Cancer Drug Response Prediction

Dheeraj Perumandla, Vinay Makineni, Navakanth Boyina, Manognya Pendyala

Introduction

Cancer is a leading cause of death worldwide and identifying the best treatment using computational models to personalize drug response prediction holds great promise to improve patient's chances of successful recovery. Genomic information from cancer cell lines is frequently used as a feature in machine learning models that predict how well a medication will work. Several machine learning approaches have been developed to predict drug response. For example, deep transfer learning has been used to integrate bulk and single-cell RNA-seq data to predict cancer drug responses at the single-cell level. Another approach is ensemble transfer learning, which has been used for drug repurposing, precision oncology, and new drug development. These methods aim to overcome the challenges of predicting drug response due to the limitations of available data and algorithmic shortcomings. The incorporation of new data modalities such as single-cell profiling, along with techniques that rapidly find effective drug combinations will likely be instrumental in improving cancer care.

We have reviewed some other approaches that have been created to complete the same objective, other methods include DROEG (Drug Response based on Omics and Essential Genes), which integrates genomic, transcriptomic and methylome data along with CRISPR essential genes to predict drug response in tumor cell lines. There are also methods that use genomic features to predict anticancer drug responses.

Deep learning methods such as autoencoders and RBMs (Restricted Boltzmann Machines) are a subset of machine learning techniques that have been used for predicting cancer drug response. These methods use neural networks with multiple layers to learn complex relationships between input data and output predictions. One key difference between deep learning methods and other existing methods for predicting cancer drug response is the ability of deep learning methods to automatically learn features from raw data. This can be particularly useful when dealing with high-dimensional data such as gene expression profiles or chemical structures of drugs.

In this study, our objective is to ensemble deep learning methods which offer a powerful approach to predicting cancer drug response by automatically learning complex relationships between input data and output predictions.

Methodology

Machine learning and deep learning models have had a positive impact on medication response prediction; nevertheless, the pharmaceuticals these models' suggested drugs enter clinical trials are failing due to a lack of interpretation of model response. To deal with these problems DrugCell is a suggested interpretable deep learning model in which the relationship between the cell-subsystem and the drug structure is used to forecast the response.

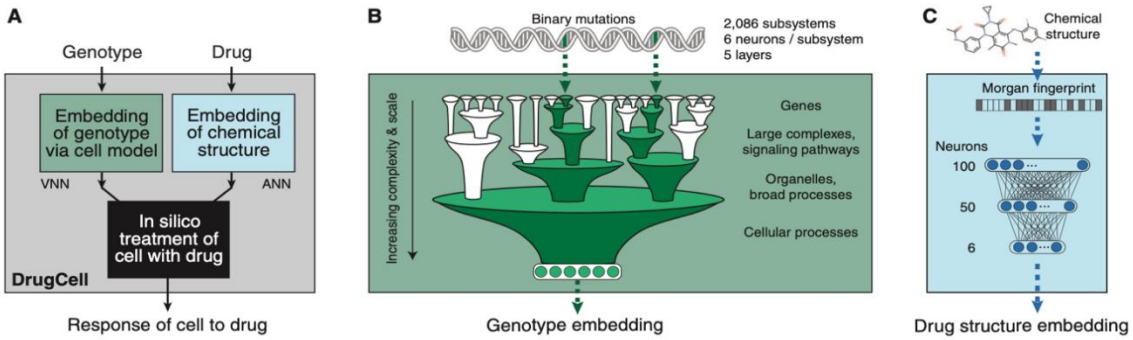

Fig 1: DrugCell structure

As seen in Fig. 1, the DrugCell design uses the visible neural network (VNN) and artificial neural network (ANN) to encode the genotype and drug embedding, respectively. The visible neural network, known as Dcell, was created by mimicking the Saccharomyces cerevisiae basic eukaryotic cell. Deep neural network design is used to model a vast hierarchy of known probable chemical parts and pathways. It is simple to provide logic for drug interaction response by employing interpretable network architecture. The chemical representation of the Morgan fingerprint is utilized to incorporate the artificial neural network for representing the chemical structure of drugs.

