Analysis of Protein-Ligand Binding Through Machine Learning Techniques

Introduction

In drug discovery campaigns, the use of protein structural prediction and computational screening of ligands has become increasingly valuable. This is in part due to the relative speeds of computational methods in comparison to wet lab experiments: it is much more efficient to screen thousands of ligands as a potential drug candidate for a target protein than it is to run those thousands of experiments in a lab. On the other hand, the growth of machine learning and the broadening of its applications has enabled a significant expansion of artificial intelligence and machine learning algorithms into the chemical and biological disciplines. This is especially evident in protein chemistry, where machine learning has direct applications in drug discovery campaigns and other pharmaceutical purposes.

When discussing applications in the pharmaceutical industry, one significant issue that is often investigated is the binding affinity of drug candidates, as that metric can correlate to the effectiveness of a certain drug candidate. In addition, a serious consideration that drug manufacturers must consider are the potential side effects that come with a certain drug. In this paper, I will be investigating both of these considerations by drawing on the PDBbind+ database and studying protein-ligand interactions. The PDBbind+ database contains a plethora of proteins along with a ligand that is known to bind to that protein. For my final project, I decided to explore how machine learning algorithms could be used to analyze the PDBbind+ database, extract structural similarities, and predict the binding affinity between a ligand and its protein.

Methods

Data Extraction

The PDBbind+ database provides thousands of folders labeled by their Protein Data Bank (PDB) ID, with each folder containing a pdb file containing the full protein sequence, a pdb file containing all amino acids found within ten angstroms of the ligand, and files describing a ligand known to bind to the protein.

For this paper, the database used was the “PDBbind v2020 Protein-ligand complexes: The general set minus refined set,” which consisted of 14,127 protein-ligand complexes. Given the constraints of time and computational resources, this study only used approximately 3,000 of the 14,127 protein-ligand complexes. In order to extract the necessary information from these files, the folders were iterated through and each file was individually analyzed for certain parameters. In this study, we focused on analyzing the ligands and the protein pockets. This is because the full protein sequence, while can be influential to knowing the full protein three-dimensional geometry, seemed to be not as useful to predicting the protein binding affinity and computationally inefficient to completely process. Thus, it was computationally more efficient to focus on the protein pocket files, as they contained the necessary information to determine protein-ligand binding affinity.

From the ligand files, a variety of parameters were extracted, including physicochemical properties such as molecular weight, partition coefficient, topological polar surface area, number of hydrogen donors and acceptors, and molecular volume. However, these properties do not sufficiently describe the

structural features of the molecule, such as functional groups and elements present. In order to include these descriptors, Morgan fingerprints were generated for each molecule. One issue with Morgan fingerprints is that they are very sparse data structures, which can be memory-expensive and inefficient. To address this issue, the Morgan fingerprints were dimensionally reduced using principal component analysis (PCA) into five dimensions, allowing for more efficient data construction. From the protein pocket files, the frequency of each amino acid was calculated, along with physicochemical properties of the protein such as extinction coefficient, aromaticity, and instability index. Furthermore, some spatial parameters were included such as the convex hull surface area and volume. Overall, the ligand extracted data has 15 features and the protein pocket extracted data has 31 features.

Included in the PDBbind+ database is a file containing the binding affinities of each protein-ligand pair. The binding affinity metric chosen for this study was the logarithmic ratio between the dissociation coefficient and the inhibition coefficient:

$$-\log\left(\frac{K_d}{K_i}\right)$$

This label is used to determine the relative dissociation of the ligand in comparison to its inhibitory activity. This metric is the most generalizable metric across the different protein-ligand pair mechanisms, such as noncompetitive, competitive, or uncompetitive inhibitory mechanisms.

Machine Learning Techniques

After performing data extraction and preprocessing, the data was input into a clustering algorithm to determine structural similarities between each data point. The clustering was split into two separate datasets: one for all of the ligands and one for all of the protein pockets. After sampling multiple dimensional reduction representations as well as clustering techniques, the methods that yielded the most distinct clusters was using t-distributed stochastic neighbor embedding (t-SNE) and density-based spatial clustering of applications with noise (DBSCAN). By generating labels for each cluster in the protein pocket, these same labels were applied to the ligand distributions.

