BDKANN - Biological Domain Knowledge-based Artificial Neural Network for drug response prediction

Oliver Snow
School of Computing Science
Simon Fraser University
Vancouver, BC
oliver_snow@sfu.ca

Hossein Sharifi Noghabi
School of Computing Science
Simon Fraser University
Vancouver, BC
hsharifi@sfu.ca

Jialin Lu
School of Computing Science
Simon Fraser University
Vancouver, BC
jialin_lu@sfu.ca

Olga Zolotareva
Faculty of Technology and Center for Biotechnology
Bielefeld University
Germany
ozolotareva@techfak.uni-bielefeld.de

Mark Lee
School of Computing Science
Simon Fraser University
Vancouver, BC
mark_lee_3@sfu.ca

Martin Ester
School of Computing Science
Simon Fraser University
Vancouver, BC
ester@sfu.ca

1 Introduction and related work

The ultimate goal of precision oncology is to provide cancer patients with the most effective therapies given their genomic features. One of the major challenges to achieve this goal is to build an accurate computational model to predict drug response given the genomic data of a patient [1, 2]. Unfortunately, current clinical datasets with drug response labels are not large enough to train an accurate model while gathering more data is expensive and time consuming. In addition, it is impossible to treat a patient with multiple drugs independently to separately study their effects. However, numerous pre-clinical datasets have recently been created based on human cancer cell lines and patient-derived xenografts [3–6]. These pre-clinical datasets are larger than clinical datasets and are screened with tens or hundreds of drugs which makes them useful resources for training a computational model.
For drug response prediction on pre-clinical datasets, previous studies have proposed different methods such as regression [7], kernel learning [8], and deep neural networks (DNNs) [9] to map the genomic data–often gene expression–to a measure of response such as the IC50 or area above the dose-response curve (AAC) [10]. Although these methods have shown promising results, they have not considered prior biological knowledge in their models. These methods used genes as the features of their models in isolation, however, genes do not operate in isolation, but rather work in biological networks or pathways. In the context of drug response prediction, genes and the corresponding proteins work together to form protein complexes, which then operate in different pathways, and eventually the function of a pathway is disrupted by a drug. Therefore, our hypothesis is that incorporating this domain expert knowledge in a computational model should improve the accuracy of drug response prediction by avoiding over-fitting on a rather small dataset.
The other major challenge is that despite the relatively accurate predictions of existing models, most are so called 'black boxes' in which the reasoning behind the prediction is unknown. This is particularly undesirable in the context of cancer therapy as clinicians require a rational explanation

of why a drug is expected to work for a given patient, however, current models do not provide this explanation. This motivates the use of domain expert knowledge as the structure of the model so that predictions can then be traced back to specific nodes in the network that represent real biological entities such as pathways, complexes, or specific genes.

Different studies have incorporated various biological knowledge into DNNs. For example, Ma et al. [11], developed DCell to use gene ontology (GO) terms for cell growth prediction. Gaudelet et al. [12], used pathways and protein complex data for disease diagnosis. Hao et al. [13], used pathway information to predict survival of cancer patients. Finally, Kang et al. [14], used gene regulatory network knowledge to predict the response to different therapies. Together, these previous works show the potential of biological domain expert knowledge for defining the architecture of DNNs, both for improving prediction accuracy and for creating interpretable models with functional explanations for the predictions made. Drawing on the strengths of these past works we build an interpretable DNN based on biological domain expert knowledge for the task of drug response prediction.

2 The BDKANN method

BDKANN overview. In this paper, we employ domain expert knowledge from genes to protein complexes to pathways and finally to drugs as layers of a DNN. The structure of BDKANN is as follows: 1) in the gene layer of this network, a node represents a gene for which the expression data is available. 2) In the protein complex layer, a node represents the complex that genes in the previous layer can form. 3) In the pathway layer, a node represents a pathway that a protein complex (or multiple complexes) in the previous layer is (are) a part of. Lastly, 4) in the drug layer, a node represents a drug that targets a given pathway(s) in the previous layer. We connect these layers in the neural network using the biological knowledge obtained from Reactome, an open source, peer-reviewed, and manually curated database containing pathway, complex, and even drug-target information [15]. The advantage of this approach is that the DNN model predicts response to multiple drugs using prior biological knowledge. However, it cannot discover new associations between different layers of information in the context of drug response. Therefore, we add edges to make the network fully connected but also add a regularization on these additional edges, which are not supported by the biological domain expert knowledge, to force their weights towards zero. Thus, the proposed method exploits the biological knowledge and also is capable of discovering new associations if an edge which is not supported by the biological domain expert knowledge "survives" the regularization.

