Deep Drug Repurposer

Sarthak Kothari, Noah Demoes, Omair Shafi Ahmed, Forrest Hooton

Summary

In 2017, the international pharmaceutical industry generated over USD 934 billion in revenue and is expected to grow to over $1,170 billion in 2021 [1]. Although people are willing to pay for pharmaceuticals, the cost associated with creating new, breakthrough drugs are costly and often unsuccessful. Only 12% of drugs that are in clinical phase will be approved for the market [4]. Drug development for a single drug costs USD 2.6 billion and 17 years to develop on average [4]. Development regulations drive over 10 percent of this cost, averaging USD 339.3 million dollars. The U.S. Food and Drug Administration (FDA) requires three phases of clinical testing: determining human safety and dosing (phase 1), initial reading of efficacy and side effects (phase 2), and larger clinical trials to determine the safety and efficacy in larger numbers (phase 3) [3]. Bypassing any stage in this process could create cost and time savings for drug development and reduce the burden on consumers.

One method that can reduce drug development cost is drug repurposement. Drug repurposement explores new applications for approved drugs or drugs that have passed at least one of the three FDA approval phases. Since these candidate drugs have already been approved by the FDA and information about pharmacology, formulation, dose, and potential toxicity is already available, it is easier for drug researchers to fulfill phase 1 and phase 2 of clinical trials. This allows a pharmaceutical firm to focus on drug efficacy in more probable drugs and save costs associated with drug development.

Drug repurposing is currently pursued through investigating drug side-effects to determine if the drug could be useful for other conditions. A classic example for this method is *sildenafil*, which was originally developed to treat hypertension. The drug also had side effects that could be used to treat another disease - erectile dysfunction in men. Sildenafil was repurposed to treat erectile dysfunction, and is now well known as Viagra. However, the intense manual analysis required for the drug side-effect investigation strategy limits the procedure for broad drug development. We propose a better way to automatically predict a drug and disease target interaction potential.

Our project aims to develop a deep learning model and web portal for drug repurposing. Specifically, we seek to use information about the molecular structure of drugs and their target receptors to predict whether or not the two inputs react, i.e. a binary classifier. A breakthrough in drug-receptor repurposing could bypass toxicity stages in drug trials, and save tens of millions of dollars for a single effort to treat a disease. There are other active efforts for drug-repurposing, however these efforts use limited datasets that comprise a small fraction of

chemical-target reactions [7]. We plan to couple deep learning techniques with a more robust dataset to develop our drug structure based receptor model. Our primary datasets will be DrugBank for model development, PubChem chefor mical metadata, and the BindingDB database as a validation dataset [9][6][10].

Methods

Stage 1 of our project focused on establishing core components for our analysis. We acquired and transformed data, created molecule embeddings based on Mol2Vec, generated a negative sample set, and developed an external validation set from BindingDB. [5] This report focuses on Stage 2, where we productionalized our drug repurposing model by enhancing our modeling and creating a web application interface. We used DrugBank and blending in BindingDB as our training and validation set, respectively. The DrugBank data is comprised of 6,011 number of unique drugs, 3,738 number of unique genes, and 3,738 unique amino acids. We determined in Stage 1 that the DrugBank dataset was limited in that there were only positive samples that existed. We employed our negative sampling methodology, outlined in our Stage 1 Report, to balance our training set. The BindingDB data set consists of 12,000 samples is comprised of 11,693 number of unique drugs and 830 number of unique aminos.

