Posing with the Coronavirus Spike Protein

Mahathi Vempati (20161003) * **Vivek Iyer (20161188)** * **Shubh Maheshwari (20161170)** *

Abstract

We apply a two-fold machine learning approach to predict the binding affinity of a given ligand with the novel coronavirus (SARS-CoV-2) spike protein. Firstly, given a ligand we predict its best binding mode with the coronavirus spike protein. This is done by predicting the rigid body transformation (rotation and translation) it has to undergo to bind to the protein. Secondly, once we have the binding mode, we predict the docking score using an ensemble of fully connected neural networks. For the purpose of this project, we use the extensive dataset created with Autodock on the supercomputer SUMMIT as ground truth.

1. Introduction

Drug discovery involves disrupting any part of the infecting organism's pathway. In the case of the novel coronavirus, one of the first stages is the binding of the spike protein to the ACE-2 receptor of the cell. This interface is shown in Fig.1. One of the ways to disrupt this stage would involve finding a molecule that has a high affinity to this interface, therefore breaking up the interaction. With this goal in mind, we look at the stages of drug discovery itself.

High Throughput Screening (HTS) is the initial stage of drug discovery that involves testing large numbers of molecules for their affinity against a receptor in question. This stage has been heavily accelerated, and has yielded better results with the advent of computational tools to aid the process. Prominent among these is 'Docking and Scoring', which is the computational method of sampling and running simulations to find the best fit for a given ligand with a protein, and then ranking several ligands' best fit to find the best binding ligand. Indeed, docking has predicted new ligands for 50 targets in a period of just five years. Not only this, docking has also illuminated false negatives that arose in the experimental procedures of High Througput Screening (1). More recently, there have been instances of docking proving to be significant in discovery of medication (2).

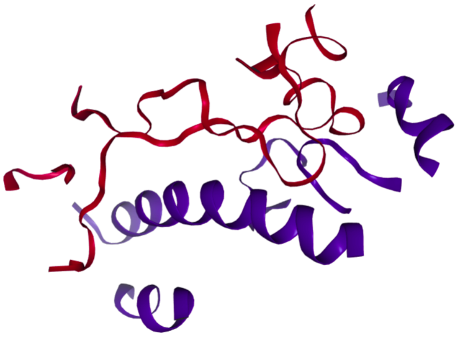

Figure 1. The portion of the interface of novel Coronavirus and ACE-2 receptor under consideration in our project. The purple residue belongs to the ACE-2 receptor and the red residue is of SARS-CoV-2.

The process of docking inevitably involves sampling the space of orientations with respect to the binding pocket of the protein; it is this that constitutes the computationally heavy and time consuming part. Attempts (3) have been made to make the sampling process more efficient, but sampling is still present nonetheless. Even in the latest Grand Challenge (4), a pose prediction competition, most winning submissions relied on docking. For scoring, however, given a pose, predicting the affinity seems to be a much more attainable endeavour, with protein+ligand based featurization such as SPLIF (5). The DL-Score machine learning technique claims to do even better than Autodock (6).

In this landscape, we attempt to use machine learning to evade sampling altogether, and then test several techniques to find the one that best scores our predicted poses. Our paper is divided as follows: In the Methods Overview section, we elucidate the three main divisions of our work. In the Experimentation section, we describe the dataset, our experiments, an the results obtained. We then conclude with how our project can be improved further.

2. Method Overview

Our process involves 3 phases. Ideally, we first shortlist a set of active ligands. Due to the nature of our dataset, this was not necessary for our main experiment. After finding a set of suitable ligands, the second phase involves predicting the

