frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 02 December 2022
DOI 10.3389/fbinf.2022.1025783

Check for updates

OPEN ACCESS

EDITED BY
Swarup Roy,
Sikkim University, India

REVIEWED BY
Pietro Hiram Guzzi,
Magna Græcia University, Italy
Assaf Gottlieb,
University of Texas Health Science
Center at Houston, United States

*CORRESPONDENCE
Roded Sharan,
roded@tauex.tau.ac.il

SPECIALTY SECTION
This article was submitted to Network Bioinformatics,
a section of the journal
Frontiers in Bioinformatics

RECEIVED 23 August 2022
ACCEPTED 21 November 2022
PUBLISHED 02 December 2022

CITATION
Gervits A and Sharan R (2022), Predicting genetic interactions, cell line dependencies and drug sensitivities with variational graph auto-encoder.
Front. Bioinform. 2:1025783.
doi: 10.3389/fbinf.2022.1025783

Predicting genetic interactions, cell line dependencies and drug sensitivities with variational graph auto-encoder

Asia Gervits and Roded Sharan*

School of Computer Science, Tel Aviv University, Tel Aviv-Yafo, Israel

Large scale cancer genomics data provide crucial information about the disease and reveal points of intervention. However, systematic data have been collected in specific cell lines and their collection is laborious and costly. Hence, there is a need to develop computational models that can predict such data for any genomic context of interest. Here we develop novel models that build on variational graph auto-encoders and can integrate diverse types of data to provide high quality predictions of genetic interactions, cell line dependencies and drug sensitivities, outperforming previous methods. Our models, data and implementation are available at: https://github.com/aijag/drugGraphNet.

KEYWORDS

deep learning, variational graph auto-encoder, genetic interaction, drug sensitivity, cell-line dependency

1 Introduction

Large scale cancer genomics data provide crucial information about the disease and reveal points of intervention. However, systematic data haven been collected in specific cell lines and their collection is laborious and costly. Hence, there is a need to develop computational models that can predict such data for any genomic context of interest.

Several large scale data sets that can be used for developing such models exist. The cancer dependency map project pinpoints potential drug targets in cancer lines whose knockdown leads to decreased cell fitness (DepMap, 2021). Systematic genetic interaction screens conducted in yeast and in human provide complementary information on potential cell-specific targets (Lee et al., 2018). The GDSC project performs systematic drug screens to identify sensitive cancer cell lines (Iorio et al., 2016).

There is a plethora of previous methods to predict these large scale data sets. Genetic interactions have been predicted based on gene ontology information (Yu et al., 2016; Ma et al., 2018), mutation and expression data (Lee et al., 2018), protein-protein interaction (PPI) data (Hao et al., 2021) and by specifically-designed deep learning models (Ma et al., 2018; Cai et al., 2020). Gene dependencies have been predicted based on expression information (Itzhacky and Sharan, 2021; Lin and Lichtarge, 2021), pathway information (Lin and Lichtarge, 2021), genetic essentiality profiles (Wang et al., 2019), and PPI and genomic alteration information (Benstead-Hume et al., 2019). Drug sensitivity data have been predicted based drug structure information combined with gene ontology information (Kuenzi et al., 2020) or gene expression data (Wang et al., 2017; Zhang et al., 2018; Choi et al., 2020; Itzhacky and Sharan, 2021). However, each method uses different information sources and most are geared toward a single prediction task.

Graph convolution networks (GCNs) and variational graph auto encoders (VGAEs) are powerful neural network architectures on graphs that can effectively capture the graph structure, perform node classifications and link prediction and are widely applicable (Scarselli et al., 2008; Kipf and Welling, 2016; Li et al., 2019). These techniques were also employed in the cancer genomics domain but again targeting a single task each time (Cai et al., 2020; Fan et al., 2020; Kuenzi et al., 2020; Ding et al., 2021; Hao et al., 2021).

In this paper, we develop an integrated model that combines VGAE with gene ontology information to perform a wide range of predictions spanning genetic interactions, gene dependencies and drug sensitivities. Our model is the first to propagate gene ontology information within a combined network of genetic interactions, gene-cell line relations and drug-target relations. It is shown to outperform previous methods for each of the prediction tasks. Its unique features include a new normalization layer and a modular architecture that allows the prediction of multiple attributes, represented as links in this model.