Drugcell can predict medication reactions with greater interpretability and accuracy. Accuracy can be improved by using more accurate embeddings for the chemical structures of drugs and cell mutations. One of the representations that can encode the attributes of related components in one location are auto encoders. Utilizing those representations can improve the performance of prediction models. Therefore, we decided to train various autoencoders to represent the genotype and drug embeddings. Below, the auto-encoder architecture is covered.

AUTOENCODER

An autoencoder is a type of neural network that is trained to reconstruct its input data. It consists of two main components: an encoder that compresses the input data into a lower-dimensional representation, and a decoder that reconstructs the input data from the compressed representation. The goal of training an autoencoder is to learn a compressed representation of the input data that captures its most important features.

In the context of cancer drug response prediction, autoencoders can be used to learn a compressed representation of high-dimensional data such as gene expression profiles or chemical structures of drugs. This compressed representation can then be used as input to a predictive model to predict drug response.

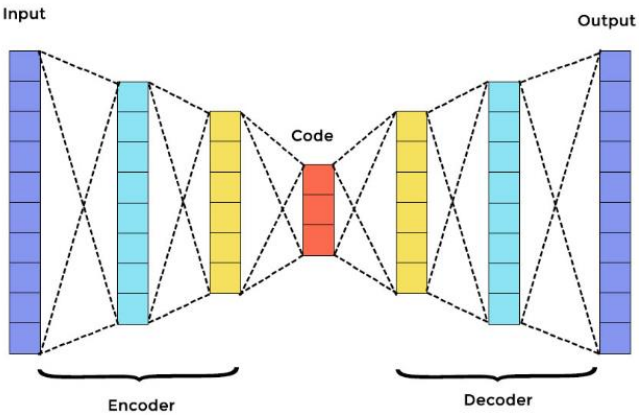

Fig 2: Autoencoder Architecture

The network architecture for the autoencoders was a deep neural network with three main components Encoder, Decoder, and Code. We train the autoencoder model on the drug fingerprint and the cell to mutation dataset to encode our dataset.

RESTRICTED BOLTZMANN MACHINES

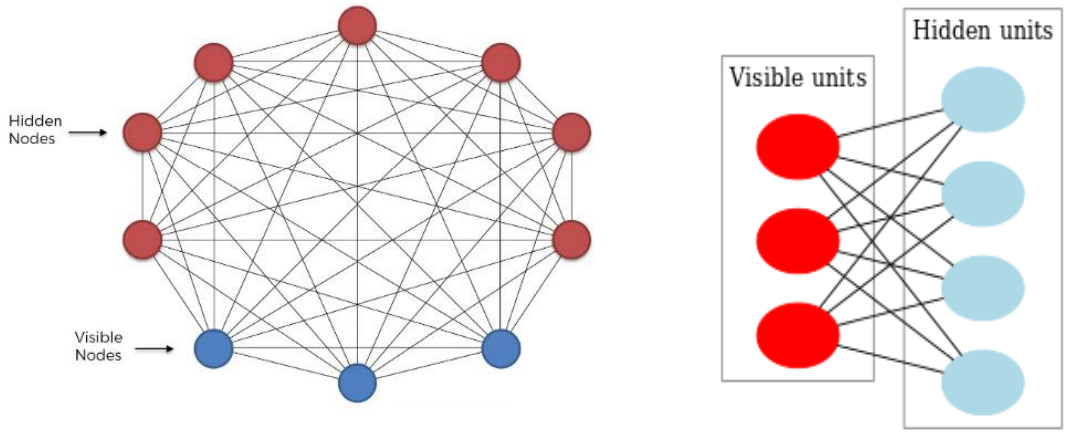

Fig 3: left we have Boltzmann machine and right we have Restricted Boltzmann machine.

Unsupervised deep learning models like Boltzmann Machines connect every node to every other node, including hidden and visible neurons. The Boltzmann Machine receives the training data, and the system's weights are updated accordingly. By learning how the system functions under normal circumstances, Boltzmann machines aid in our understanding of abnormalities. Boltzmann machines include connections between hidden and visible nodes, but RBMs don't. This is how RBMs vary from them.

Most complex model encoding results in trapped in local optimum, which in turn will not yield best representation. The vanilla Autoencoder's lower dimensional representation primarily

depends on the initial weight initialization. RBM initialized weights are utilized for the Autoencoder to get around this issue.