Problem definition. Given gene expression data $X \in M \times N$, where M is the number of cell lines and N is the number of genes, and AAC labels $Y \in M \times K$, where K is the number of screened drugs, the objective is to learn a model $f(X)$ to predict Y. In this paper, f is a DNN parameterized by Θ that receives gene expression data from cell lines and predicts the response to K drugs such that it minimizes the total Mean Squared Error. The goal is to incorporate the domain expert knowledge in designing the structure and connections of different layers in the DNN implementing f, and at the same time explore new connections and associations between different layers beyond the available domain expert knowledge. For example, for any two consecutive layers such as $L_i$and L_j, the connections (weights of the neural network) are $\theta_E^{L_{ij}}$ and $\theta_D^{L_{ij}}$ such that $\theta_E^{L_{ij}} \cup \theta_D^{L_{ij}} = \Theta^{L_{ij}}$, where $\theta_E^{L_{ij}}$ are connections supported by the domain expert knowledge, and $\theta_D^{L_{ij}}$ are the additional connections (i.e. unknown based on available domain expert knowledge) for discovering new associations between L_i and L_j. To avoid overparameterization of f, $\theta_D^{L_{ij}}$ will be regularized such that only connections with direct impact on the predictions survive by forcing the others without any impact to become zero.

BDKANN architecture. The architecture of BDKANN is defined by the acquired biological knowledge from the cell-line data and the network data from Reactome. The general architecture can be seen in figure 1, where the input layer consists of genes, the first hidden layer consists of protein complexes, the second hidden layer consists of pathways, and the output layer corresponds to the drugs for which we predict AAC. The dimension of the input layer is 7,436, defined by the set of genes whose expression was measured in all studied datasets and reduced to the genes for which there is complex information. The dimension of the first hidden layer is 11,836, corresponding to the number of protein complexes that the input genes map to. The dimension of the second hidden layer is 1,262, corresponding to the number of pathways that the previous complexes map to. The final output layer dimension is 6, corresponding to the number of overlapping drugs between the datasets for which we have studied and choosing those drugs with the least NA values.

BDKANN cost function. A key part of BDKANN is the discovery regularization. The goal of this regularization is to discover new associations between different layers of the network. We employ l_1

