Application of Deep Learning Based Drug Sensitivity Prediction Model on a Novel Anticancer Drug

3527

Nischal Mahaveer Chand, Vivek K. Vishnudas, Michael A. Kiebish, Anjan Thakurta, Niven R. Narain, Stephane Gesta, and Gregory M. Miller

[1]BPGbio, Framingham, MA USA

ABSTRACT

- Understanding drug sensitivity is a critical steps in selection of a candidate drug compound and involves repetitive testing on cell-lines that can take up significant time and effort. *In-silico* methods for predicting drug sensitivity would enable faster and cheaper testing and help guide molecule optimization.
- DeepDSC [1] proposed by Li Min, et al. is a deep learning model that integrates cell-line RNAseq and compound fingerprints [2] to predict the half-maximal inhibitory concentration (IC50) of the pair.
- We retrained DeepDSC and evaluated the model on a novel anti-cancer drug – BRG399. We observed poor generalizability of DeepDSC and examined potential limitations, including lack of compound diversity in the GDSC2 [5] sensitivity dataset and inflexibility of compound fingerprints for neural network models.
- To address these limitations, we trained a Graph Convolutional Network (GCN) on NCI60 sensitivity data [3] and observed a reduction in RMSE error by 80.4% and an improvement in R^2 by 48.9%.

DEEPDSC

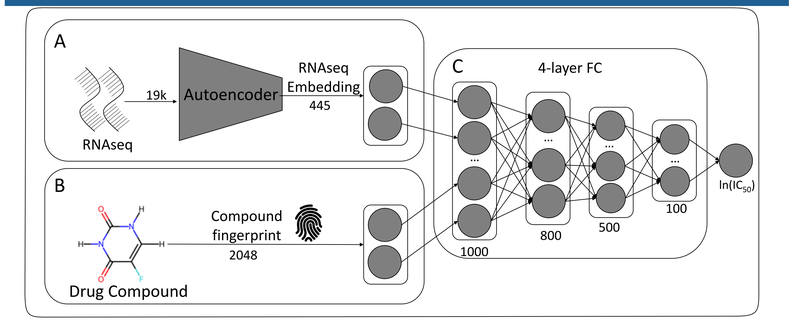


FIGURE 1. Schematic diagram of DeepDSC [1]. (A) First, an autoencoder was trained on RNAseq data for 1450 cell-lines characterized by CCLE [4]. (B) Next, extended circular fingerprints (ECFP) [2] of length 2048 were computed for all compound in the GDSC2 [5] dataset using RDKit. (C) Finally, a 4-layer fully-connected network (FC) was trained to predict the natural log (ln) transformed IC50 in μM of each drug – cell-line pair. Models were trained with 10% dropout and ReLU activation function for hidden layers. The RNAseq embeddings of the cell-line and drug fingerprints were concatenated and passed as input to the network. The 4-layer FC had 3.5M learnable parameters, and the models were trained using TensorFlow.

DEEPDSC RESULTS

DeepDSC Model	Training Set		Heldout Set		Test Set: BRG399	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Baseline	0.42 ± 0.028	0.98 ± 0.003	1.01 ± 0.014	0.87 ± 0.004	1.63 ± 0.065	-5.06 ± 0.4
+ weight decay	0.42 ± 0.034	0.98 ± 0.004	1.01 ± 0.009	0.87 ± 0.003	1.48 ± 0.102	-3.78 ± 0.58
+ L2 regularization	0.63 ± 0.08	0.95 ± 0.014	1.03 ± 0.015	0.86 ± 0.004	1.84 ± 0.205	-7.14 ± 1.996

TABLE 1. Model performance on training (80%), heldout (20%), and BRG399 (test set) (n=101). We retrain DeepDSC as described in [1] with no modification (baseline). Next, we chose to optimize the weight decay of the optimizer or the L2 regularization constant for kernel and bias regularization to improve generalizability of DeepDSC on BRG399. Metrics are reported as average and standard deviation from 5 reruns of the training pipeline with a different training and heldout splits. Hyperparameter were optimized on the training set using 5-fold cross validation (CV). The model with lowest CV RMSE was chosen as best and retrained on the full training set. Finally, the best model was evaluated on the heldout set and on BRG399. Underlined in the lowest RMSE on BRG399.