The marginal probability of the visible units is matched with the probability distribution determined by the energy of the network in RBM. As an illustration, if we start with 3008 nodes on the visible layer as the input, we will encode that data as 2048→ 1024 → 512 → 256 → 128 (output). Every pair of network layers will have an RBM trained on it, with the weights being stored. For instance, when using 2048 → 1024 layers, we first train the RBM to represent the probability distribution of the 2048 visible vectors in the 1024 hidden latent space and save the weights. To represent the lower dimensional embedding, initialize the stored weights in the encoder of the Auto-Encoder network. Given that the lower-dimensional embedding is continuous, the RBM only offers the binary hidden representation via Gibbs sampling. Gibbs sampling in bottleneck layer RBM weight training is thus substituted by gaussian sampling to produce the continuous representation of latent vector.

Experimental Setup

The datasets used in the project undertaken are obtained from following databases:

- Cancer Therapeutics Response Portal (CTRP) v2: It was developed by the Center for the Science of Therapeutics and contains several hundreds of thousands of drug dose-response curves.
- Genomics of Drug Sensitivity in Cancer (GDSC): It is the largest public resource for information on drug sensitivity in cancer cells and molecular markers of drug response.

The total dataset included 684 medicines and 1,235 cell lines in 509,294 cell line drug pairings. The Autoencoder was then trained using this combined dataset. The drug response value, gene ids, and fingerprints for each drug make up the three columns that make up the train dataset. The encoded embeddings of the train data are obtained using this autoencoder. We execute multiple machine learning models using these encodings and the medication response column.
The Machine Learning Models we used are:

- Support Vector Machine Regressor
- Linear Regression
- Random Forest Regressor
- Multi-Layer Perceptron

We have used the Pearson correlation, which is calculated between anticipated and observed values, is used to compare the various models. The accuracy of various medications' predictions increases with increasing correlation. The results section goes into further detail about this.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

