)	0.01	0.01	0.1	0.0001
Omics 2 generator learning rate	5e-6	5e-6	5e-6	5e-6
Omics 2 critic learning rate	5e-5	5e-5	5e-5	5e-5
Omics 2 L_2-norm coefficient (α)	0.001	0.001	0.001	0.001

selected through grid search and details of the hyperparameters used in this study are listed in Table 2. In Table 2, Omics 1 is the mRNA/gene expression data for both interaction networks, Omics 2 is miRNA expression in miRNA–mRNA interaction network and TF in TF–gene interaction network. 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 for 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. The validation set described in Section 2 were employed for tuning all hyperparameters. All updated mRNA and miRNA (or gene and TF) datasets ($k = 1, 2, .., 5$) are sequentially fed into the classifier. The SVM-based classifier described in Section 2 was used for classification in all experiments. In the classifier, the dataset was divided into five folds with three folds for training, one fold for validation (parameter tuning and synthetic data update selection) and one fold for testing. We repeated the five-fold splitting 50 times on each dataset. The updated mRNA/gene expression (k^*) with the highest AUC score for validation samples was selected as the final synthetic mRNA/gene expression output from the model and similarly the updated miRNA/TF expression with the highest AUC score for validation samples was selected as the final synthetic miRNA/TF expression output. Figure 2 illustrates the process of selecting the final synthetic mRNA and miRNA datasets from all available updates for TCGA breast cancer patients outcome prediction. $k = 1$ gives the best validation AUC for synthetic mRNA expression whereas $k = 2$ gives the best validation AUC for synthetic miRNA expression. Therefore, mRNA update 1 and miRNA update 2 are used for predicting the test samples and the corresponding results are reported in this study.

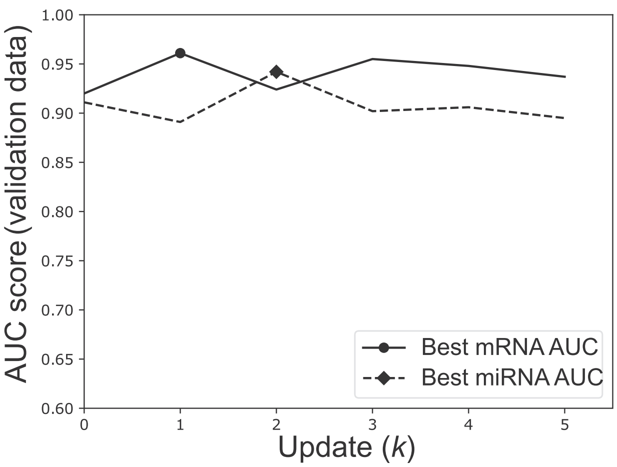

Fig. 2. Prediction results of triple negative (TN) status on TCGA breast cancer patients using validation samples. AUC of the prediction results using validation samples of synthetic mRNA and miRNA for $k = [1, 2, 3, 4, 5]$. Update k^* with the best validation AUC is selected as the final synthetic data for each omics profile

One synthetic data is generated for breast cancer ER and TN status prediction based on the average validation AUC of the two clinical variables.

3.3 Integration of mRNA and miRNA expression

We generate the synthetic mRNA and miRNA datasets by integrating the two omcis profiles and their interaction network and assess the quality of the synthetic data through three experiments.

3.3.1 omicsGAN improved cancer outcome prediction

To evaluate the quality of the synthetic datasets generated by omicsGAN, we designed cancer outcome prediction and significant predictive signature identification tasks on the TCGA breast cancer, lung cancer and ovarian cancer datasets under the assumptions: (i) The synthetic datasets learned in omicsGAN consider the expressions in both mRNA and miRNA profiles and the biological interactions between them. So they will provide better predictive signatures compared to mRNA and miRNA expressions. (ii) The better predictive signatures will improve the disease phenotype prediction.