2 Methods

2.1 Data collection

2.1.1 Genetic interaction (GI) data

We used genetic interactions from three different sources: (i) A yeast GI dataset from (Costanzo et al., 2016) downloaded from https://thecellmap.org/costanzo2016/. We used the provided thresholds of p-value threshold ≤0.05 and GI score ≤0.08, to extract ~240K negative GIs. For the neutral pairs we used the same p-value threshold and score higher than 0.08. (ii) A human GI dataset collected in 2 cell lines, K562 and Jurkat, from (Horlbeck et al., 2018). We focused on the larger K562 dataset due to the high correlation between the two datasets. We used the reported threshold of -3, resulting in 1,678 negative GIs. (iii) SynLethDB collection of human synthetic lethality (SL) interactions from (Guo et al., 2016) with 19,667 SL pairs among 6,375 genes.

2.1.2 Achilles gene dependency data

We downloaded gene dependency data from https://depmap.org (Dempster et al., 2019; DepMap, 2021), version 21Q4. We used a dependency threshold of 0.5 as in the recently published method (BioVNN, described below) (Lin and Lichtarge, 2021). For constructing our model we also downloaded CCLE (Ghandi et al., 2019) mutations for each cell line and selected the damaging mutations by the variant annotations. We excluded genes which were either nearly all dependent (up to 6) across cell lines, as in BioVNN. The final constructed dataset contains 922 cell lines and their affect among 5,975 genes, spanning ~1.4M dependent pairs.

2.1.3 Drug sensitivity data

We downloaded the GDSC binarized IC50 dataset from http://www.cancerrxgene.org/(Iorio et al., 2016). The dataset consists of 1,001 cancer cell line and 265 tested drugs, spanning ~20K sensitive pairs. In addition, we downloaded from the same site the gene targets for each drug and the mutated genes for each cell line that were used to construct our model. Due to lack of variant annotation information in this datasource, we focused on nonsense, frame shift, exonic splicing silencer and gene fusion mutations that we considered as harmful.

2.1.4 Gene ontology (GO)

For our feature generation we downloaded the latest version of ontology file from http://geneontology.org/. We used all the terms from the three GO subnetworks: biology process (BP), cellular components (CC) and molecular functions (MF). Following the original publication we removed terms with the evidence code ‘inferred by genetic interaction’ (IGI), to avoid potential circularity in predicting genetic interactions. In addition, terms that do not connect to any gene in the model’s graph were removed.

2.2 Link prediction algorithm

Graph autoencoder (GAE) and variational graph autoencoder (VGAE) models have been demonstrated as efficient tools to learn graph embeddings in an unsupervised way and serving as an infrastructure for link prediction (Kipf and Welling, 2016). Here we combined the VGAE model with gene ontology information to generate a modular graph structure that models the connections between drugs, cell lines and genes and could be used to a wide range of tasks: predicting genetic interactions, cell line dependencies and drug sensitivities.

Given a graph G on a set of n vertices V and a real adjacency matrix A, a graph convolutional network (GCN) model receives two matrices as inputs: $A \in R^{n \times n}$ and $X \in R^{n \times f}$ as the feature matrix of V. The output of a single layer is $\sigma(\hat{A}\delta(X)W)$ Where σ is the activation function, δ is a dropout applied on the input, $\bar{A} := A + I$, $\bar{D}$ the corresponding diagonal degree matrix and $\hat{A} := \bar{D}^{-\frac{1}{2}}\bar{A}\bar{D}^{-\frac{1}{2}}$, W the learnable weight matrices.

2.2.1 Encoder

In our model we use a normalization layer that we find to improve model performance, followed by a 2-layer GCN which computes the node embedding distribution by Eqs. 1, 2: $\mu \in R^{n \times f}$