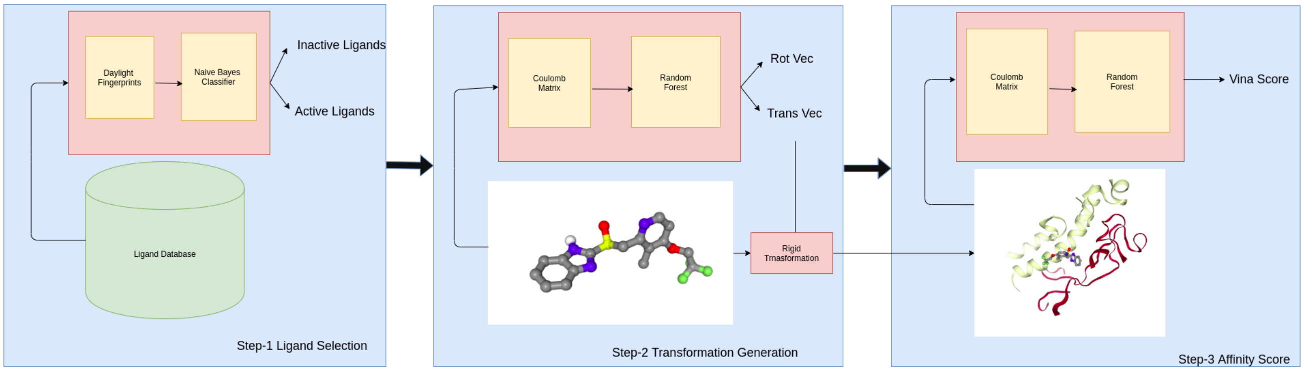

Figure 2. Pipeline of our method. 1) For a large chemical database, we use naive bayes to select possible candidates. 2) Given the spike-protein of the corona-virus and the initial pose of the ligand we estimate the rigid transformation for to dock between the spike protein and the ACE-2 receptor. 3) Lastly we evaluate the docking by predicting the protein-ligand affinity score.

best pose for each ligand. Lastly we predict the free energy using the protein-ligand interaction to be able to compare between various ligands. The full code for our project can be found here.

2.1. Drug selection

Before we apply our methods to pose and score a drug, we do a preliminary assessment using a Naive Bayes (12) classifier trained on a list of 280,000 molecules published by PubChem (11) on whether it is active or inactive. We use 'Daylight Fingerprints' to featurize the molecules. Our code can be found here.

2.2. Predicting transformation

Consider the task at hand: Predicting a binding pose without sampling. Given the feature set of a ligand, one would have to predict the final coordinates of all the atoms with respect to the protein. This is cumbersome for two reasons. Firstly, each input has a different output size based on the number of atoms. Secondly, the size of the predicted output is extremely large for big molecules: we would have to predict around 500 values for some molecules.

Instead, we approximate this task to a rigid body transformation problem. What if, all we did was predict how a given molecule rotated and translated to fit into the binding pocket of the protein? Then, irrespective of the molecule size, we would have to predict only 6 parameters for each (the translation and rotation vector)! Note that this would be possible to do if the initial orientation of the molecule was standard and "trainable" in some sense.

For this purpose, we first orient the molecules such that their principal axes (eigenvectors of the Moment of Inertia tensor) align with the x, y and z axes. Once all molecules are oriented correctly, we then compute the translation and rotation vectors.

Now, for every ligand we have a translation and rotation vector, which constitute our Y values for training. For the X values, we consider three different types of features. First, the circular fingerprints: Circular fingerprints are bitstrings that uniquely identify a ligand based on its features, by iteratively taking into consideration the neighbours of every atom using the Morgan algorithm (8). Second, we use eigenvalues of Coulomb matrices and thirdly we use Chemical descriptors: an array of 111 values for every ligand based on several chemical properties. We compare how the three features perform.

To predict rotation and transformation matrices from these features, we consider various ML models such as a Support Vector Machines, Random Forest Regressors and Multi Layer perceptrons. In the former, we use epsilon-Support Vector Regression to individually predict every element of the transformation and rotation matrices independently. Random Forest Classifiers use an ensemble of classifying decision trees on various subsets of the original dataset in an effort to improve accuracy and reduce overfitting whilst predicting the multi label output of transformation and rotation matrices. Lastly, we used a Multi Layer perceptron, which is essentially a feed-forward neural network that attempts to learn weights using 4 linear layers and a ReLU activation function.

2.3. Scoring Rigid Transformation

After estimating the rotation euler angles and transformation vector, rigid transformation is performed on the ligand to get the new position of each molecule: $\hat{X}$

$$\hat{X} = \hat{R} \cdot X + \hat{T} \tag{1}$$

where $\hat{R}$ is the rotation matrix created using the predicted euler angles and $\hat{T}$ is the predicted docking location w.r.t to the spike protein receptor.

The prediction of protein-ligand binding affinities is crucial for drug discovery research. Essentially the goal is to provide a score to each protein-ligand interaction. More the score higher the changes of successful docking during simulation.

Inspired from the work by (9) the process is divided into 2 steps.