We ran the classifier with above mentioned five-fold splitting 50 times to select the best synthetic data among the 5 updates based on validation samples and classify the test samples using the selected synthetic data. The average AUC scores of 50 splittings are reported in Table 3. 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, 95 cancer patients below the survival time cutoff (<25 months) and 64 above the cutoff (>50 months) in the lung cancer dataset as well as 61 cancer patients below the survival time cutoff (<25 months) and 77 above the cutoff (>50 months) in the ovarian cancer dataset. Table 3 illustrates that the synthetic mRNA and miRNA expression generated by omicsGAN achieved better average classification results than original mRNA and miRNA expression for phenotype predictions across all three cancer types. We also add the baseline where we perform the classification with concatenated miRNA and mRNA expression to see whether addition of more omics data is the reason for the improvement. We can see that concatenated data has similar or better prediction ability compared to the original mRNA and miRNA expression dataset; however, synthetic dataset from omicsGAN always outperforms the concatenated data by a significant margin. This signifies that even though the addition of more omics data improves the outcome prediction performance, omicsGAN relies on the interaction network to generate synthetic data with better predictive signal.

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 Student's t-test on the expression datasets with different clinical variables. The number of features with a P-value smaller than 0.001 in each dataset except miRNA

Table 3. The classification performance on TCGA breast cancer, lung cancer and ovarian cancer datasets

Input data	Breast cancer		Lung cancer	Ovarian cancer
	ER	TN	Survival time	Survival time
mRNA	0.913	0.91	0.675	0.651
synthetic mRNA (omicsGAN)	0.948[a]	0.949[a]	0.733[a]	0.708[a]
miRNA	0.878	0.904	0.595	0.627
synthetic miRNA (omicsGAN)	0.945[a]	0.938[a]	0.733[a]	0.721[a]
mRNA+miNRA	0.905	0.921	0.67	0.658

Note: Average AUC scores of classify cancer patients clinical variables on the synthetic mRNA, miRNA datasets generated from omicsGAN and the original mRNA, miRNA expression datasets.

[a] The difference between the results on the original expression data and the synthetic data is statistically significant (P-value < 0.001).

Table 4. Number of significant features

Input data	Breast cancer		Lung cancer	Ovarian cancer
	ER	TN	Survival time	Survival time
mRNA	4144	3893	227	133
synthetic mRNA (omicsGAN)	4566	4241	372	142
miRNA	91	91	23	20
synthetic miRNA (omicsGAN)	136	127	58	12

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

expression for lung cancer patients are presented in Table 4. *P*-value cutoff of 0.05 is set for miRNA expression for lung cancer patients as no feature had a *P*-value smaller than 0.001 in either the real miRNA expression or the synthetic one. We can see an increased number of significant features in synthetic mRNA compared to the original one for all three cancer types. Synthetic miRNA on the other hand has more significant features for breast cancer and lung cancer, but less for ovarian cancer compared to the original miRNA expression datasets. Therefore, omicsGAN enriches the features of synthetic datasets with better predictive signatures that results into improved cancer outcome prediction.

3.3.2 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. Tables 3 and 4 already illustrates the ability of omicsGAN to improve the cancer outcome prediction performance. However, we hypothesized that omicsGAN harnesses the information of biological interaction between two omics layers from multi-omics interaction network to generate the synthetic datasets with better predictive signals. Hence, we want to investigate whether the improvement in performance is because of the additional omics data or the model can exploit the interaction network for data integration. We design an experiment to explore the effects of the interaction network on synthetic omics data and their predictive performance where we ran the framework 10 times with same settings and input $\boldsymbol{X}$ (mRNA expression), $\boldsymbol{Y}$ (miRNA expression) as before but a different interaction network on TCGA lung cancer datasets. We replaced the true network with 10 different randomized networks with same density as the true one. The prediction results for synthetic mRNA and miRNA expression using true and random networks are shown as boxplots in Figures 3 and 4, respectively. Prediction results using original mRNA/miRNA expression, synthetic mRNA/miRNA expression generated using the true network, and synthetic mRNA/miRNA expression generated using random network are plotted in each figure. The first two boxplots display the same results for lung cancer outcome prediction as shown in Table 3. Fifty dots in each of these two boxplots represent the AUC corresponding to 50 random splittings. The third boxplot illustrates the results using 10 random networks, each with 50 splittings. The statistics (mean, median and standard deviation) of the prediction performance of the splittings are shown above each boxplot. In Figures 3 and 4, we see a reduction in performance of synthetic mRNA/miRNA 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.