DEEPDSC MODEL EVALUATION

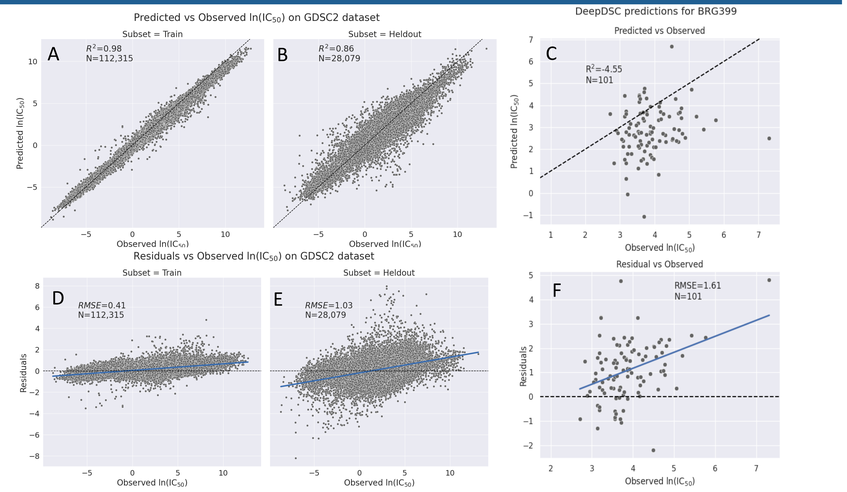


FIGURE 2. Performance of best DeepDSC w/ weight decay model (model with lowest heldout set RMSE across the 5 repeats). Predicted vs Observed ln(IC_{50}) on (A) training set (B) heldout set (C) BRG399 (test set). For BRG399, we observed that the model overestimated the sensitivity of the compound. Next, we observed a non-random scatter of residuals in (D) training set (E) heldout set (F) BRG399 (test set), as signified by the blue regression lines. These results indicated that the model was unable to explain all the variance in the dataset with the given input data (RNAseq and compound fingerprints).

GRAPH CONVOLUTIONAL NETWORK

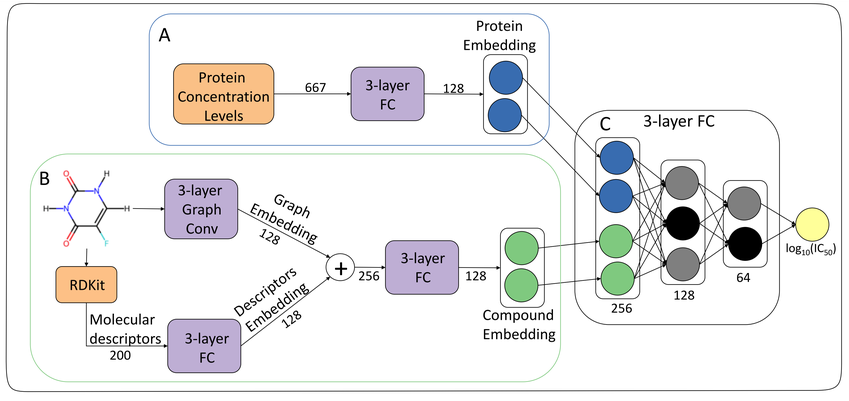
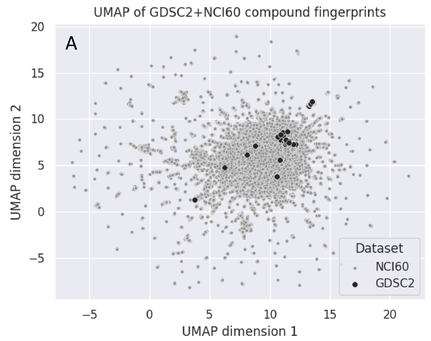