Molecular Fingerprints

Molecular fingerprints represent molecules in low resolution, machine readable string. The most common fingerprint method is chemical formulas, but other methods exist such as SMILES or Extended-Connectivity Fingerprints (ECFP). The downside to chemical formulas is that they do not capture bond angle or bond type. For example, benzaldehyde's chemical formula is C_6H_5CHO, but this does not capture the bond structure that is central in drug repurposement identification. SMILES are another string representation of a molecule. They capture more information than chemical formulas by including characters that represent the bond type. For example, the SMILES representation of benzaldehyde is C1=CC=C(C=C1)C=O. The third main method utilized ECFP, designed to capture molecular features relevant to molecular activity. [8] ECFP, paired with the Morgan Algorithm, represent molecules based on circular representations of substructures. The Morgan Algorithm, when given a drug and a specified radius, breaks the molecule down into substructures between a single atom and the given radius. It then represents each substructure numerically based on: heavy atom connections, non-hydrogen bonds, atomic number, sign of charge, absolute charge, and number of attached hydrogens. [8] The resulting numerical representations of each substructure are concatenated to form Morgan Fingerprint representation of the specified molecule. It is this numerical representation that we utilize to represent each drug as input into our model. Figure 2 shows an example of glycine being represented as a Morgan Fingerprint [5]. This methodology means that each fingerprint's length will be determined by the complexity of the molecule. More complex molecules will have longer fingerprints while simple molecules will have shorter molecules.

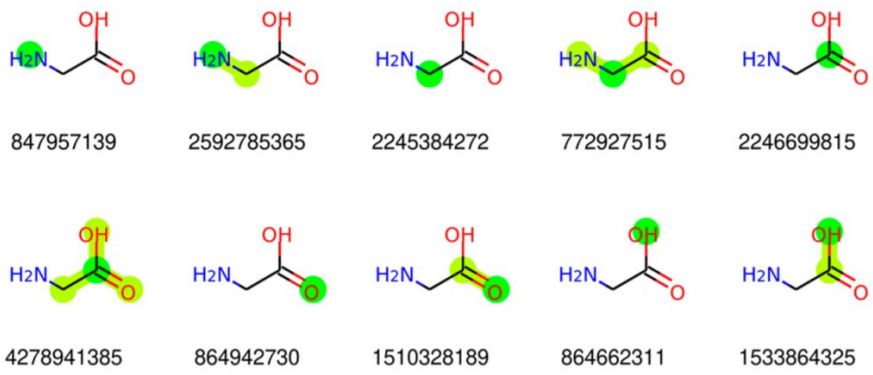


Figure 1: *Representing A Morgan Fingerprint of Glycine.*

Modeling

Stage 1 of this project assessed a variety of models to establish baselines, including a logistic regression model, deep neural network model, a common convolutional neural network model, and converging convolutional neural network model with Mol2Vec embedding inputs. Surprised that our convolutional neural networks compared so poorly with other models, we realized using Mol2Vec as the input to a convolutional neural network decreases model performance. Convolutional neural networks learn higher order covariances and features, and leverage them for downstream learning tasks. By design, Mol2Vec eliminates the covariance necessary for a convolutional neural network to successfully learn. Therefore, our primary efforts for model enhancement comprised of changing a convolutional model input from embeddings to molecular fingerprints.

Model Inputs

Diverging from earlier embedding input methodologies, we chose to use raw fingerprints and encoded genes/amino acids for stage 2 of our project. Per drug molecule, we concatenated all fingerprint substructure codes into a single vector. The molecule vectors were different sizes, as molecules have different numbers of subcomponents. We padded each molecule vector with even padding on the left and right to make all vectors an equal size. However, vector size followed a long tailed distribution before padding, hence most vectors contained primarily padding tokens. We trimmed the vectors by iteratively removing a percentage of vector ends until the number of padding tokens comprised less than half of the dataset. For genes and amino acids, we assigned each unique token a vocabulary integer. Using the generated vocab hash scheme, we transformed each target into a padded vector, again applying the trimming methodology.

Model Structure Optimization

Our model structure was defined by iteratively improvising over the previous dense model. While the previous models largest layer had 200 units that mapped to the 200

dimensional embedded latent space, the input layer of the current setup has 512 filters of 1024 units each, mapping to the raw input. This allows the model to learn patterns and covariances in molecular fingerprints, genes and amino's at a much higher definition. The downstream dense layers after the convolutional layers mostly remained the same. Beginning at 10 epochs, we iteratively increased the epochs in steps of 5 until we reached the ceiling of validation accuracy, stopping at 50. Increasing batch size takes longer to converge and makes it more likely to underfit. However, increasing the batch size decreases the number of steps per epoch and therefore reduces the amount of time taken per epoch. This parameter had to be tuned to find the right tradeoff between reducing the likelihood of underfit and time taken for each iteration. This optimal tradeoff was found at a batch size of 500. The activation functions largely remained the same.