3.3.3 omicsGAN improved survival prediction

To further investigate the quality of the synthetic mRNA and miRNA expression data produced by omicsGAN, the patient's overall survival was predicted on breast cancer, lung cancer and ovarian cancer datasets. The Cox proportional hazards model with elastic net penalty as described in Section 2.3 evaluates the correlation between patient's overall survival and genomic features, i.e. the original mRNA, miRNA expressions and the synthetic mRNA, miRNA expressions in this study. The relative weight r in Equation 9 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 (***PI***) as mentioned in Section 2.3. The survival predictions were visualized by Kaplan–Meier plots and compared by the log-rank test *P*-values. The Kaplan–Meier plots in Figures 5 and 6 exemplify the improved patient survival predictions on lung cancer using the synthetic mRNA, miRNA expressions generated by omicsGAN compared to the original mRNA, miRNA expressions. The log-rank

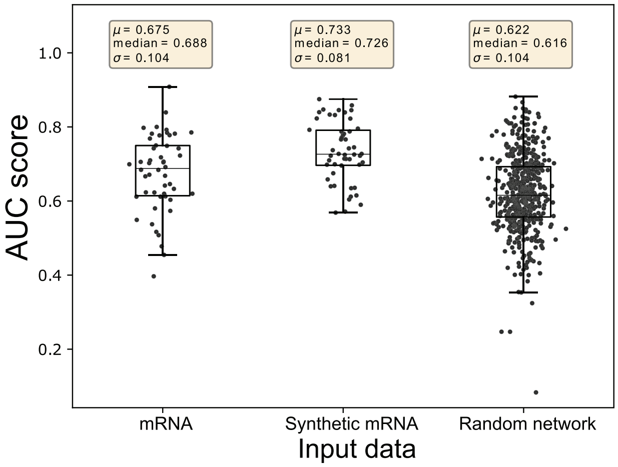

Fig. 3. Prediction results of the survival time on TCGA lung cancer patients using original and synthetic mRNA expression. Prediction results using original mRNA expression, synthetic mRNA expression generated using true interaction network, and synthetic mRNA expression generated using random interaction network are plotted, respectively. Each dot represents the AUC score from one splitting. The statistics (mean, median and standard deviation) of the prediction performance of the 50 splittings are shown above each boxplot

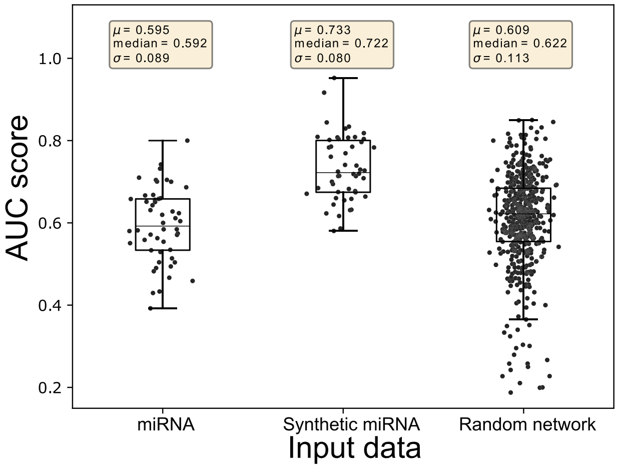

Fig. 4. Prediction results of the survival time on TCGA lung cancer patients using original and synthetic miRNA expression. Prediction results using original miRNA expression, synthetic miRNA expression generated using true interaction network, and synthetic miRNA expression generated using random interaction network are plotted, respectively. Each dot represents the AUC score from one splitting. The statistics (mean, median and standard deviation) of the prediction performance of the 50 splittings are shown above each boxplot

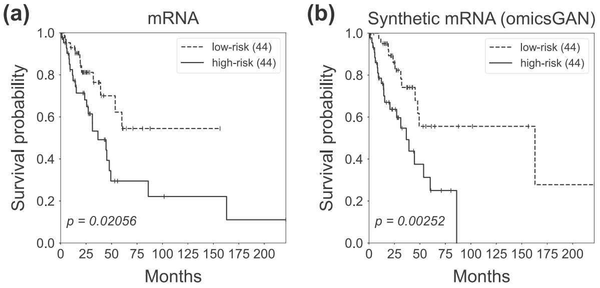

Fig. 5. Survival prediction on lung cancer patients with mRNA profiles. Kaplan–Meier survival plots for high (solid line) and low (dashed line) risk groups generated by (**a**) original mRNA, (**b**) synthetic mRNA expression data on lung cancer patients. The number in the parenthesis indicates the number 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

