Graph attention network based representation learning for cancer drug response prediction and interpretation

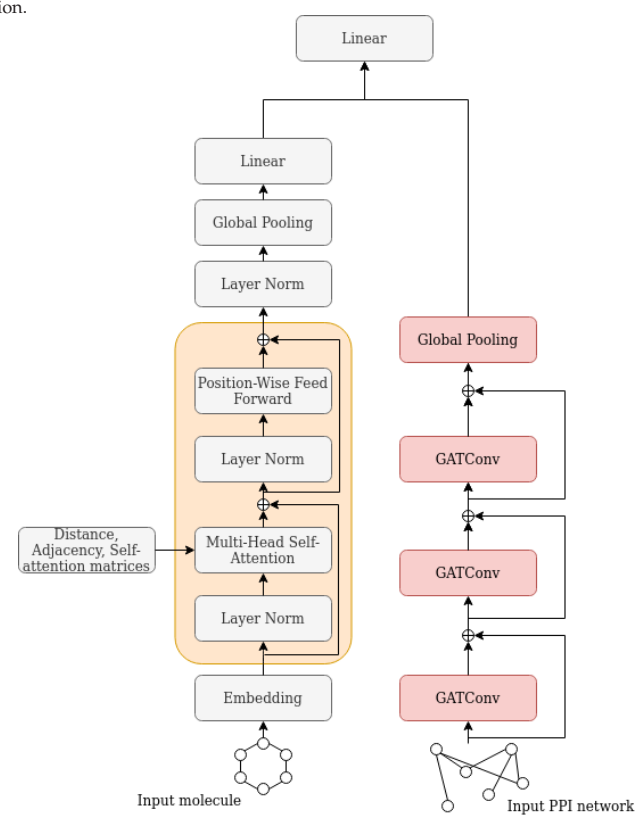
Dionizije Fa*, Fran Supek**, Tomislav Šmuc*
*Rudjer Boskovic Institute
**IRB Barcelona

Introduction

In recent years it has been shown that protein-protein interactions are targetable by drugs which expands the potential pool of drug targets (Goncearenco et al., 2017). In this work we integrate protein-protein interactions and genomic features for modeling cancer drug response, which allows us to discover cell line specific interactions that are most predictive of drug response in cancer cell lines. To that end we construct a multimodal neural network for prediction of drug response based on molecular graph structure and cancer cell line protein-protein interactions. Our model gives insight into drug response related protein-protein interactions, all the while improving on the state of the art on the common benchmark dataset.

PPI interaction model

In our approach, shown in the figure below we have combined molecular transformers (Łukasz Maziarka et al., 2020) for learning drug representations and graph attention neural networks (Veličković et al., 2018) for learning cell line representations, i.e. exploiting protein-protein interaction (PPI) network as a basic graph representation.



Graph attention networks allow us to examine interpretable relevant interactions between nodes of the protein-protein interaction graphs. We show that this approach improves cancer drug response prediction in pharmacogenomic databases, and allows for interpretation of the interactions. This approach also solves the omics integration challenge, as additional gene-wise features such as miRNA expression, methylation, etc., can simply be added as node features of the graph.

To construct cancer cell lines representation for drug response prediction we used the Homo sapiens protein-protein interaction network featurized by expression (TPM), copy number variation (CNV) and mutation data from DepMap. Firstly, from the whole PPI network provided by STRING we kept only experimentally confirmed interactions regardless of their combined interaction score. Secondly, we dropped all the proteins in the PPI network which don't have the associated protein coding genes in the landmark genes of the L1000 assay. Next, we merged such PPI network of landmark genes on DepMap data, where TPM, CNV and mutation data(variant classification) were available. Variant classification was one-hot encoded. To summarize, this procedure produced an undirected graph shared across all cell lines with 813 nodes and 4325 edges.

Pharmacogenomic datasets used were NCI-60, GDSC and CTRP. For CTRP and GDSC data we used recalculated IC50 values, provided by PharmacoGX R-package, version 2.1.10 (Smirnov et al., 2016). Cancer cell lines omics data were extracted from DepMap, 20q3 release. List of landmarkgenes was obtained from the L1000 assay. "Homo sapiens" protein-protein interaction network was downloaded from the STRING database, version 11.0b

Funding

This work has been fully supported by the "Research Cooperability" Program of the Croatian Science Foundation funded by the European Union from the European Social Fund under the Operational Programme Efficient Human Resources 2014-2020. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Croatian Science Foundation, Ministry of Science and Education and European Commission.

Results

We compared our model trained to predict pIC50 values to the state of the art models, which were trained on the GDSC2 dataset. To our knowledge, our model outperforms the state of the art model DeepCDR (Liu et al., 2020).