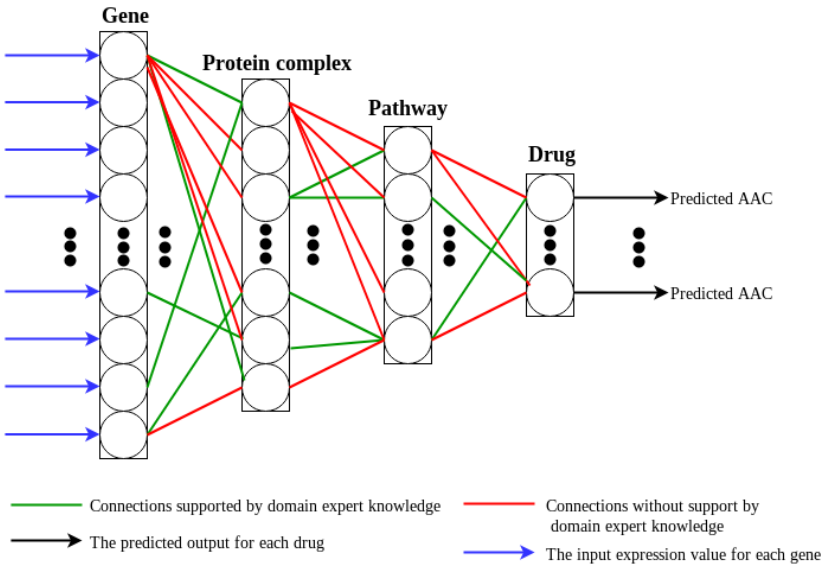


Figure 1: Overview of the structure of BDKANN

regularizations to force the weights corresponding to the edges that are not supported by the biological domain expert knowledge to become zero. The rationale behind regularizing them is that since they are not supported by the domain expert knowledge, we do not want them to exist, on the other hand, we also want to discover new interactions. Therefore, if a weight survives this regularization by the end of the training (has a non-zero value), its corresponding edge was a contributing interaction to predict the drug response output. To train BDKANN in an end-to-end fashion, we employ the following cost function: $J(\Theta) = MSE(f(X), Y) + \beta l_1(\theta_D)$.

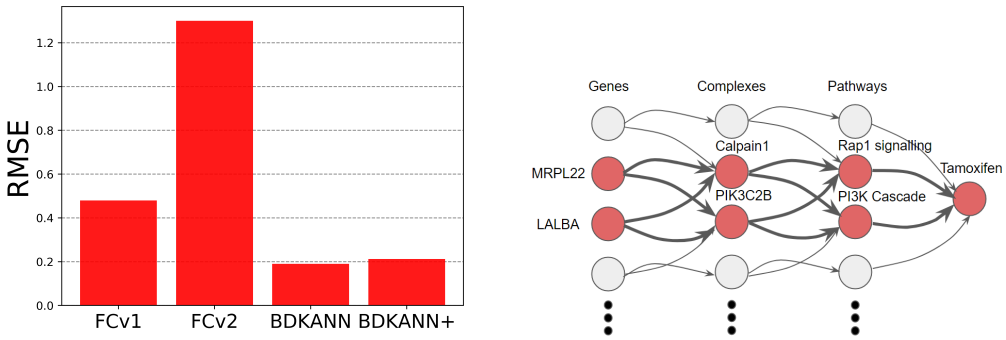
Interpretation. For interpreting the learned BDKANN model, we wish to extract contributing genes, complexes, and pathways that are highly relevant for predicting the AAC of a given drug. To achieve this goal, we leverage Layer-wise Relevance Propagation (LRP) [16], a node attribution method [17] which assigns attribution scores to nodes in a layer for predicting the response to a certain output drug for an input sample. A high attribution score indicates that this node plays an important role in the prediction. Given a drug to interpret, we apply LRP for each node in each layer on the entire test set and take the average value over the test set. By doing this, we obtain a list of averaged attribution scores for the nodes in gene, protein complex, and pathway layers for each drug. Then, we sort this list and select the top k nodes from each layer for which we investigate their biological relevance.

3 Experimental results

Experimental questions. Our goal is to answer two questions:
1) Does BDKANN outperform knowledge-based and DNN baselines in terms of RMSE?
2) Does the biological knowledge-based structure of BDKANN lead to a more interpretable model and are these interpretations informative?

To answer the first question, we compared BDKANN with the discovery connections (denoted as BDKANN+) against three baselines. First, a fully-connected DNN with tuned number of layers and nodes in each layer as hyper-parameters (denoted as FCv1). Second, a fully-connected DNN with the same architecture as the BDKANN network (denoted as FCv2). We used dropout for the regularization of these baselines. Third, a knowledge-based DNN with the same architecture as the BDKANN network but only with the connections supported from the domain expert knowledge (illustrated in green in Figure 1-denoted as BDKANN). We did not use dropout for BDKANN and BDKANN+ because we don't want to remove nodes supported by the domain expert knowledge.

To answer the second question we return the top 10 nodes for each layer based on their average attribution score on the test set. Then, we investigate the biological entities these nodes correspond to and see which are shared across drugs and which are drug-specific. Additionally, we can see if we simply recapitulate the known contributors to cancer or if we identify some high-attribution nodes



(a) Test RMSE of BDKANN models and baselines (b) Example set of high-attribution nodes for Tamoxifen

Figure 2: Obtained results for the prediction performance and the interpretation phase

that can point to new cancer drivers or drug targets.