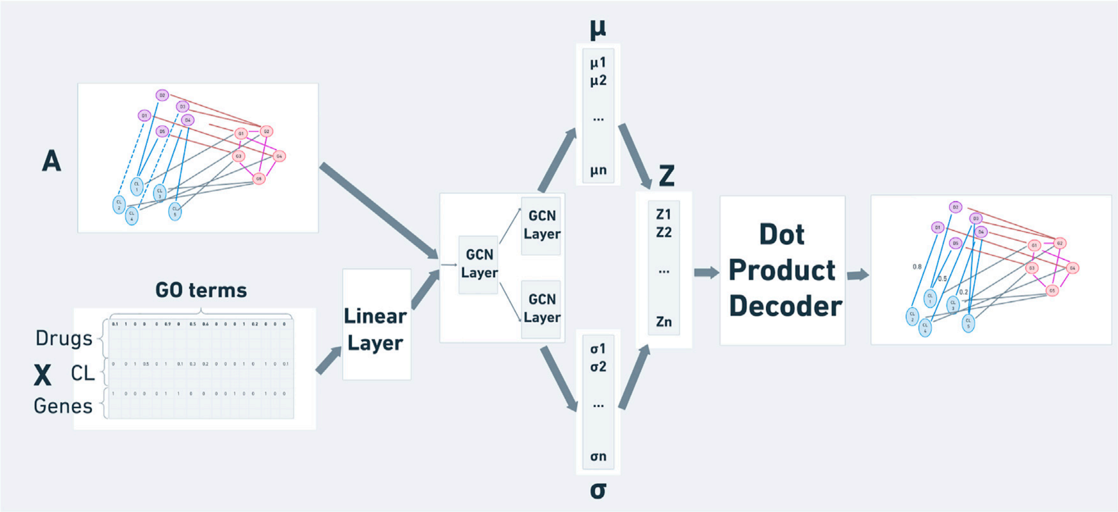

FIGURE 1
An illustration of our final constructed model for drug sensitivity prediction. The adjacency matrix *A* with the masked sensitivity interactions and feature matrix *X* serve as inputs. The final output is the reconstructed $\tilde{A}$ adjacency matrix, allowing the prediction of new drug-cell line interactions.

is the matrix of mean vectors, log $\sigma^2 \in R^{n\times f}$ is (log of) the variance matrix, and f represents the dimension of the embedding node vectors. From those distribution matrices we draw the embedding matrix Z by $z = \mu + \sigma * \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$. In the equations, σ_j are the activation functions, W_is are learnable weight matrices and b is a learnable bias vector. The contribution of the additional normalization layer, compared to the standard VGAE performance, is summarized in Supplementary Figure S1.

$$\bar{X} = XW_0 + b \tag{1}$$

$$\mu = \sigma_2\left(\hat{A}\sigma_1\left(\hat{A}\bar{X}W_1\right)W_2\right); log\ \sigma^2 = \sigma_2\left(\hat{A}\sigma_1\left(\hat{A}\bar{X}W_1\right)W_3\right) \tag{2}$$

2.2.2 Decoder

The decoder is defined by the dot product between latent Z variables, and the output is a reconstructed adjacency matrix $\tilde{A}$ as follows:

$$\tilde{A} = \sigma_3(ZZ^T) \tag{3}$$

where σ_3 is the sigmoid function.

2.2.3 Loss function

The loss function of VGAE includes two parts. The first part is the binary cross-entropy between the target A and the model output, while the second part is the KL-divergence between $q(Z|X,A) = \Pi_{i=1}^{N} q(z_i|X,A) = \Pi_{i=1}^{N}\mathcal{N}(z_i|\mu_i, diag(\sigma_i^2))$ and $p(Z) = \Pi_i p(z_i) = \Pi_i \mathcal{N}(z_i|0,I)$, this part aims to generate the latent dimension with Gaussian distribution. The final loss function is defined as follows:

$$L = E_{q(Z|X,A)}[\log p(A|Z)] - KL[q(Z|X,A)\|p(Z)] \tag{4}$$

The full model we developed is illustrated in Figure 1. Details about the underlying graphs and features are provided below.

2.2.4 Graph construction

The graphs underlying our models are gradually built in three parts. The first part is a graph of GIs only (Figure 2A), here the nodes represent genes and the edges represent GIs. We will represent this model as $VGAE_G$. The second part are nodes representing cell lines that are connected to the first part nodes using dependency relations (Figure 2B), we will represent this model as $VGAE_{CD}$. The final part consists also of drug nodes which are connected to the gene part using drug-target relations. The connections between drugs and cell lines represent the drug sensitivity, which is the target of our prediction in this model, we will represent this model as $VGAE_{DS}$ (Figure 2C). In the $VGAE_{DS}$ graph, cell lines are connected to their mutated genes (rather than using dependency relations like in $VGAE_{CD}$).