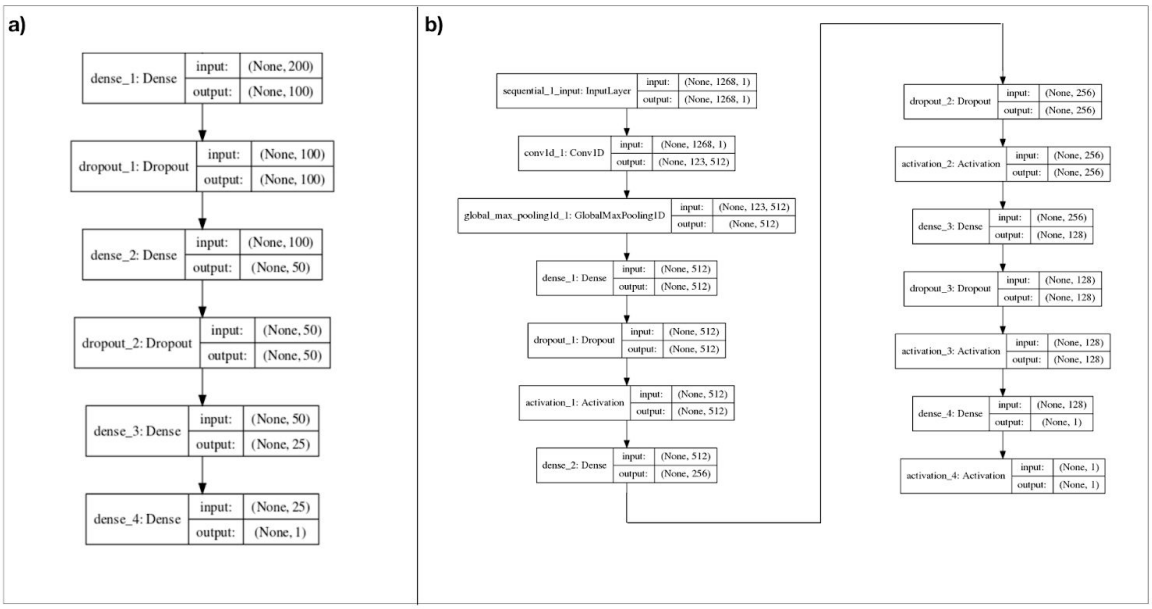


Figure 2: *Comparing previous model vs current model. a) The previous model that ran on the drug fingerprint and gene embeddings. b) The current model that runs convolutions over the raw input.*

Web Application

The intent of building a drug repurposement model is to enable researchers to identify drugs that can be repurposed by reducing the number of candidate drug-gene / drug-amino-acid combinations they must test. We created a web application that gives researchers the ability to input an FDA approved drug in DrugBank and receive a list of all candidate genes in DrugBank and their corresponding match likelihood. The web application consisted of two main entities, the front-end and back-end. Figure A1 diagrams our web application architecture. The web app was hosted on GCP's compute engine. The front end was built using React. We provided users

with a search box where users could search both drugs and targets via their name or chemical composition/target sequence. Once they selected the drug, a POST request was made to our backend, built using Flask in python, with the selected item in order to get the predictions from our model. These predictions were then displayed in a table format. We load our trained keras model and all the data required in the model's input format. For every POST request, we either get a drug or a target. Based on the input query, we use all the combinations of the alternative category and run our model to generate the probability of each interaction. For example, if a drug was sent in the POST request, we predict the likelihood of interaction with all our targets present in the system. These predictions are then converted into appropriate JSON strings and returned to service that made the POST request.

Results

Of all models we examined during our project, we concluded with the convolutional neural network designed in stage 2. Although the gene f1-score was lower than previous models (appendix Fig A2), it had the highest precision of any model. In an applied setting, the most important facet of a model is precision. An incorrect positive causes heavy losses, while an overlooked positive is only a missed opportunity. After discovering the potential of the convolutional neural network with raw fingerprint inputs, we dove deeper into its results.