Datasets. In this work, we utilized two datasets:

1) The Genomics of Drug Sensitivity in Cancer (GDSC) dataset [3], consisting of more than 1000 pan-cancer cell lines, screened with 265 targeted and chemotherapy drugs.

2) The Cancer Therapeutic Response Portal (CTRPv2) dataset [18], consisting of more than 800 pan-cancer cell lines, screened with 481 targeted and chemotherapy drugs.

We employed GDSC dataset for training and CTRPv2 dataset for test and obtained them via PharmacoGx R package [19] and PharmacoDB [10]. These two datasets have 93 drugs in common, however, we focused on 6 of those drugs including Doxorubicin, Tamoxifen, Masitinib, 17-AAG, GDC-0941, and PLX4720 because these drugs had pathway information available in REACTOME and we filtered out drugs with a high fraction of NA values for the outcome.

Experimental design. BDKANN and all of the baselines were trained and tuned for the hyper-parameters using 10-fold cross validation on the GDSC dataset. The best hyper-parameters were selected based on RMSE and each model was re-trained on the entire GDSC dataset using the selected parameters. Finally, the trained model was tested on the CTRPv2 dataset. All of the methods had the same number of genes as the input. The trained BDKANN model with discovery regularization was then passed to the interpretation method as explained above, giving us a list of high-attribution nodes for each drug.

Prediction results. Figure 2a reports the RMSE of the studied baselines and two versions of BDKANN. BDKANN and BDKANN+ achieved lower RMSE on the test data compared to the baselines. We observed that FCv1 and FCv2 over-fitted more than the BDKANN models. The reason is that all of the four models achieved comparable train RMSE (0.01 ± 0.004) but different test errors. We believe that the domain expert knowledge in BDKANN networks reduces over-fitting.

We observed that BDKANN also achieved better performance compared to BDKANN+. We acknowledge that BDKANN+ has more to offer in terms of interpretability because of the additional discovery connections, but on the other hand, it is also more susceptible to over-fitting due to having additional parameters to learn. We believe validating this requires further experiments with different discovery regularizations in the future.

Interpretation results. First, we find the attributions scores are unsurprisingly very sparse, meaning relevant information are concentrated on only a few nodes. Interestingly, the high-attribution nodes for the gene/complex/pathway layers were distinct for different drugs, but with some nodes being shared across many drugs. This suggests that some nodes are generally important for all drugs while other nodes are drug-specific.

Figure 2b presents a case example of the analysis we can perform based on the attribution scores of the network. For the purpose of this work we focus only on a single drug, Tamoxifen, primarily prescribed to treat breast cancer and we only discuss one of the most notable nodes for each layer and their relevance to cancer. Starting at the gene layer, we see that the gene LALBA, is uniquely important for predicting Tamoxifen response and not for the other drugs and indeed has been found to be highly expressed in some advanced breast cancers [20].

Moving to the protein complex layer, a high-attribution complex for predicting Tamoxifen response is PIK3C2B complexed with calcium and magnesium ions. PIK3C2B codes for a protein that is part of the PI3K/AKT/mTOR pathway, a pathway responsible for regulating cell growth and linked to a

number of cancers [21].
Indeed we see that this pathway is one of the ten high-attribution nodes in the pathway layer for Tamoxifen as well as three other drugs, indicating the general importance of this node for cancer drug response prediction. Other nodes corresponding to the gene MRPL22, the complex Calpain1, and the pathway Rap1 were common to several drugs suggesting that they represent more general cancer mechanisms [22–25]. Interestingly, for Tamoxifen, none of the 10 high-attribution pathways had connections based on biological domain expert knowledge to Tamoxifen, meaning those connections survived regularization, indicating the potential of BDKANN+ to discover new knowledge.

4 Conclusion and future directions

Prediction error results show that BDKANN is capable of having comparable or better predictions than baseline methods in a multi-drug setting. Furthermore, the network is interpretable given that its structure is based on biological domain expert knowledge and we show that these interpretations can highlight both general and drug-specific mechanisms. Future work will further optimize the predictive capability of BDKANN, with the inclusion of more drugs and biological knowledge, as well as refine the workflow for analyzing the interpretation results and their relation to cancer treatment.