- **1. Feature Extraction**: We use BINANA(10) to extract protein-ligand feaatures. For every ligand and protein atoms within a distance of 2.5 A - 4.0 A electrostatic interactions, binding pocket flexibility, hydrogen bonds, salt bridges, van der Waals, rotatable bonds, π interactions, among others 20 . A total of 348 features were considered for each protein-ligand complex.
- **2. Score Prediction using MLP** A fully connected neural network is trained to predict the score. Mean square error loss is used to train the network. Adding regularization such as L2-norm and Dropout significantly improve the performance of the model.

3. Experimentation

3.1. Dataset

Our dataset (7) consists of around 8500 ligands. Each ligand has around 10 final poses with the interface of the Spike protein and the ACE-2 receptor. Each pose is given an Autodock score. For our project, we use the best pose for every ligand, and its corresponding score.

3.2. Quantitative analysis

In the following section, we use the RMSE and R2-score metrics to analyse the performance of our models. The Root Mean Square Error (RMSE), or the Root Mean Square Deviation (RMSD) is a standard deviation of the residuals. In other words, it is a measure of how far from the ground truth the predicted values are. The R2 score, or the coefficient of determination is the square of the correlation between predicted labels and ground truth labels, and in other words it is used to analyze how strong of a relationship there is between two variables.

As part of our project, we calculate the R2-score and RMSE for each of the 12 model-feature pairs and report them as a metric to compare the performance of each model. Since R2-score indicates degree of correlation, we prefer models with high positive correlation, i.e close to +1. Zero indicates no correlation, and negative values indicate negative correlation. RMSD is also quite similar, with RMSD values close to +1 being preferred.

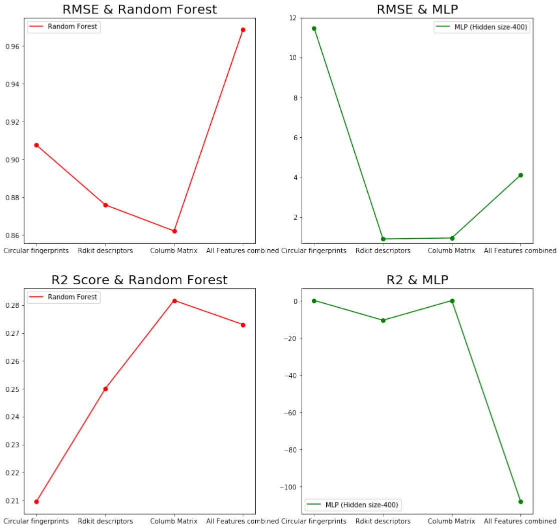

Figure 3. A plot depicting the performance of different models for different features

Figure 3 shows the performance of different models across different feature spaces, displayed for various metrics. As shown in the figure, Random Forest classifier significantly outperforms the MLP, with MLP giving RMSE as high as 11.48 and an R2-score of 0.13. In contrast, the Random Forest classifier gave an RMSE of 0.97 and an R2-score of 0.28 as its best performance. The Coulomb matrix was observed to give the best R2-scores for both models, while Circular fingerprints give the best RMSE values. Overall, we observed the best performance using a Random Forest classifier, with an RMSE of 0.97 and R2-score of 0.273.

For reference, note that atom coordinates are all of the order 1 to 10 units.

In the scoring phase, we score our poses based on the DL-Score method (as cited above), and the BINNANA method, as described above. The RMSE for BINANA is 1.29, and the RMSE for the DL-Score method is 5.4. This refers to the scores obtained for the correct poses vs the scores obtained for the poses we predicted.

The ranges for these scores were around -3 to 3, so we can see that our poses did well on the BINANA scoring criteria but did very poorly on the DL-Score criteria.

3.3. Qualitative

We study one of the predictions made by our model. Consider the drug Lansoprazole. The initial position with respect to the spike protein interface is as shown in Fig.4. The final position (the ground truth) is shown in Fig.5. Note the upward placement of the benzene-like ring structure com-

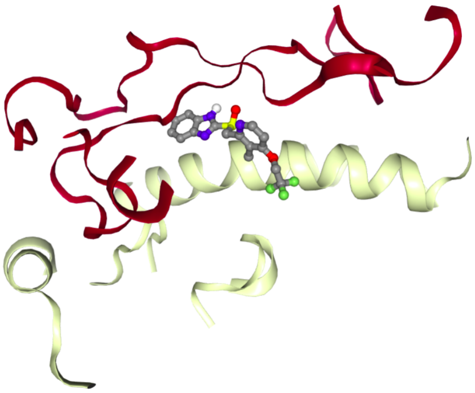

Figure 4. Initial position of drug Lansoprazole with respect to the spike protein interface. This is a randomly generated configuration.

