Binding Affinity Prediction of Protein-Ligand complexes using Machine Learning

MSc Project

Abdus Salam Khazi

Supervisors:
Simon Bray & Alireza Khanteymoori

September 27, 2021

Table of Contents

Biological Background

What are a proteins and ligands?

- **Proteins:** Complex molecules that are work-horses (machines) of a living organism.
- **Ligands:** Molecules that bind to particular proteins, called receptor proteins.
- Proteins and ligands bind together to form protein-ligand complexes.

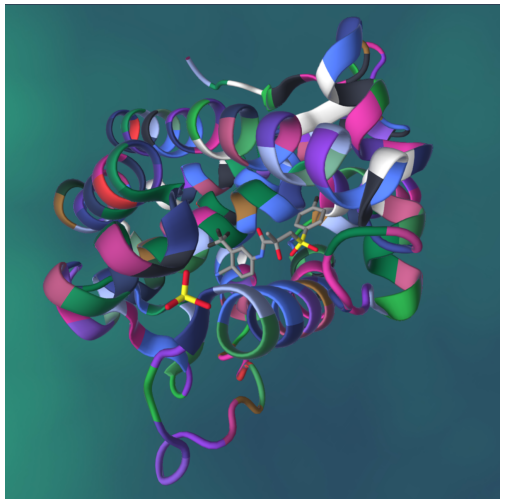


Figure: Haemoglobin transporter protein [6].

Biological Background

Protein-Ligand complexes

- Any potential binding location in the 3D structure of a protein is called a pocket.
- The pockets of proteins only bind to ligands of complementary shape.
- Drugs are just ligand molecules that bind to protein to cause a therapeutic effect.

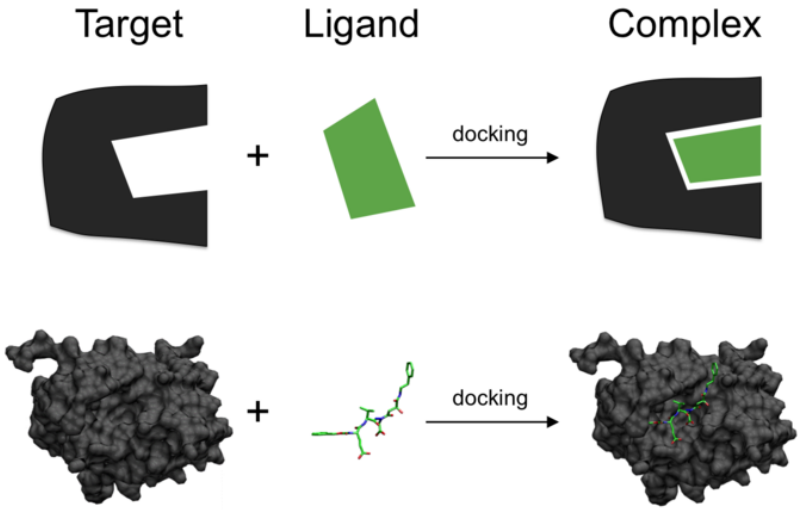


Figure: Lock and Key hypothesis in molecular docking [7].

Biological Background

Understanding Protein-Binding Affinity.

- Binding affinity between a protein and a ligand is quantified by the K_d, K_i and IC_{50}. Here K_d refers to the dissociation constant, K_i to inhibition constant, and IC_{50} to inhibitory concentration 50%.
- K_d can be quantified by using protein concentration $[P]$ and ligand concentration $[L]$ at equillibrium [1].

$$K_d = \frac{[P][L]}{[PL]}$$

- K_i and IC_{50} are similarly defined.

Problem Definition

- Determining if a potential drug (ligand) can bind to a target protein is very costly processes [2].
- The project aims to predict the ligand affinity based on previously recorded data ("In-Silico" method). This reduces the drug discovery costs.
- We use PDB databank, which holds PL affinity data collected over many decades.