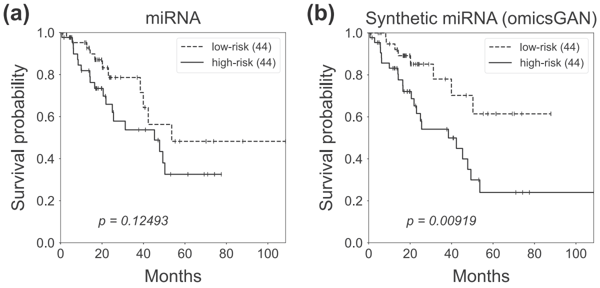

Fig. 6. Survival prediction on lung cancer patients with miRNA profiles. Kaplan–Meier survival plots for high (solid line) and low (dashed line) risk groups generated by (**a**) original miRNA, (**b**) synthetic miRNA expression data on lung cancer patients. The number in the parenthesis indicates the 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

test *P*-values clearly demonstrate a strong additional prognostic power of the synthetic omics profiles beyond the original signatures. Similar observations are identified on breast and ovarian cancer patient samples (Supplementary Figs S1–S4).

3.4 Integration of TF and gene expression

The experiments above show the ability of omicsGAN to generate synthetic data with better predictive power by harnessing the information from miRNA–mRNA interaction network. Here, we design another experiment using TF–gene interaction network to evaluate whether omicsGAN can show similar improvement in integrating other omics data and their interaction network. We performed the lung cancer phenotype prediction based on the same classification setup as described in Section 3.3.1 on TFs and their target gene expression datasets. The average AUC scores of 50 splittings are reported in Table 5. Both the synthetic TF and target gene expression performed better in classifying the lung cancer patients based on their survival time than the original TF, gene expression and concatenated TF and gene expression. These findings signify that our proposed framework can work with varying set multi-omics data.

4 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 data 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 inadequate 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. In most studies, these inter-omics relations are neglected and it is inefficient to predict phenotype using integrated multi-omics data without considering the interactions. Therefore, the integrating of the bipartite interaction network with multi-omics data can result in improved disease phenotype prediction and designing frameworks capable of such integration is gaining importance.

Table 5. The classification performance on TCGA lung cancer dataset

Input data	Lung cancer
Gene	0.645
Synthetic gene	0.727[a]
TF	0.656
Synthetic TF	0.743[a]
Gene+TF	0.682

Note: Average AUC scores of classification performance between synthetic gene, TF generated from omicsGAN and the original gene, TF expression on lung cancer datasets.

[a]The difference between the results on the original expression data and the synthetic data is statistically significant (P-value < 0.001).

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. All synthetic datasets generated in this study with two interaction networks (i.e. miRNA–mRNA and TF–gene) perform better in cancer outcome prediction compared to the original expression datasets; however, the same model using a random interaction network with same density does not perform as good as the synthetic datasets obtained through true network. It signifies that omicsGAN does not fuse information from the two omics data directly; rather functionally incorporate the interaction network into the integration. Synthetic miRNA expression using random interaction network works better than the original miRNA expression (Fig. 4) but synthetic mRNA using random interaction network does not perform better than original mRNA expression (Fig. 3). The reason is, without the true interaction network, omicsGAN can still integrate information from the two omics data to generate synthetic datasets. In that case, the performance of one synthetic data will depend on the additional information received from the other omics data. Synthetic miRNA receives information from mRNA expression, which is significantly better in lung cancer outcome prediction compared to miRNA and thus improves the performance of synthetic miRNA. Synthetic mRNA on the other hand receives information from miRNA that is worse at prediction compared to mRNA and thus results in a decreased performance. An 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.