Modeling Results

The modelling was done in 3 stages, in steps of increasing complexity of the dataset.

1) Modelling Gene - Drug Interactions on DrugBank Data

   Genes constituted our original target inputs for all models. For the convolutional model designed in stage 2, we observe a strong precision of .94 in Fig 3A, although the recall is significantly lower at .49.

2) Modelling Amino Acid - Drug Interactions on DrugBank Data

   While DrugBank has gene - drug interactions, most other open databases map interactions between amino acids and drugs. Biologically, it is the amino acid that interacts with a drug and not the gene itself. As amino acids are translated by RNA, which in turn is transcribed by DNA, which consists of genes, we anticipated a strong correlation between genes and proteins. Furthermore, since each character in the amino sequence can be any 1 in 20 amino acids, whereas the characters in the genetic sequence can be 1 in 4, there is more information per character in the amino sequence than in the genetic sequence. Using amino acids as inputs, our model scored a higher recall of .96 (Fig 4b) compared to using genes as inputs.

3) Modelling Amino Acid - Drug Interactions on DrugBank + Pubchem Data

After incorporating amino acids and drugs from DrugBank, we expanded the dataset to other, larger databases. We mix the interaction data from DrugBank and BindingDB and then separate out training and testing samples; mixing data preserves the type of chemicals being modeled. We then train the model on the training set and evaluate the performance of the model on the test set. To our surprise, the model recall increases significantly compared with only using DrugBank (Fig 3C). We hypothesize that the doubled size of training data and BindingDB labels directly tied to binding score could have influenced this change. The mixed validation set with our new model scored the highest f1-score across all models and experiments, suggesting that our efforts do capture general patterns for classification.

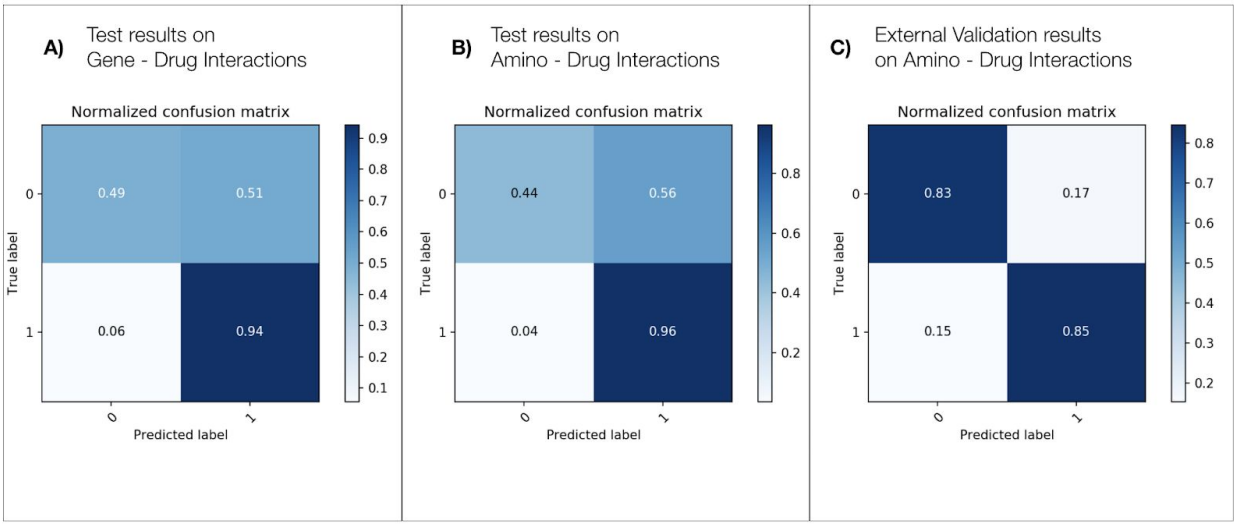


Figure 3: *Confusion matrix of the stage 2 convolution model for raw fingerprint inputs with A) target gene inputs, B) target amino acid inputs, and C) target amino acid inputs with the mixed BindingDB and DrugBank dataset.*