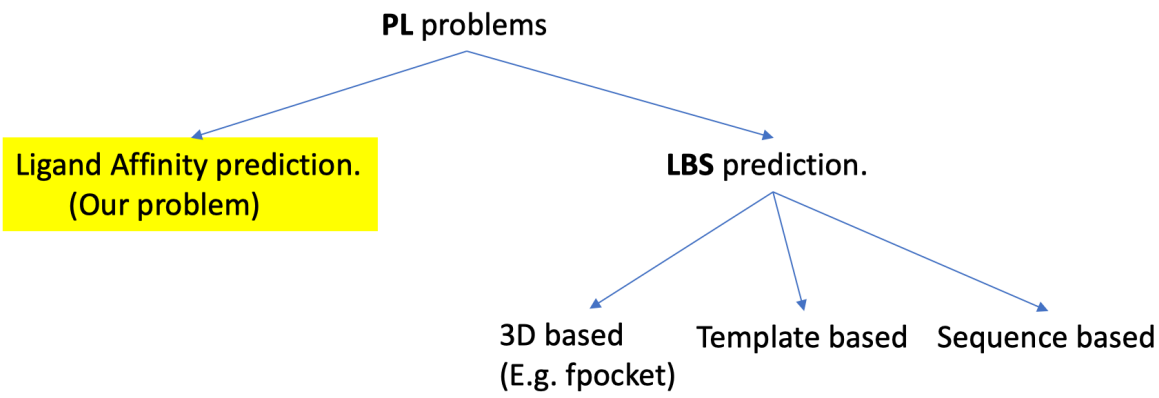


Figure: Protein-Ligand problem classifcation.

Feature Extraction

PDB databank (v2019) was used to extract input features.

- We use *fpocket/dpocket* ligand binding site prediction library to get the features of pockets pockets in proteins.
- *RDKit* library is used to extract features for each ligand.
- **Ligand Features:** Using *RDKit.Chem.Descriptors*, 402 features were extracted for each ligand. Hence the ligand features space was $\mathbf{R}^{402}$.
- **Protein Features:** For every pocket, 55 descriptors are obtained in total. Hence, the input space for protein features is $\mathbf{R}^{55}$

The concatinated input feature space before input feature elimination $\mathbf{R}^{457}$.

Feature Selection

We only had 16000 data points to train a feature space of $\mathbf{R}^{457}$. We reduced our features using the following feature selection strategies:

- **Output Correlation**: The input features that have the best *Pearson* and *Spearman* correlation were selected. [5].
- **Genetic Algorithms**: Genetic algorithms with the following score function was used to select the best features [4]:

$$\text{score} = \mathbf{R}^2\text{score} * \text{Features Eliminated}$$

- **Manual Feature Selection:** A selected list of 121 ligand descriptors was used with all protein descriptors as input to the model.

Testing strategy

We use the following methods to determine the quality of results and reproducing them:

- **Reproducibility**: To reproduce the results, we use report random seed (Execution ID) for every execution.
- R^2 **score** (*Coefficient of determination*) [3]: . $R^2 \in (-\infty, 1.0]$ where 1.0 is the best score.
- **Visualization:** Our model's approximated function $f : \mathbf{R}^n \mapsto \mathbf{R} \;\; \text{where} \;\; n \in \mathbf{I}^+$ is visualized as a 2D scatter plot.

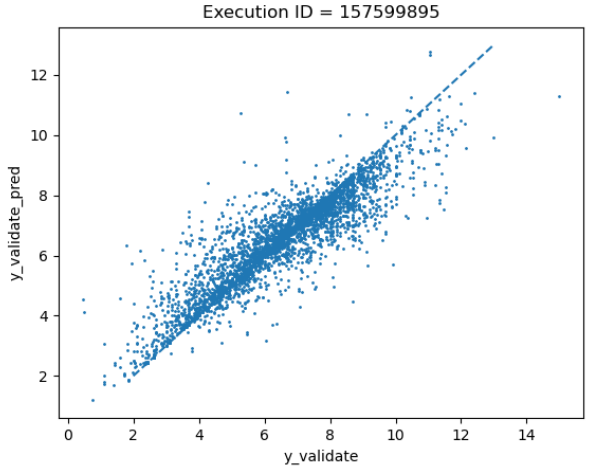
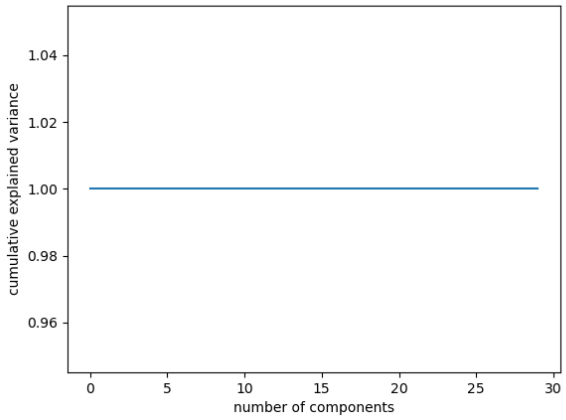