In order to predict binding affinity, multiple machine learning architectures were used to determine which would be most viable. The first method used was a random forest model, as it is an effective learning algorithm that also has an accessible explainability. After testing the random forest model, a deep learning model was used. In particular, the structure of the network was to have the ligand properties be inputted into a network, while the protein pocket properties were inputted into a separate network. Then, the outputs of the two networks were concatenated and processed through a third network, which would output a prediction for the binding affinity.

Results

Protein pocket features are more clusterable than ligand features

After performing the t-SNE dimensionality reduction on the protein pocket features, the clusters formed looked like *Figure 1*, where there were several distinct cluster-like structures.

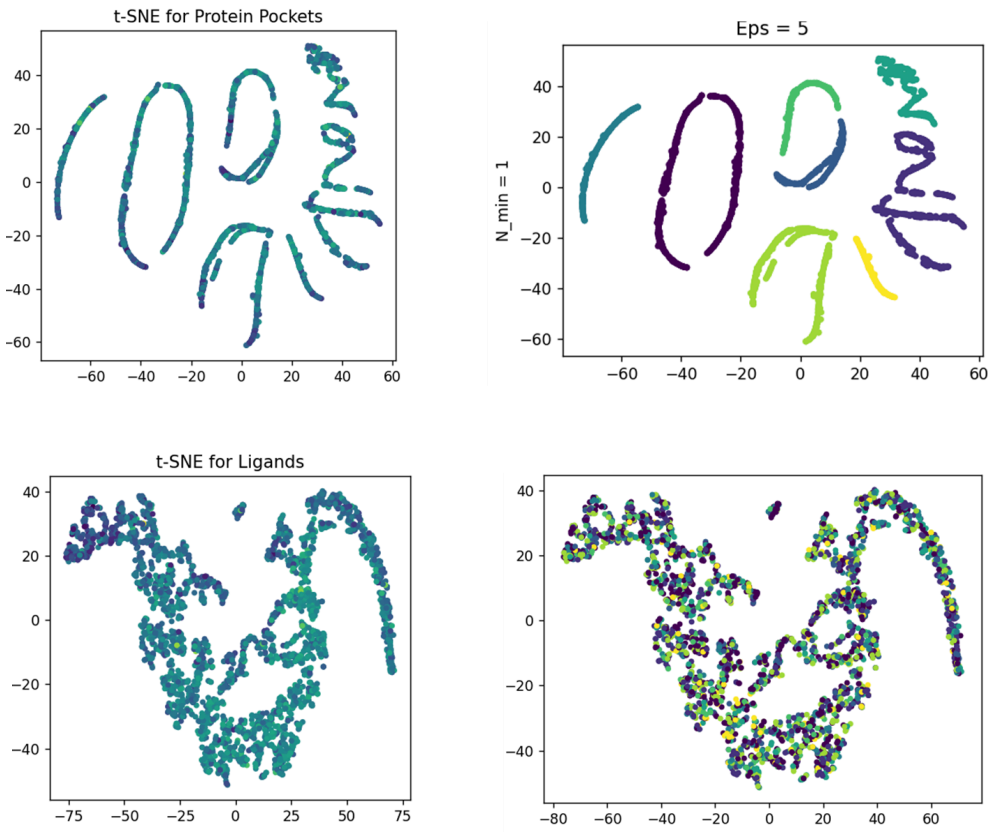

Figure 1. *The top left graph displays the dimensionally reduced protein pocket features, with the color scheme indicating the binding affinity. Lighter colors indicate a smaller binding affinity value and darker colors indicate larger values. The top right graph displays the clusters identified by the DBSCAN algorithm. The bottom left graph displays the dimensionally reduced ligand features. The bottom right graph displays the dimensionally reduced ligand features with the same labels as the top right graph.*

As seen in *Figure 1*, the clusters in the protein pocket feature space are visually distinct, but when using the same labels in the ligand feature space, the clusters are not distinguishable.

Random forest model performs better than deep learning architectures

In order to predict binding affinity, a test: validation: training set distribution of 20:20:60 was used to train the models. After testing multiple different deep learning structures, the most optimal structure was able to predict the training set binding affinities with an R^2 value of 0.748 and the testing set binding affinities with an R^2 value of 0.359. This is a significant disparity between the testing performance compared to the training performance, which indicates the presence of overfitting. This is further confirmed in *Figure 2*, where the validation loss is largely stagnant.