Web App

The final web application enables users to find potential genes/amino acids that might interact positively with a given drug. First, users will access the website using the host IP address. They will be greeted with a home page that consists of a search bar (see Appendix). One a user clicks on the search bar, they can start typing a either a drug or target amino or gene. Once a character is typed the platform begins recommending potential drugs or targets that can be searched via a drop down menu (Fig 4). They can initiate a search by clicking the desired drug or target or by typing the desired entity and hitting enter. The model will then generate target names, probabilities, and whether or not it is a candidate via the Results

category if the search was a drug. The same will occur if a target is searched, but drugs will appear in the results table (Fig 4).

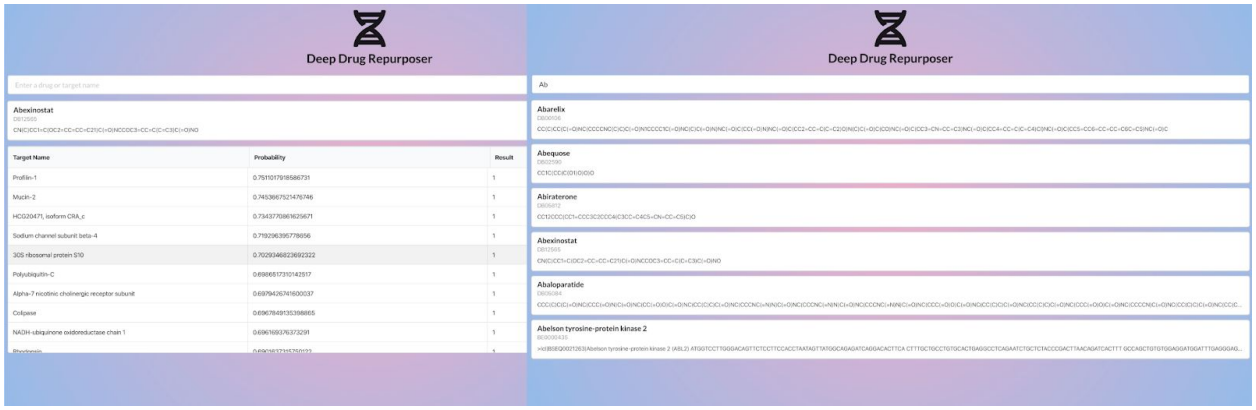


Figure 4: *Depiction of the user interface search results and recommended drug list when attempting to search.*

Discussion

In view of both stage 1 and stage 2 of our project, this project successful fulfills the objective of creating a drug repurposing classifier using DrugBank data. We implemented a novel network-based decoy sampling method to generate negative samples from a database with purely positive relationship, and used the synthesised dataset to train a model with reasonably high accuracy. We also determined trimmed chemical fingerprint and gene vectors to be the most successful inputs for convolutional neural networks, rather than our earlier attempt of using chemical embeddings. Finally, we implemented a web app that could be used by researchers to leverage our model insights with DrugBank drug-target relationships.

This project has clear limitations, particularly in data synthesis and model prediction. Our mode is inherently tied to DrugBank's perspective of what is a drug-target relationship. Although we attempted to integrate BindingDB data as an external source, the threshold for the 0 and 1 labels were determined based on DrugBank's reactivity distribution (discussed in the midterm report). In hindsight, we would have used a more objective metric such as the chemical-target reactivity scores. Not only would predicting drug-target reactivity more directly relate to objective molecular properties, but it would also have allowed us to compare our model with other research that predicts molecule-target reactivity. Second, we were not able to try all combinations of input types and models due to time constraints. Other models could have performed better with the raw fingerprint input vectors, particularly the converging convolutional neural network. Although our project does not produce fully conclusive results, it was an initial step towards making a drug repurposing model as a binary classifier.

Statement of Contributions

Forrest Hooton casted the high level vision of the project and implemented data pipeline for model inputs.

Noah DeMoes contributed to model development, and headed report writing.