2.2.5 Feature generation

To generate the features vector for each node we used the ontotype method (Yu et al., 2016), where each gene is represented by a binary vector of its GO terms and a gene set by the sum of its member gene vectors. Specifically, we associated cell-lines nodes with their sets of mutated genes, and drug nodes with their sets of gene targets. All nodes vectors were normalized by dividing by the number of genes they represent. The final features included the terms that were connected to at least one of the drugs, cell lines or genes, so all the three types of nodes is sharing the same

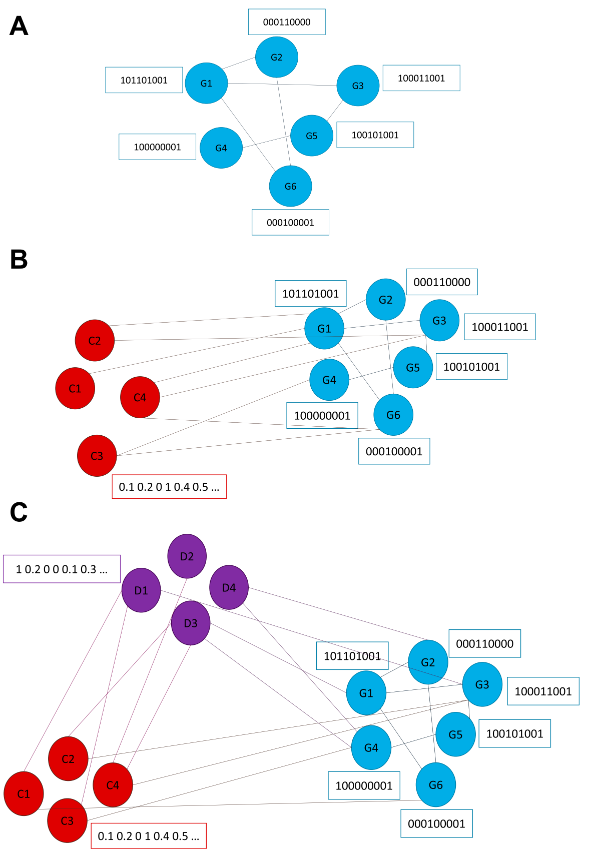

FIGURE 2
Graph model structures that underlie our VGAE model. **(A)** $VGAE_G$—GI part with gene nodes. **(B)** $VGAE_{CD}$—Cell line dependency part with nodes representing genes and cell lines. **(C)** $VGAE_{DS}$—Drug sensitivity part with nodes representing genes, cell-lines and drugs.

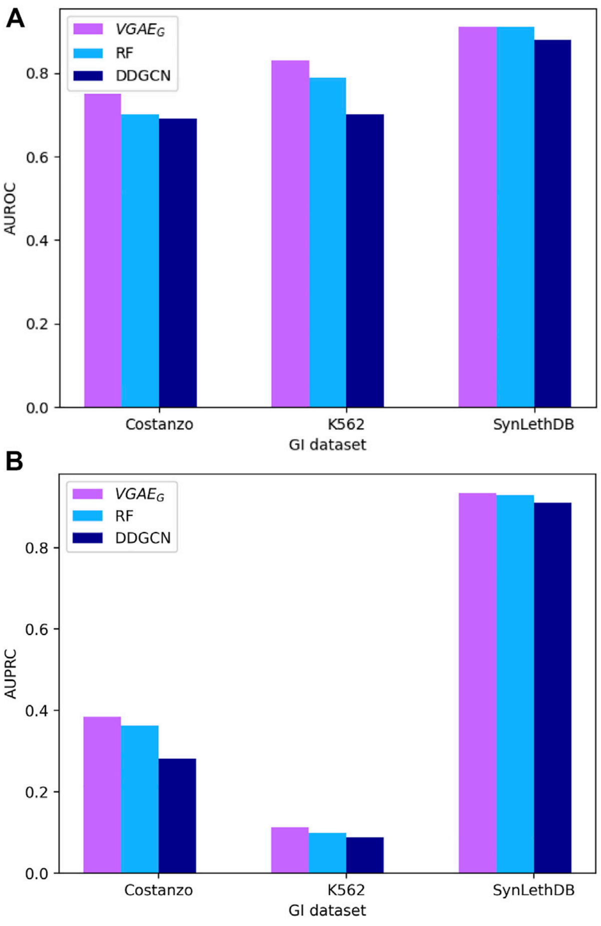

FIGURE 3
Performance of GI prediction across different datasets. **(A)** AUROC results. **(B)** AUPRC results.

features dimension. GO annotations with the evidence code “inferred by genetic interaction” (IGI) were removed to avoid potential circularity in predicting genetic interactions.