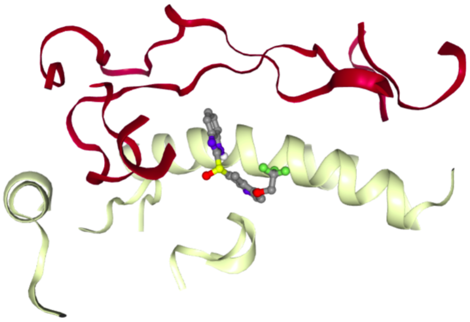

Figure 5. The final position of the drug as given in the dataset.

pared to the initial position. Our predicted configuration as shown in Fig.6 also gets the upward placement of the benzene-like ring structure correctly, however, the Fluorines (green) are not tilted as upward as they should be. Perhaps relaxation of rigidity would help the transformation.

4. Conclusion + Future Work

The two areas for improvement for our work are

- The lack of a ”standard” orientation for our model: We highlight once again that for a model that predicts only transformations, rather than the final configuration, the initial configuration is of paramount importance - to ensure that it is sensible for training, and to ensure that any molecule being tested is first oriented to this standard configuration. In our work, we picked the configuration where the principal axes of the molecule are oriented with the coordinate axes, but this may not be the best standard orientation to have. A project that naturally emerges from our work is to evaluate different notions of a standard orientation of a molecule. A standard orientation could depend on specific features, depend on weights, etc.

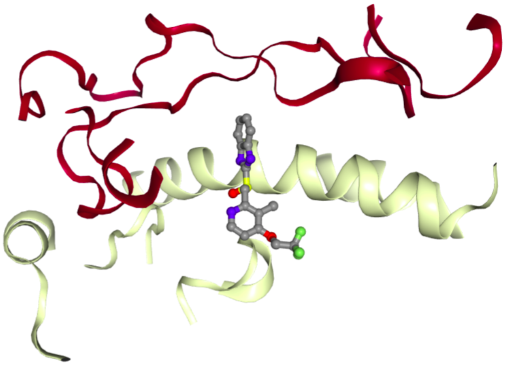

Figure 6. The predicted position of the drug by our model.

- Going beyond rigid body transformations: as illustrated in an example in the paper, the model is unable to fully predict the final configuration because of rigid body constraints. Careful relaxation of this constraint, perhaps taking into account what parts of a molecule maybe considered flexible may yield better results. This also gives rise to the problem of possibly not having same sized outputs for all inputs, which needs to be taken care of.

References

[1] Coleman RG, Carchia M, Sterling T, Irwin JJ, Shoichet BK (2013) Ligand Pose and Orientational Sampling in Molecular Docking. PLoS ONE 8(10): e75992. https://doi.org/10.1371/journal.pone.0075992

[2] Mark Andrew Phillips, Marisa A. Stewart, Darby L. Woodling and Zhong-Ru Xie (July 11th 2018). Has Molecular Docking Ever Brought us a Medicine?, Molecular Docking, Dimitrios P. Vlachakis, IntechOpen, DOI: 10.5772/intechopen.72898. Available from: https://www.intechopen.com/books/molecular-docking/has-molecular-docking

[3] J. Phys. Chem. B 2018, 122, 21, 5579-5598 https://doi.org/10.1021/acs.jpcb.7b11820

[4] Grand Challenge 3 https://drugdesigndata.org/about/grand-challenge-3

[5] Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. J Chem Inf Model. 2014;54(9):2555-2561. doi:10.1021/ci500319f

[6] J. Chem. Inf. Model. 2017, 57, 4, 942-957 https://doi.org/10.1021/acs.jcim.6b00740

[7] Smith, Micholas; Smith, Jeremy C. (2020): Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.11871402.v3

[8] Morgan algorithm https://depth-first.com/articles/2019/01/11/extended-connectivity-fingerprints/

[9] DLSCORE: A Deep Learning Model for Predicting Protein-Ligand Binding Affinities https://chemrxiv.org/articles/DLSCORE_A_Deep_Learning_Model_for_Predicting_Protein-Ligand_Binding_Affinities/6159143/1

[10] BINANA: A Novel Algorithm for Ligand-Binding Characterization https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3099006/

[11] National Center for Biotechnology Information. PubChem Database. Source=The Scripps Research Institute Molecular Screening Center, AID=1706, https://pubchem.ncbi.nlm.nih.gov/bioassay/1706

[12] medRxiv 2020.04.05.20054254; doi: https://doi.org/10.1101/2020.04.05.20054254