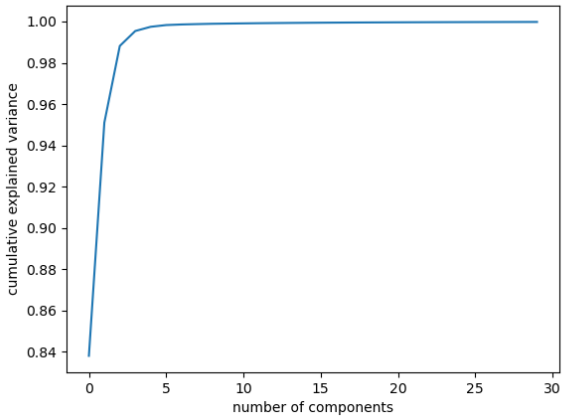
Figure: (Sample) Visualizing accuracy. $R^2 \approx 0.805$.

Data Preprocessing

- Anomalies such as NaN (Not a number) values were removed from the data before sending them as input to the model.
- We used PCA (principle component analysis) to find that the ligand feature *IPC* was having log scale values.



(a) With original Ligand feature IPC.



(b) With log scaled Ligand feature IPC.

Figure: Cumulative PCA of ligand features.

Dealing with measurement resolution

- In the PDB databank, each complex also has a corresponding measurement resolution.
- The structural detail of the 3D image is inversely proportional to the measurement resolution.
- The weighting of each data point was done according to hyperbolic formulae and linear formulae.
  $W_i = \frac{\max \mathrm{R}_{1\ldots n}}{R_i}$; $W_i = (\max \mathrm{R}_{1\ldots n} + 1) - R_i$

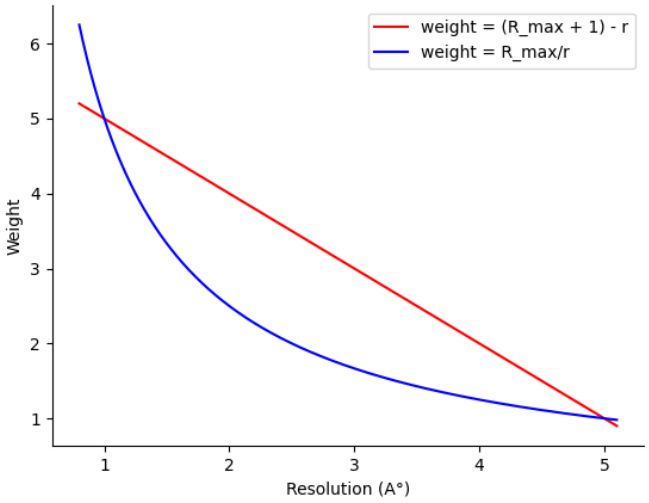


Figure: Weight calculation formulae.

Q & A

References

- Du, Li, Xia, Ai, Liang, Sang, Ji and Liu; Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods (2016)
- DiMasi, Grabowski and Hansen; nnovation in the pharmaceutical industry: New estimates of R & D costs (2016)
- scikit-learn; R2 score, the coefficient of determination,
- John H. Holland. Genetic Algorithms. (1960)
- Schober, Boer and Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. (2018)
- Protein–Ligand complex; Wikipedia.
- Docking (molecular); Wikipedia.