The framework presents an innovative way for multi-omics data integration incorporating their biological interaction. 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 and TF–gene interaction, the same technique can be extrapolated to any two omics data if their interaction network is biologically meaningful. 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. In this study, all missing data is imputed by zero. The prediction performance can be further improved using advanced data imputation frameworks (Nagpal *et al.*, 2019; Song *et al.*, 2020; Zhou *et al.*, 2020) and multi-omics pre-processing methods (Sharifi-Noghabi *et al.*, 2019).

5 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 GAN 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 was supported by grants from the National Science Foundation (NSF) [NSF-III1755761]; National Institutes of Health (NIH) [2R01GM113952-06, DK097771].

Conflict of Interest: none declared.

References

Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *elife*, **4**, e05005.

Ahmed,K.T. *et al.* (2020) Network-based drug sensitivity prediction. *BMC Med. Genomics*, **13**, 1–10.

Ahmed,K.T. *et al.* (2021) In silico model for miRNA-mediated regulatory network in cancer. *Brief. Bioinf.*

Argelaguet,R. *et al.* (2018) Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.

Arjovsky,M. *et al.* (2017) Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Davidson-Pilon,C. (2019) lifelines: survival analysis in Python. *J. Open Source Softw.*, **4**, 1317.

Gao,J. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.

Ghahramani,A. *et al.* (2018) Generative adversarial networks simulate gene expression and predict perturbations in single cells. *BioRxiv*, 262501.

Goldman,M.J. *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.*, **38**, 675–678.

Goodfellow,I. *et al.* (2014) Generative adversarial nets. *Adv. Neural Inf. Process. Syst.*, **27**, 2672–2680.

Goodwin,S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.

Hentze,M.W. *et al.* (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, **19**, 327–341.

Kim,M. *et al.* (2018) An improved method for prediction of cancer prognosis by network learning. *Genes*, **9**, 478.

Kipf,T.N. and Welling,M. (2016) Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Koh,H.W. *et al.* (2019) iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst. Biol. Appl.*, **5**, 22.

Krzyszczyk,P. *et al.* (2018) The growing role of precision and personalized medicine for cancer treatment. *Technology*, **6**, 79–100.

Liu,Z.-P. *et al.* (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, **2015**, 15806.

Nagpal,S. *et al.* (2019) TIGAR: an improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *Am. J. Hum. Genet.*, **105**, 258–266.

Nguyen,H. *et al.* (2019) PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, **35**, 2843–2846.

Nussbacher,J.K. and Yeo,G.W. (2018) Systematic discovery of RNA binding proteins that regulate microRNA levels. *Mol. Cell*, **69**, 1005–1016.

Park,J. *et al.* (2020) A practical application of generative adversarial networks for RNA-seq analysis to predict the molecular progress of Alzheimer's disease. *PLoS Comput. Biol.*, **16**, e1008099.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pölsterl,S. (2020) scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.*, **21**, 1–6.

Rappoport,N. and Shamir,R. (2019) NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, **35**, 3348–3356.

Sharifi-Noghabi,H. *et al.* (2019) MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, **35**, i501–i509.

Simon,N. *et al.* (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.*, **39**, 1.

Song,M. *et al.* (2020) A review of integrative imputation for multi-omics datasets. *Front. Genet.*, **11**, 570255.

Subramanian,I. *et al.* (2020) Multi-omics data integration, interpretation, and its application. *Bioinf. Biol. Insights*, **14**, 1177932219899051.

The Cancer Genome Atlas Network. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.

The Cancer Genome Atlas Research Network. *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609.

The Cancer Genome Atlas Research Network. *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543.

Wang,B. *et al.* (2014a) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Wang,C. *et al.* (2014b) Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods*, **67**, 304–312.

Xu,Y. *et al.* (2020) scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res.*, **48**, e85.

Yeh,H.-S. *et al.* (2017) Analyses of alternative polyadenylation: from old school biochemistry to high-throughput technologies. *BMB Rep.*, **50**, 201–207.

Zhang,H. *et al.* (2017) Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy. Oct 22, 2017 – Oct 29, 2017, pp. 5907–5915.

Zhou,T. *et al.* (2019) Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Hum. Brain Map.*, **40**, 1001–1016.

Zhou,X. *et al.* (2020) Imputing missing RNA-seq data from DNA methylation by using transfer learning based neural network. *bioRxiv*, 803692.