References

[1] Mehreen Ali and Tero Aittokallio. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical reviews*, 11(1):31–39, 2019.

[2] John Marquart, Emerson Y Chen, and Vinay Prasad. Estimation of the percentage of us patients with cancer who benefit from genome-driven oncology. *JAMA oncology*, 2018.

[3] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[4] Hui Gao, Joshua M Korn, Stéphane Ferretti, John E Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, Christian Schnell, Guizhi Yang, Yun Zhang, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature medicine*, 21(11):1318, 2015.

[5] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

[6] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[7] Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, 15(3):R47, 2014.

[8] Xiao He, Lukas Folkman, Karsten Borgwardt, and Jonathan Wren. Kernelized rank learning for personalized drug recommendation. *Bioinformatics*, 1:9, 2018.

[9] Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz318.

[10] Petr Smirnov, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-alawi, Zhaleh Safikhani, and Benjamin Haibe-Kains. Pharmacodb: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic acids research*, 46(D1):D994–D1002, 2017.

[11] Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15(4):290, 2018.

[12] Thomas Gaudelet, Noel Malod-Dognin, Jon Sanchez-Valle, Vera Pancaldi, Alfonso Valencia, and Natasa Przulj. Unveiling new disease, pathway, and gene associations via multi-scale neural networks. *arXiv preprint arXiv:1901.10005*, 2019.

[13] Jie Hao, Youngsoon Kim, Tae-Kyung Kim, and Mingon Kang. Pasnet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC bioinformatics*, 19(1):510, 2018.

[14] Tianyu Kang, Wei Ding, Luoyan Zhang, Daniel Ziemek, and Kourosh Zarringhalam. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. *BMC bioinformatics*, 18(1):565, 2017.

[15] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2017.

[16] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[17] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks! *CoRR*, abs/1808.04260, 2018. URL `http://arxiv.org/abs/1808.04260`.

[18] Matthew G Rees, Brinton Seashore-Ludlow, Jaime H Cheah, Drew J Adams, Edmund V Price, Shubhroz Gill, Sarah Javaid, Matthew E Coletti, Victor L Jones, Nicole E Bodycombe, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology*, 12(2):109, 2016.

[19] 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, 2015.

[20] Vincent K Tuohy. Retired self-proteins as vaccine targets for primary immunoprevention of adult-onset cancers. *Expert review of vaccines*, 13(12):1447–1462, 2014.

[21] Yiqun Zhang, Patrick Kwok-Shing Ng, Melanie Kucherlapati, Fengju Chen, Yuexin Liu, Yiu Huen Tsang, Guillermo de Velasco, Kang Jin Jeong, Rehan Akbani, Angela Hadjipanayis, et al. A pan-cancer proteogenomic atlas of pi3k/akt/mtor pathway alterations. *Cancer cell*, 31 (6):820–832, 2017.

[22] Zhenzhen Huang, Huilong Duan, and Haomin Li. Identification of gene expression pattern related to breast cancer survival using integrated tcga datasets and genomic tools. *BioMed research international*, 2015, 2015.

[23] Federica Sotgia, Marco Fiorillo, and Michael P Lisanti. Mitochondrial markers predict recurrence, metastasis and tamoxifen-resistance in breast cancer patients: Early detection of treatment failure with companion diagnostics. *Oncotarget*, 8(40):68730, 2017.

[24] Shadia M Al-Bahlani, Ruqaya M Al-Rashdi, Shiyam Kumar, Shadia S Al-Sinawi, Maiya A Al-Bahri, and Asem A Shalaby. Calpain-1 expression in triple-negative breast cancer: a potential prognostic factor independent of the proliferative/apoptotic index. *BioMed research international*, 2017, 2017.

[25] Elaine A McSherry, Kieran Brennan, Lance Hudson, Arnold DK Hill, and Ann M Hopkins. Breast cancer cell migration is regulated through junctional adhesion molecule-a-mediated activation of rap1 gtpase. *Breast Cancer Research*, 13(2):R31, 2011.