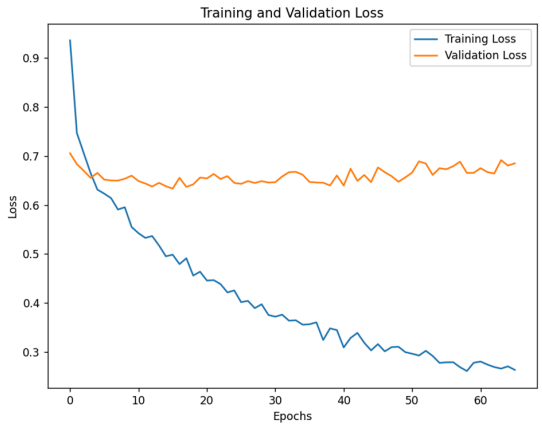

Figure 2. *A plot of the validation and training loss throughout the training process of the deep learning architecture.*

Since the deep learning model experienced overfitting, a random forest model was evaluated to determine whether it would perform better. After optimizing the hyperparameters of the random forest model, it was found that the random forest model performed marginally better than the deep learning model, with predictions of binding affinities in the test set with an R^2 value of 0.386. While this model does not perform much better than the deep learning model, it is possible to extract significance values of each feature by calculating the Shapley additive explanations (SHAP) values. As shown in *Figure 3*, the most influential feature in the determination of predictions in the random forest model is the convex hull volume of the protein pocket. This is an interesting observation because, intuitively, three-dimensional descriptors of the protein pocket and ligand should be significantly influential to the prediction of binding affinities.

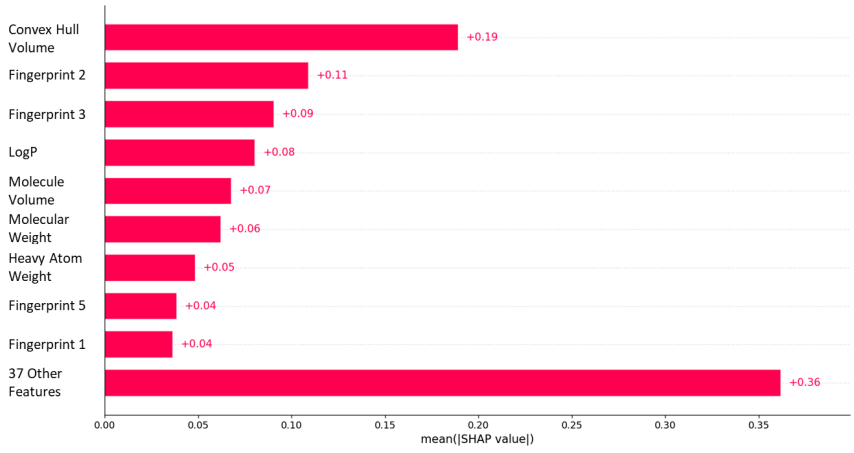

Figure 3. *A bar graph showing the SHAP values of the random forest model.*

Conclusion

This study demonstrates the potential of machine learning techniques to analyze protein-ligand interactions, particularly for predicting binding affinities using the PDBbind+ database. By leveraging

extracted physicochemical and structural features of protein pockets and ligands, the clustering analysis revealed that protein pocket features are more amenable to cluster formation than ligand features. This insight underscores the significance of protein structural properties in predicting binding interactions.

When comparing machine learning models, the random forest algorithm marginally outperformed the deep learning architecture, achieving a test set R^2 value of 0.386. The deep learning model's overfitting highlights the challenges of applying neural networks to relatively small datasets with high-dimensional features. The interpretability of the random forest model through SHAP analysis further revealed the convex hull volume of the protein pocket as the most influential feature, aligning with the intuitive importance of three-dimensional spatial descriptors in binding affinity prediction. For future advancements in this study, three-dimensional spatial descriptors should be further explored. In particular, one possible way to incorporate three-dimensional structural features is by plotting the protein and ligand atoms in a Cartesian coordinate and inputting them into a convolutional neural network.

Nonetheless, while the models' performances leave room for improvement, this analysis emphasizes the promise of integrating dimensional reduction, clustering, and explainable machine learning techniques in drug discovery campaigns. Future work could focus on refining feature selection, incorporating additional spatial and dynamic descriptors, and leveraging larger datasets to enhance model performance and generalizability. Ultimately, this research highlights the power of computational methods to complement experimental approaches, accelerating the identification and optimization of potential drug candidates.