2.3 Training procedure and performance evaluation

To evaluate our model we performed five-fold cross-validation (CV). For highly imbalanced datasets, like drug sensitivity, we also generated temporary balanced datasets by randomly sampling neutral samples of the same size as the positive samples, and performed 10 repetitions of the CV. For our model training we used randomly selected 10% edges from the training data as a validation set. The validation set was used for the evaluation of the model during the training for selecting the model from the epoch best performances, and for early stopping of the training process in the case of overfitting. We optimized the model’s hyperparameters using the validation data, choosing the hyperparameter configuration that performed best.

As a benchmark in all prediction tasks we compared to the ontotype method (Yu et al., 2016) that we build on. In this method, the ontotype feature vectors are fed into a random forest (RF) classifier in the prediction phase rather than being propagated in a graph as in our new model. For the RF training, we used the same split on the edges. The input features for the cell dependency and drug sensitivity tasks are the sum of the cell line and drug features that we used in our model.

We calculated true positive rate (TPR), false positive rate (FPR), precision and recall by varying the preset thresholds to construct receiver operating characteristic (ROC) and precision–recall (PR) curves. We then generated two metrics, namely the area under the ROC curve (AUROC) and the area under the PR curve (AUPRC), to evaluate the performance of our model and other methods. We averaged the results from all cross-validation splits to calculate the overall AUROC and AUPRC.

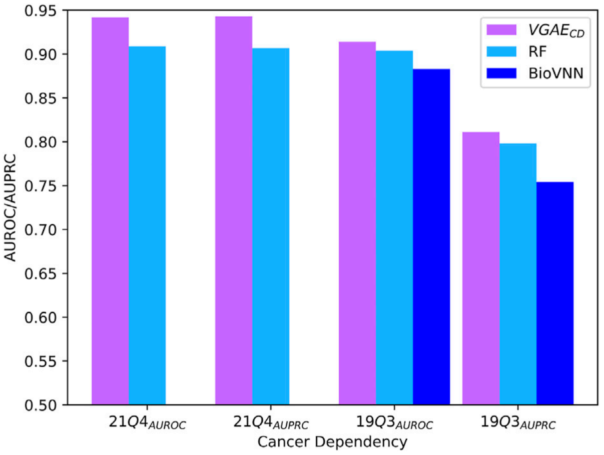

FIGURE 4
Performance evaluation for cancer dependency prediction on the two evaluated datasets. The presented BioVNN results are taken from (Lin and Lichtarge, 2021).

3 Results

3.1 $VGAE_G$ model for predicting genetic interactions

To evaluate our VGAE model we applied its first variant $VGAE_G$ to predict genetic interactions and compared to the ontotype method (Yu et al., 2016). We tested three data sets from yeast (systematic GI data) and human (systematic data from the K562 cell-line and interactions from SynLethDB) as described in the Data section of the Methods. Our model outperformed the previous method (Figure 3; Supplementary Table S1). To analyze the contribution of our model on top of the ontotype method, we split the systematic datasets to two equal parts based on the ontotype sparsities, observing that the higher the sparsity the higher the contribution (Supplementary Table S2). In addition we also compared our model to a recent GCN model, DDGCN (Cai et al., 2020), that does not use the GO knowledge or the additional normalization layer. The results show the benefit of integrating GO information within a GCN and motivate the more complex models below.

3.2 $VGAE_{CD}$ model for predicting cancer dependencies

Next, we tested our second model variant $VGAE_{CD}$ on the recent 21Q4 cell-line dependency dataset. In this application the input genetic interactions were taken from SynLethDB. For comparison purpose, we adapted the ontotype method to this setting, representing a cell line by the normalized sum of ontotype vectors of genes it depends on. In addition, we compared our model to a recently published method (BioVNN) which uses a visible neural network over pathway knowledge to predict dependencies (Lin and Lichtarge, 2021). For this comparison, we analyzed the same 19Q3 dataset used in the previous paper containing 609 cell-lines and 683 genes. We also employed the same cross validation procedure where cell lines are distinct between the train, validation and test folds. The AUROC and AUPRC results are summarized in Figure 4; Supplementary Table S1, and again show the superiority of our model.