Model	PPI multimodal	DeepCDR	GraphDRP	tCNNs
RMSE	**1.009±0.005**	1.058±0.006	1.091	1.782±0.006
Pearson correlation	**0.931±0.001**	0.923±0.006	0.929	0.885±0.008

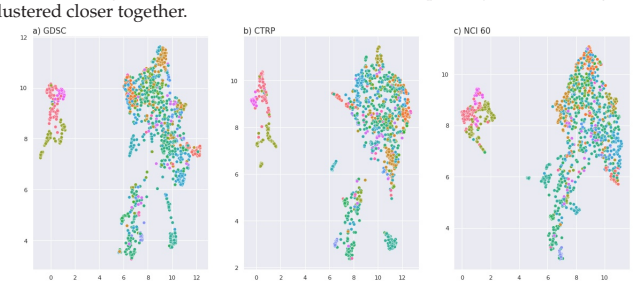
We have also evaluated the regression model on different experimental settings: predicting response of unknown cell lines (blind cells), predicting response of unknown drugs (blind drugs), and predicting response where both drugs and cell lines are unknown (double blind).

Experimental setting	Blind cells	Blind drugs	Double blind
RMSE	**1.548**	2.998	3.098
$\mathbf{R^2}$	**0.703**	-0.323	-0.372

While in the regression settings the predicted values for blind cells show correlation to true values, in other settings the models performed worse than a horizontal line would. From these results it is also obvious that a random test set, which is often evaluated on the benchmark GDSC dataset, does not carry meaningful information on the usefulness of these models in practical settings, i.e. precision medicine and in discovery of new therapeutics.

Interpretation

For intepretation of the attention coefficients we trained models to predict binary response of the cell line to a drug, i.e. is a cell line sensitive or resistant at a threshold concentration of 1 µM. Attention coefficients learned by the graph attention layers assign importance to the edges of the PPI network. Since the graph structure is fixed, importance is directly related to the genomic features used. To visualize these attention coefficients we performed UMAP dimensionality reduction (n_neighbors=10, min_dist=0.1, learning_rate=0.1) on the attention coefficients of the cell lines which are colored by primary disease classification according to DepMap, e.g. breast cancer, in the figure below. Since cell line representations are learned by minimizing the cost function over the available drugs, we would expect that the learned coefficients would be a function of the sampling of the chemical and cell line space. However, across three datasets (NCI, CTRP, GDSC), even with different chemical and cell line spaces we observe that the clustering of the interactions is preserved across datasets, e.g. blood cancers are separated in an isolated cluster across all datasets, since these cancers share a large number of interactions across their primary disease. Cell lines that share a larger number of interactions across their primary diseases are generally clustered closer together.



To further investigate drug response related protein-protein interactions, we group cell lines by primary disease and subtype, then calculate how often the top 10 differential interactions appear in each of the groups, cell lines by primary disease, by subtype, and by primary disease and subtype. We consider models with 1 GATConv layer, trained on GDSC and CTRP datasets, which show the biggest overlap between the interpreted coefficients, i.e. protein-protein interactions. Only overlapping interactions between the datasets were considered. In colorectal adenocarcinoma, we found GLI2-USP7 interaction, which has been shown to be a promising target in cancer (Bryant et al., 2014). USP7 is a deubiquitylating enzyme that activates the hedgehog signaling pathway, which is associated with multiple cancers, including colorectal cancer. Inhibition of CHEK1 shows potential as a therapy in blood cancers (Bryant et al., 2014). We have also found that ERBB3 is an important interaction hub in laryngeal squamous cell carcinoma. It has been shown that ERBB3 signaling can influence drug response in head and neck squamous cell carcinoma and that it could potentially be a therapuetic target (Erjala et al., 2006).

References

Christopher Bryant, Kirsten Scriven, and Andrew J Massey. Inhibition of the checkpoint kinase chk1 induces dna damage and cell death in human leukemia and lymphoma cells. *Molecular cancer*, 13(1):1–12, 2014.

Kaisa Erjala, Maria Sundvall, Teemu T Junttila, Na Zhang, Mika Savisalo, Pekka Mali, Jarmo Kulmala, Jaakko Pulkkinen, Reidar Grenman, and Klaus Elenius. Signaling via erbb2 and erbb3 associates with resistance and epidermal growth factor receptor (egfr) amplification with sensitivity to egfr inhibitor gefitinib in head and neck squamous cell carcinoma cells. *Clinical Cancer Research*, 12(13):4103–4111, 2006.

Alexander Goncearenco, Minghui Li, Franco L Simonetti, Benjamin A Shoemaker, and Anna R Panchenko. Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. pages 221–236, 2017.

Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2):i911–i918, 2020.

Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena MA Gendoo, Patrick Grossmann, et al. Pharmacogx: an r package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, 2016.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. 2018.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. 2020.