Sarthak Kothari was responsible for implementing the web application.

Omair Shafi Ahmed headed model development and tuning.

References

[1] Business Research Company. “The Growing Pharmaceuticals Market: Expert Forecasts and Analysis.” *Market Research Blog*, blog.marketresearch.com/the-growing-pharmaceuticals-market-expert-forecasts-and-analysis.

[2] Ciallella, John, and John Ciallella. “A New Use for Old Drugs.” *Eureka*, 7 Aug. 2019, eureka.criver.com/a-new-use-for-old-drugs/.

[3] Commissioner, Office of the. “Step 3: Clinical Research.” *U.S. Food and Drug Administration*, FDA, www.fda.gov/patients/drug-development-process/step-3-clinical-research.

[4] Dimasi, Joseph A., et al. “Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs.” *Journal of Health Economics*, vol. 47, 2016, pp. 20–33., doi:10.1016/j.jhealeco.2016.01.012.

[5] Jaeger, Sabrina, et al. “Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition.” *Journal of Chemical Information and Modeling*, vol. 58, no. 1, 2018, pp. 27–35., doi:10.1021/acs.jcim.7b00616.

[6] Kim, Sunghwan, et al. “PubChem 2019 Update: Improved Access to Chemical Data.” *Nucleic Acids Research*, vol. 47, no. D1, 2018, doi:10.1093/nar/gky1033.

[7] Nguyen, Thin, et al. “Prediction of Drug–Target Binding Affinity Using Graph Neural Networks.” 2019, doi:10.1101/684662.

[8] Rogers, David, and Mathew Hahn. “Extended-Connectivity Fingerprints.” *Journal of Chemical Information and Modeling*, vol. 50, no. 5, 2010, pp. 742–754., doi:10.1021/ci100050t.

[9] Wishart, David S, et al. “DrugBank 5.0: a Major Update to the DrugBank Database for 2018.” *Nucleic Acids Research*, vol. 46, no. D1, 2017, doi:10.1093/nar/gkx1037.

[10] Gilson, Michael K., et al. “BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology.” Nucleic Acids Research, vol. 44, no. D1, 2015, doi:10.1093/nar/gkv1072.