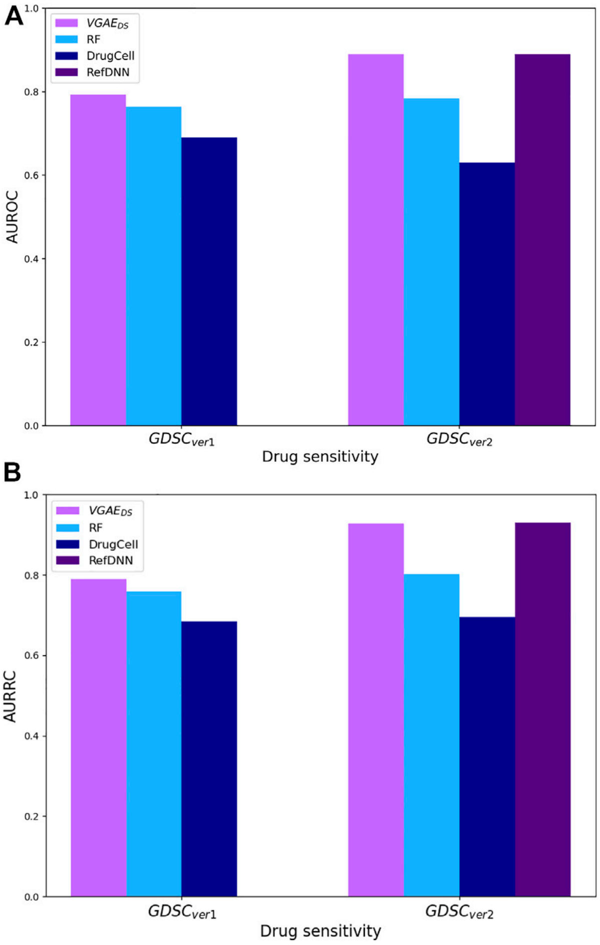

FIGURE 5
Prediction performance on drug sensitivity data. **(A)** AUROC results. **(B)** AUPRC results.

3.3 $VGAE_{DS}$ model for predicting drug sensitivity data

Last, we applied our full model to predict drug sensitivity relations. For comparison purpose we again adapted the ontotype method for this task by representing each drug (cell line, resp.) by the normalized sum of the ontotype vectors of its targets (mutated genes, resp.). We further compared ourselves to DrugCell (Kuenzi et al., 2020), a deep network that similarly to our model uses GO information and cell-line mutations, but unlike our model uses drug chemical structure as additional input. Since DrugCell is designed for a regression task, we adapted it for classification by changing: (i) the last activation function to sigmoid activation; and (ii) the loss function to binary cross-entropy. In addition, we tested our model on the different version of GDSC dataset (version 17.3) that was used for the training of a recently published method (RefDNN), a deep NN that uses gene expression and drug structure as inputs (Choi et al., 2020). In this dataset, IC50 continuous values were binarized based on the reported maximum screening concentration threshold. The comparison results are summarized in Figure 5; Supplementary Table S1, and show that $VGAE_{DS}$ outperforms the other methods or receives similar results in both datasets.

4 Discussion

We have presented a graph variational auto-encoder based model for predicting genetic interactions, cell line dependencies and drug sensitivities. The model propagates gene ontology information over a network of gene, drug and cell-line interactions, providing uniform representations to genes, cell lines and drugs, allowing the wide scale of predictions. The unique features of the model include a new normalization layer and a modular architecture that allows the prediction of multiple attributes. While our models achieved promising results, their performance in a real clinical setting where samples come from real patients will need to be assessed when such data becomes available.

For future work, we would like to create one model that can predict genetic interactions, cell line dependencies and drug sensitivities, rather than having three separate models. To this end, the connections between cell lines and genes would represent cancer dependency instead of cell lines mutations. This model structure is currently not feasible due to the low number of overlapping cell lines between dependency and sensitivity data.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

AG is the main writer of this paper as part of her master's degree, RS is her supervisor in this work.

Funding

RS was supported by a research grant from the Israel Science Foundation (grant no. 715/18) and by the Koret–UC Berkeley–Tel Aviv University Initiative in Computational Biology and Bioinformatics.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2022.1025783/full#supplementary-material

References

Benstead-Hume, G., Wooller, S. K., Dias, S., Woodbine, L., Carr, A. M., and Pearl, F. M. (2019). Biological network topology features predict gene dependencies in cancer cell lines. *bioRxiv* 2019, 751776.

Cai, R., Chen, X., Fang, Y., Wu, M., and Hao, Y. (2020). Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. *Bioinformatics* 36, 4458–4465. doi:10.1093/bioinformatics/btaa211

Choi, J., Park, S., and Ahn, J. (2020). Refdnn: A reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci. Rep.* 10, 1861. doi:10.1038/s41598-020-58821-x

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353, aaf1420. doi:10.1126/science.aaf1420

Dempster, J. M., Rossen, J., Kazachkova, M., Pan, J., Kugener, G., Root, D. E., et al. (2019). Extracting biological insights from the project achilles genome-scale crispr screens in cancer cell lines. *BioRxiv* 2019, 720243.

[Dataset] DepMap, B. (2021). Depmap 21q4 public. doi:10.6084/M9.FIGSHARE.16924132.V1

Ding, Y., Tian, L.-P., Lei, X., Liao, B., and Wu, F.-X. (2021). Variational graph auto-encoders for mirna-disease association prediction. *Methods* 192, 25–34. doi:10.1016/j.ymeth.2020.08.004

Fan, K., Guan, Y., and Zhang, Y. (2020). Graph2go: A multi-modal attributed network embedding method for inferring protein functions. *GigaScience* 9, giaa081. doi:10.1093/gigascience/giaa081

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature* 569, 503–508. doi:10.1038/s41586-019-1186-3

Guo, J., Liu, H., and Zheng, J. (2016). Synlethdb: Synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res.* 44, D1011–D1017. doi:10.1093/nar/gkv1108

Hao, Z., Wu, D., Fang, Y., Wu, M., Cai, R., and Li, X. (2021). Prediction of synthetic lethal interactions in human cancers using multi-view graph auto-encoder. *IEEE J. Biomed. Health Inf.* 25, 4041–4051. doi:10.1109/JBHI.2021.3079302

Horlbeck, M. A., Xu, A., Wang, M., Bennett, N. K., Park, C. Y., Bogdanoff, D., et al. (2018). Mapping the genetic landscape of human cells. *Cell.* 174, 953–967.e22. doi:10.1016/j.cell.2018.06.010

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell.* 166, 740–754. doi:10.1016/j.cell.2016.06.017

Itzhacky, N., and Sharan, R. (2021). Prediction of cancer dependencies from expression data using deep learning. *Mol. Omics* 17, 66–71. doi:10.1039/d0mo00042f

Kipf, T. N., and Welling, M. (2016). *Variational graph auto-encoders. arXiv preprint arXiv:1611.07308*.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., et al. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell.* 38, 672–684.e6. doi:10.1016/j.ccell.2020.09.014

Lee, J. S., Das, A., Jerby-Arnon, L., Arafeh, R., Auslander, N., Davidson, M., et al. (2018). Harnessing synthetic lethality to predict the response to cancer treatment. *Nat. Commun.* 9, 2546. doi:10.1038/s41467-018-04647-1

Li, Z., Liu, Z., Huang, J., Tang, G., Duan, Y., Zhang, Z., et al. (2019). Mv-gcn: Multi-view graph convolutional networks for link prediction. *IEEE Access* 7, 176317–176328. doi:10.1109/ACCESS.2019.2957306

Lin, C.-H., and Lichtarge, O. (2021). Using interpretable deep learning to model cancer dependencies. *Bioinformatics* 37, 2675–2681. doi:10.1093/bioinformatics/btab137

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* 15, 290–298. doi:10.1038/nmeth.4627

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. Neural Netw.* 20, 61–80. doi:10.1109/tnn.2008.2005605

Wang, L., Li, X., Zhang, L., and Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer* 17, 513. doi:10.1186/s12885-017-3500-5

Wang, W., Malyutina, A., Pessia, A., Saarela, J., Heckman, C. A., and Tang, J. (2019). Combined gene essentiality scoring improves the prediction of cancer dependency maps. *EBioMedicine* 50, 67–80. doi:10.1016/j.ebiom.2019.10.051

Yu, M. K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J. F., et al. (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell. Syst.* 2, 77–88. doi:10.1016/j.cels.2016.02.003

Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. (2018). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* 8, 3355–3359. doi:10.1038/s41598-018-21622-4