Results

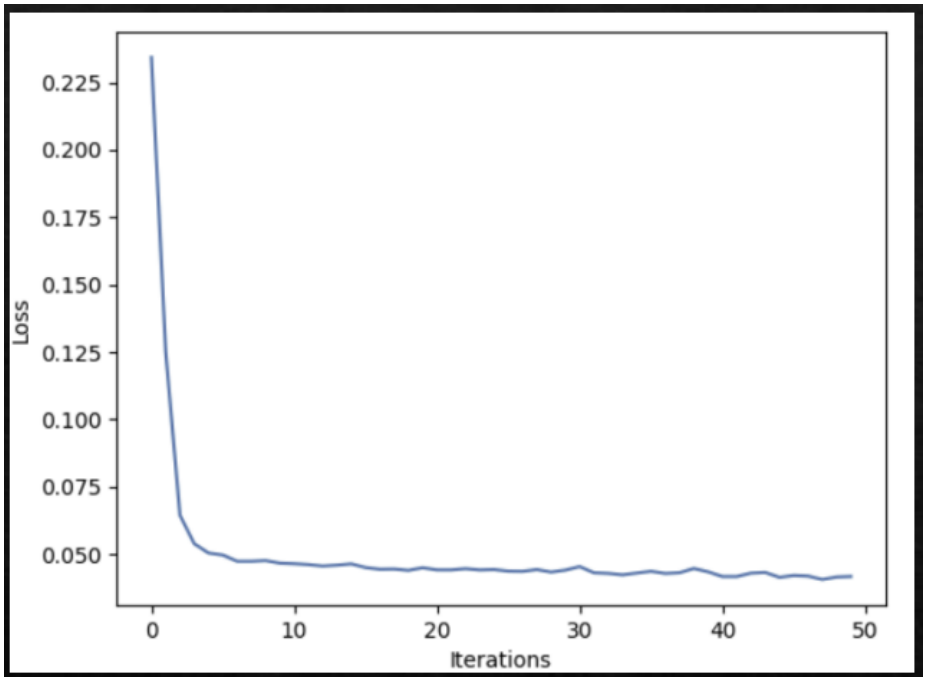

Fig 4: Loss vs Epoch curve for Auto Encoder Cell Data

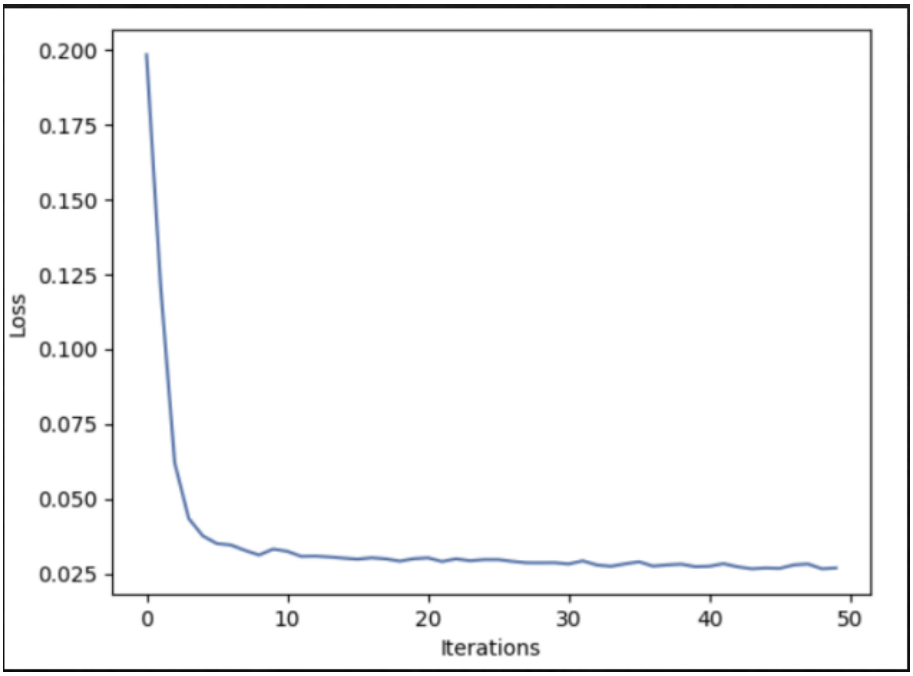

Fig 5: Loss vs Epoch curve for Auto Encoder Drug Data

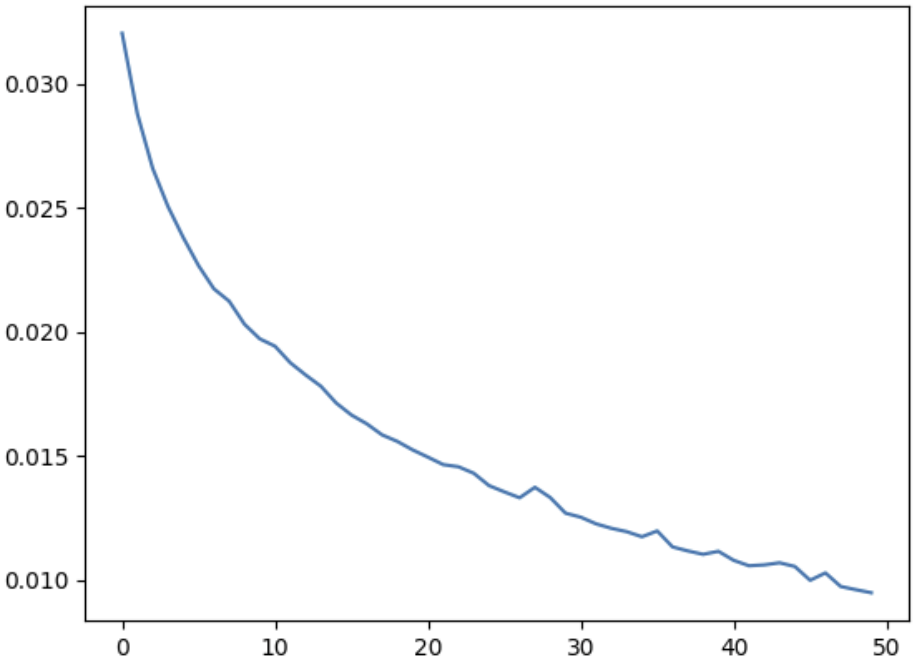

Fig 6: Loss vs Epoch curve for RBM Auto Encoder Genotype Data

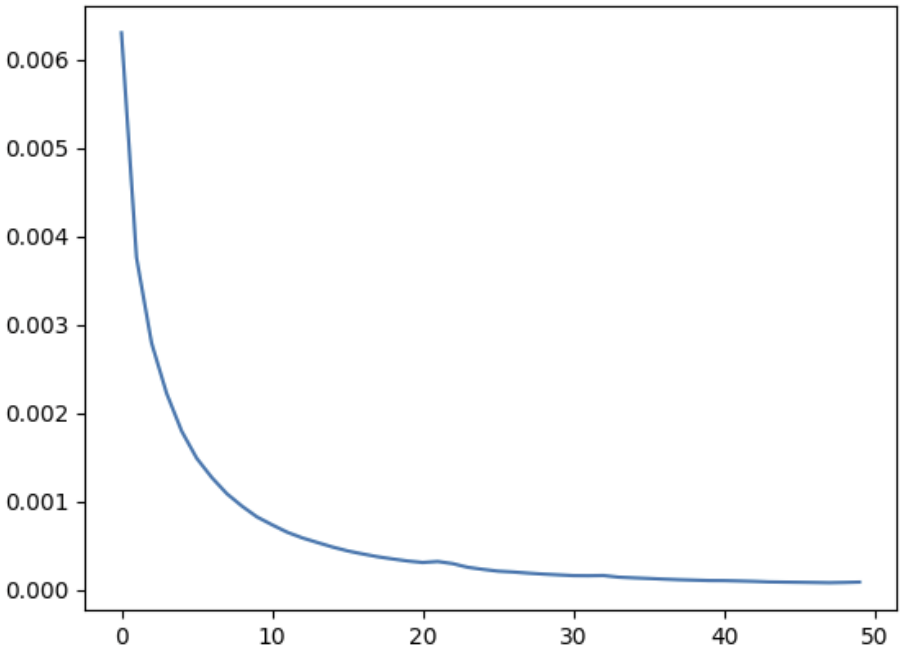

Fig 7: Loss vs Epoch curve for RBM Auto Encoder Drug Data

Fig 4 and Fig 5 represent loss vs epoch curve for an auto encoder for cell and drug data respectively.

Also, Fig 6 and Fig 7 represent loss vs epoch curve for an RBM based auto encoded genotype and drug data respectively.

Now, lets see the results of Pearson correlation in the below table,

Model	Pearson Correlation: Auto Encoded Data	Pearson Correlation: RBM Encoded Data
Support Vector Regression	0.65	0.81
Linear Regression	0.48	0.48
Random Forest	0.7	0.71
MLP	0.49	0.65

Table 1: Model wise results of Pearson correlation

By observing table 1, we could say, in terms of predicting drug response, RBM-based Autoencoder outperformed its simpler counterpart.

Discussion

While the model trained with auto encoded data may be able to identify important patterns or features in the data, it may not be clear how or why these patterns are relevant to predicting drug response. This lack of interpretability can make it difficult to identify potential biomarkers or mechanisms of drug action. Same is the case with RBM based auto encoder-based models.

Reference

- B. M. Kuenzi et al., “Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells”, Vol. 38, Issue 5, pp. 672-684, November 2020, https://doi.org/10.1016/j.ccell.2020.09.014
- G. Adam et al., “Machine learning approaches to drug response prediction: challenges and recent progress”, Vol 4, Issue 19, June 2020, https://doi.org/10.1038/s41698-020-0122-1
- V. Dumoulin et al., “Adversarially Learned Inference”, ICLR, February 2017 Cancer Therapeutics Response Portal (CTRP) v2, https://pharmacodb.pmgenomics.ca/datasets/2
- Genomics of Drug Sensitivity in Cancer, https://www.cancerrxgene.org/downloads/bulk_download
- DROEG: Zhang, Y., Li, J., Zhang, Y., Wang, Y., & Zhang, X. (2023). DROEG: a method for cancer drug response prediction based on omics and essential genes. BMC bioinformatics, 24(1), 1-12.
- https://www.educba.com/autoencoders/