FIGURE 3. Schematic diagram of graph convolutional network (GCN). (A) Cell line proteomics: Proteomics data were obtained from NCI60 website [3] for 667 proteins. Protein concentration levels were passed through a 3-layer FC to obtain the protein embeddings for a cell line. (B) Compounds: Encoded as graphs with atoms as nodes and bonds as edges: node embeddings incorporate atomic information for each atom while bond lengths were computed with RDKit and encoded for each bond. The graph was then passed through three layers of GraphConv [5] layer to obtain a graph embedding for the compound. Additionally, molecular descriptors were compound using RDKit and passed through a 3-layer fully connected network to obtain the descriptor embedding. These embeddings were concatenated for each compound and further passed through a 3-layer FC to obtain the compound embedding. (C) Finally, the compound embedding and protein embedding were concatenated and passed through a 3-layer FC with a single neuron with linear activation on the final layer to obtain the log (base 10) transformed IC50 in M. The full model had 880k learnable parameters, and the models were trained using PyTorch and PyTorch Geometric.

GDSC2 AND NCI60 COMPOUND DIVERSITY



B

Dataset	No. of compounds	U1 Variance	U2 Variance
GDSC2	234	1.24	2.70
NCI60	49,582	15.03	13.66

C

Dataset	No. of unique compounds	No. of unique cell lines	No. of available IC50s
GDSC2	232	690	140,394
NCI60	34,432	60	1,859,319

FIGURE 4. (A) Unified Manifold Approximation and Projection (UMAP) was applied on 234 GDSC2 compounds and 49,582 NCI60 compounds to project their 2048 length fingerprints in 2-dimensions. NCI60 compounds show a larger variance compared to GDSC2 compounds as shown by the larger footprint of NCI60 over GDSC2. (B) Variance of UMAP dimension 1 (U1) and UMAP dimension 2 (U2) in GDSC2 and NCI60 datasets. NCI60 shows a larger variance in both dimensions. (C) Data characteristics after cleaning compounds for GDSC2 and NCI60 sensitivity datasets.

GRAPH CONVOLUTIONAL NETWORK RESULTS

Dataset	Training Set		Heldout Set		Test Set: BRG399	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
GDSC2	1.41 ± 0.009	0.74 ± 0.004	1.44 ± 0.01	0.73 ± 0.002	1.36 ± 0.435	-3.71 ± 3.084
NCI60	0.3 ± 0.005	0.83 ± 0.005	0.3 ± 0.004	0.82 ± 0.005	0.29 ± 0.026	-1.93 ± 0.501

TABLE 2. GCN model performance trained on GDSC2 and NCI60 datasets. Metrics are reported as average and standard deviation from 5 reruns of the training pipeline with a different training and heldout splits. Hyperparameter were optimized on the training set using 5-fold cross validation (CV). The model with lowest CV RMSE was chosen as best and retrained on the full training set. Finally, the best model was evaluated on the heldout set and on BRG399. Underlined in the lowest RMSE on BRG399. We observe an improvement in RMSE by 63.5% when using GNN on NCI60 dataset.

CONCLUSION

- DeepDSC [1], based on compound fingerprints and cell-line RNAseq, showed poor generalization on heldout set and on BRG399 – a novel anti-cancer compound, not present in the training set.
- We suspected the lack of compound diversity and inflexibility of compound fingerprinting as limitations for DeepDSC.
- NCI60 [3] showed a larger compound diversity compared to GDSC2 [5].
- We trained a Graph Convolutional Network based on [6] with NCI60 sensitivity data and observed improved generalizability to BRG399, with a reduction in RMSE error by 80.4% and an improvement in R^2 by 48.9%.

REFERENCES

1. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines Li Min, Wang Yake, Zheng Ruiqing, Shi Xinghua, Li Yaohang, Wu Fang-Xiang, Wang Jianxin IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, Vol 18 (2), 575-582 DOI: 10.1109/TCBB.2019.2919581
2. Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." Journal of chemical information and modeling 50.5 (2010): 742-754.
3. Available from: http://dtp.cancer.gov.
4. Barretina, Jordi, et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." Nature 483.7391 (2012): 603-607.
5. Yang, Wanjuan, et al. "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells." Nucleic acids research 41.D1 (2012): D955-D961
6. Morris, Christopher, et al. "Weisfeiler and leman go neural: Higher-order graph neural networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.