Appendix

Link to GitHub code:
https://github.com/fhooton/DeepDrugRepurposer

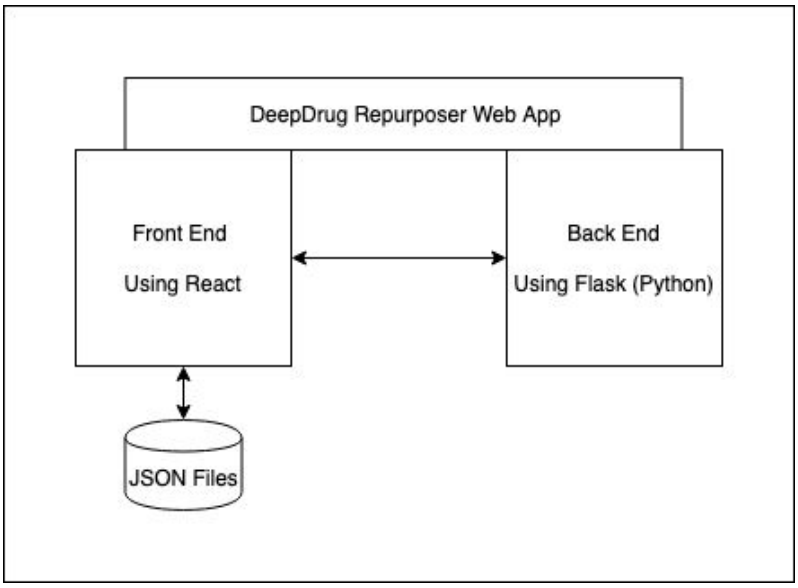


Figure A1: *Web application architecture.*

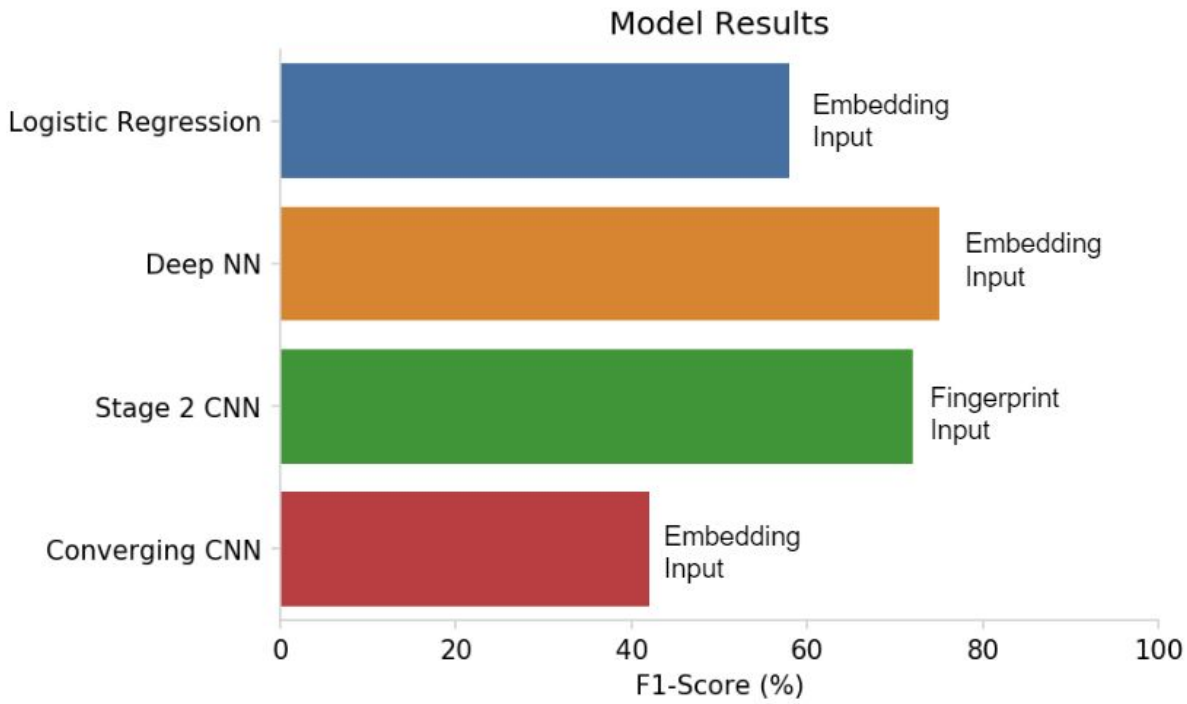
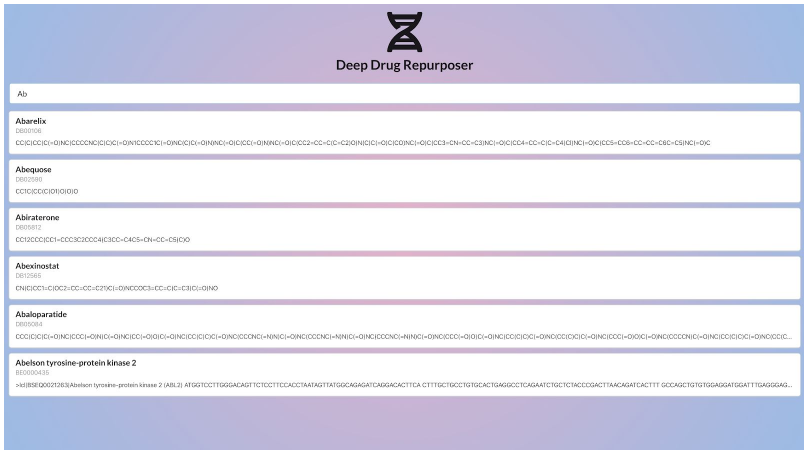


Figure A2: *Model f1-scores for all models across the semester. Earlier models used Mol2Vec embeddings as inputs, denoted by the bar text.*

Web Application Screen shots

Home Page:

Search:



Predictions/Results: