Binding Affinity Prediction of Protein-Ligand complexes using Machine Learning

MSc Project

Abdus Salam Khazi

Supervisors:
Simon Bray & Alireza Khanteymoori

October 10, 2021

Table of Contents

Biological Background

What are proteins and ligands?

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

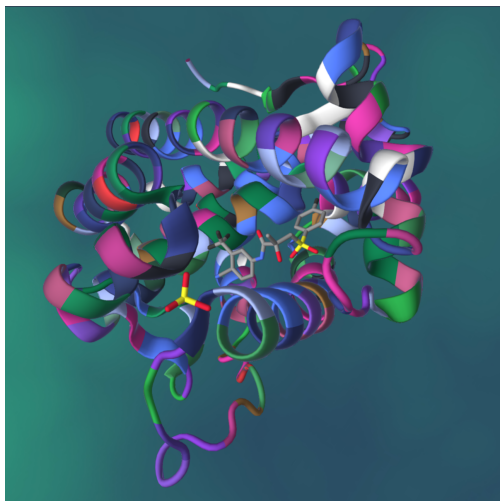


Figure: Haemoglobin transporter protein.

Protein-Ligand complexes

- Ligands bind to proteins at "cavity" like locations called pockets.
- The pockets and the ligands are complementary in shape.

Drugs

- Drugs are ligand molecules that bind to proteins.
- They cause a therapeutic effect after binding to the proteins.

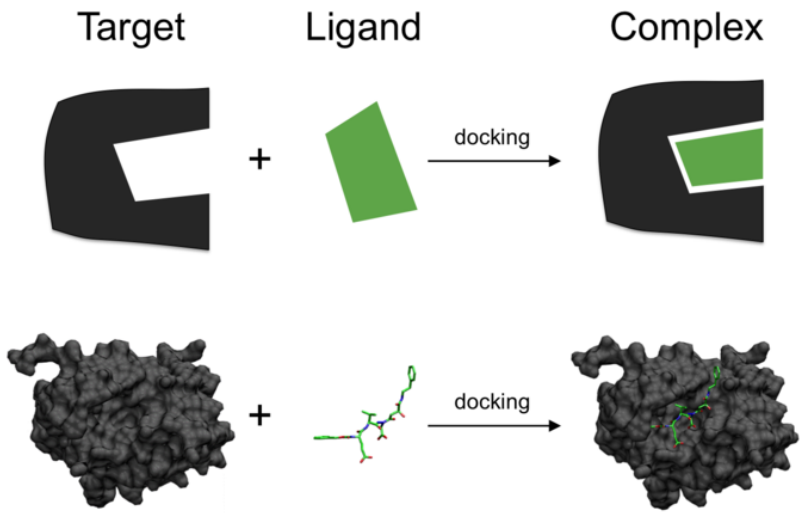


Figure: Lock and Key hypothesis in molecular docking.

Biological Background

Protein-Binding Affinity

- Assume a dynamic system in which protein P and ligand L are binding and unbinding continuously.
- Let $[P]$ be the concentration of the protein and $[L]$ be the concentration of the ligand. Let $[PL]$ be the concentration of the protein ligand complex.
- Binding affinity can be quantified by using $[P]$, $[L]$ and $[PL]$ (**at equilibrium**).

$$\mathrm{BindingAffinity} = \frac{[P][L]}{[PL]}$$

Problem Definition

- Determining if a potential drug (ligand) can bind to a target protein is very expensive [3].
- The project tries to reduce the drug discovery costs by eliminating bad leads.
- **Problem definition**: Predict protein-ligand binding affinity using "In-Silico" methods.

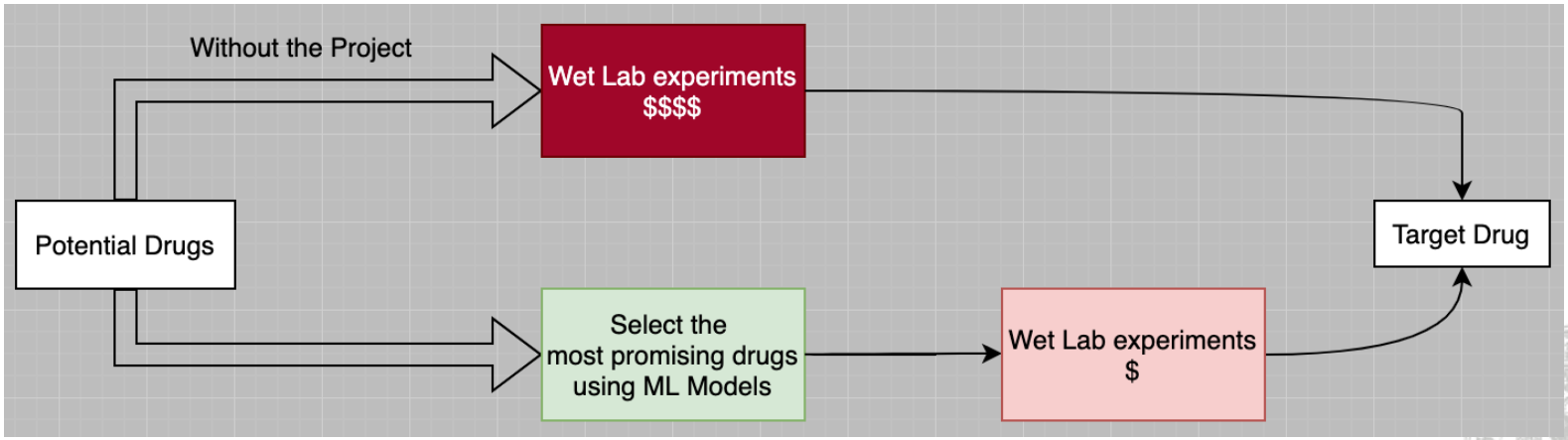


Figure: Project Overview

Problem Definition

There are various problems in the protein-ligand domain. The following figure shows the classification tree.

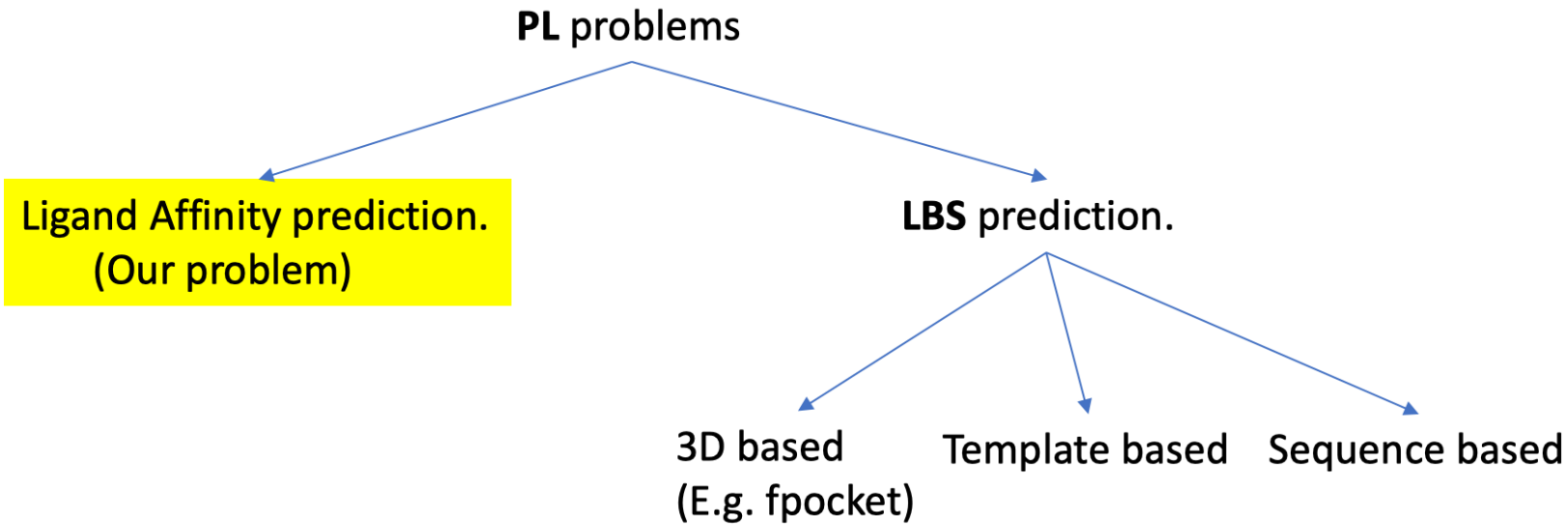


Figure: Protein-Ligand problem classifcation.

Problem Definition

- The input data to the ML model is extracted from a database called PDB Data bank.
- *fpocket* and *RDKit* were used to extract the features of proteins and ligands.
- The input features contain information about the 3D structures of the proteins and the ligands.

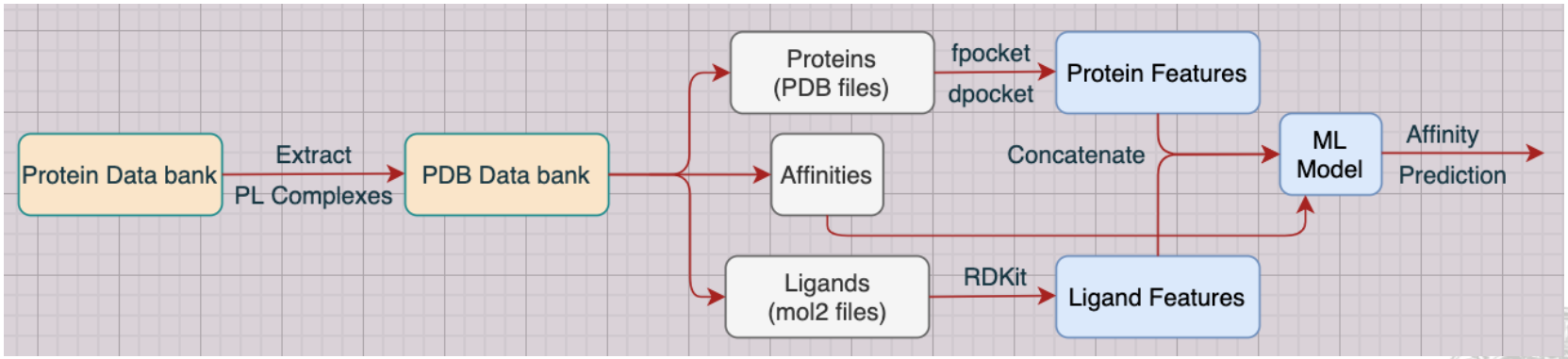


Figure: Data Input Overview.

Feature Extraction

Protein Features:

- *fpocket* is an LBS prediction algorithm used to predict ligand binding pockets.
- There can be multiple binding pockets for a PL complex.
- Using *dpocket*, 55 descriptors were obtained for every (potentially) binding pocket as real values.

Ligand Features:

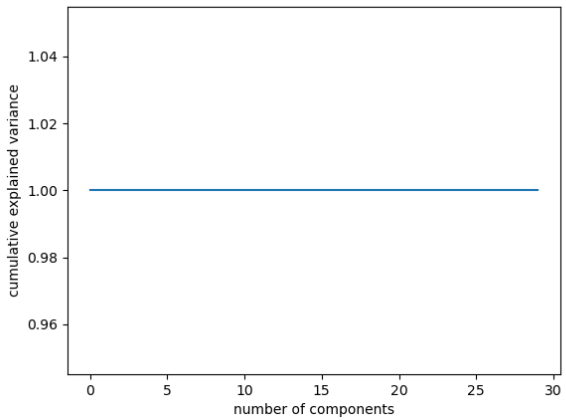
- Using *RDKit.Chem.Descriptors* module, 402 features were extracted as real values.

Concatenation:

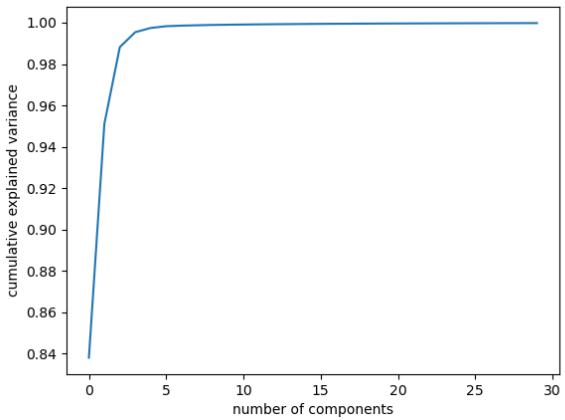
- The (concatinated) input feature space to the model was $\mathbf{R}^{457}$.
- It was less than $\mathbf{R}^{457}$ if feature selection is done before model training.

Data Preprocessing

- Data points containing NaN (Not a number) values were removed from the data.
- PCA (Principle Component Analysis) was used to find the variance contribution of the features.
- Feature *IPC* was log scaled for numerical safety during training.



(a) With original Ligand feature IPC.



(b) With log scaled Ligand feature IPC.

Figure: Cumulative PCA of ligand features.

Feature Selection

Feature selection strategies:

- **Manual:** 121 ligand descriptors + all protein descriptors.
- **Output Correlation**: Features with the best *Pearson* or *Spearman* correlation w.r.t the affinity score (output) were selected.
- **Genetic Algorithms**: Genetic algorithms with a population score function "**R2** score $*$ Features Eliminated" was used to select the best features [4]:

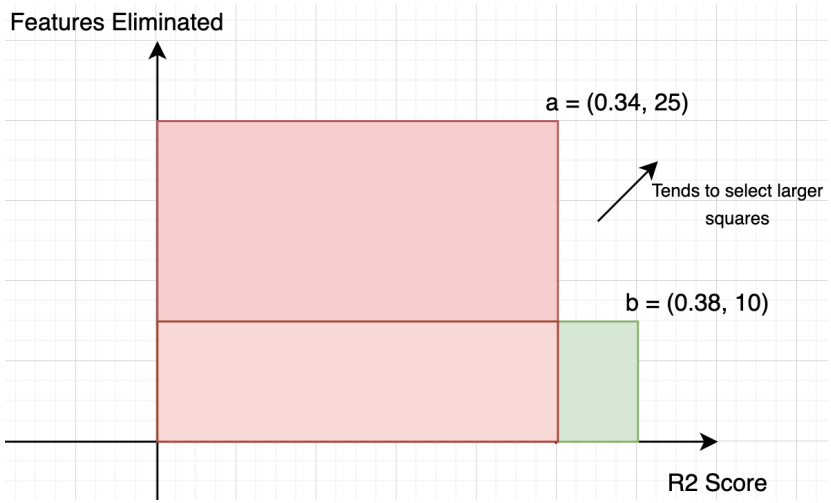


Figure: Genetic Algorithm score function representation.

Dealing with measurement resolution

- In PDB bind databank, each complex has a corresponding measurement resolution (Å units).
- 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 (or) linear formulae:

$$W_{\text{Hyperbolic}} = \frac{\max \mathrm{R}_{1\ldots\mathrm{n}}}{R_i} \quad W_{\text{Linear}} = (\max \mathrm{R}_{1\ldots\mathrm{n}} + 1) - R_i$$

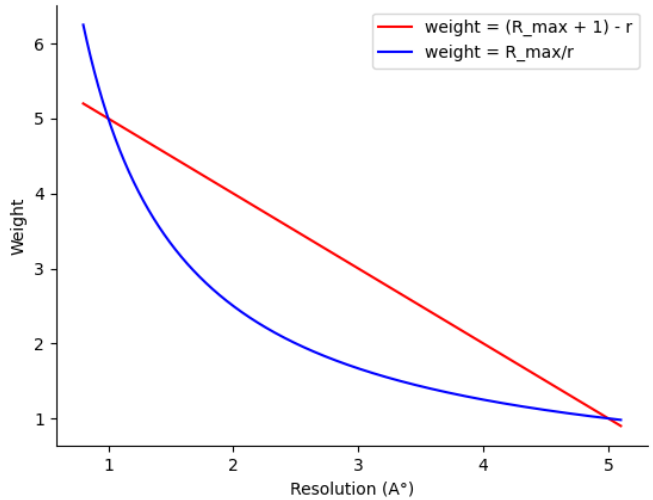


Figure: Weight calculation formulae.

Feature Family Correlations

- Features can be divided into families.
- Important ones are - AUTOCORR2d_, Chi, EState_VSA, PEOE_VSA, SMR_VSA, SlogP_VSA, VSA_EState, and fr_.
- Within Chi and AUTOCORR2d_, the features are correlated.
- ML models need to take into account this issue.

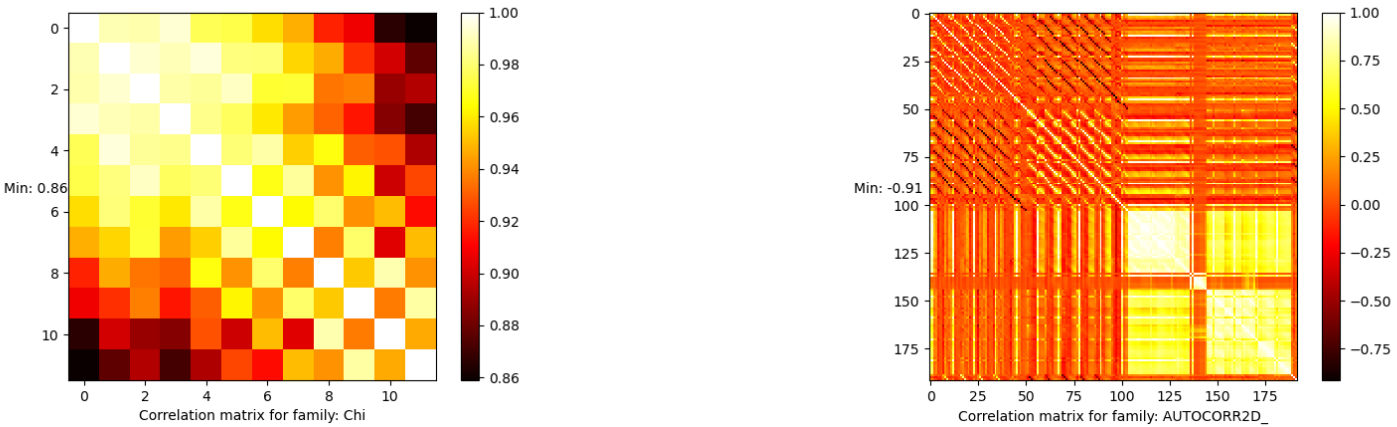


Figure: Correlation Heat Map.

Testing strategy

For every execution, we report the **random seed** used in it. This random seed can be used as a script argument (Execution ID) to reproduce the exact results. To determine the result quality we use:

- **R^2 score** (*Coefficient of determination*): $R^2 \in (-\infty, 1.0]$ where 1.0 is the best score.
- **Visualization:** The model's prediction is visualized as a 2D scatter plot. The best plot is $y = x$ line which corresponds to the best R^2 score of 1.0.

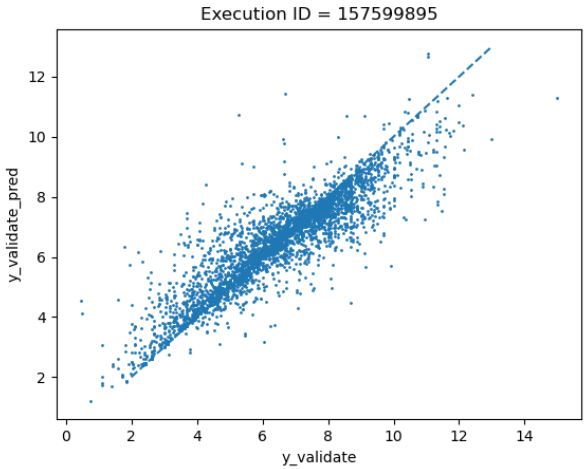


Figure: (Sample) Visualizing accuracy. $R^2 \approx 0.805$.

Machine Learning models

The ML model should approximate the following function:

$$\text{Binding affinity prediction} : \mathbf{R}^n \mapsto \mathbf{R} \text{ where } n \in \mathbf{I}^+$$

The following ML models were studied

- Simple Linear Regression
- Random Forest Regression
- Support Vector Regression
- Rotation Forest Regression

Also Note:

- DNNs cold not be used due to lack of data.
- The project only had $\approx$ 16000 data points to train.
- A simple DNN a model of size $[457, 20, 10, 1]$ has 9350 parameters.
- The DNN would overfit drastically.

Simple Linear Regression

- This model approximates the binding affinity using a linear hyperplane. (By minimizing the square of errors).
- It is the cheapest computational model.
- Assumes strong linear relationship between input features and binding affinity.
- Genetic algorithms sucessfully used for feature selection.

No. features	Feature selection	Weighting	Training	Validation	Testing
457	-	-	0.461	0.415	0.320
457	-	Hyperbolic	0.454	0.427	0.337
457	-	Hyperbolic duplication	0.465	0.416	0.326
457	-	Linear	0.458	0.419	0.328
457	-	Linear Duplication	0.460	0.428	0.327
49	**Genetic**	**Hyperbolic**	**≈0.377**	**≈0.374**	**≈0.364**
40	Pearson Correlation	Hyperbolic	0.287	0.278	0.285
40	Spearman Correlation	Hyperbolic	0.289	0.294	0.290
176	Manual	Hyperbolic	0.362	0.346	0.331

Table: R^2 scores of the Linear Regression Model.

Random Forest Regression

Support Vector Regression

Rotation Forest Regression

Discussion

Notable points:

- Best models: Linear Regression and Random Forest Regression.
- RF uses correlated features to make itself more robust.
- RF can deal with both discrete and real valued features.

Limitations:

- Linear regression assumes data linearity.
- RF has heavy reliance on ligand features.
- Both models were black box models.
- Testing results were sometimes better than validation results. This is because test data $<$ validation data. But the difference is minimal.

Discussion

Further work:

- A new weighting strategy: Weighting a pocket descriptor based on the overlap between the pocket and the ligand. For example,

$$W_{\text{Total}} = W_{\text{Hyperbolic}} * W_{\text{Overlap}}$$

- Improvement of feature selection: Build 1 model per family of features. Use the best feature as a family surrogate.
- A more explainable model can be built.

Q & A

References

- Du, Li, Xia, Ai, Liang, Sang, Ji and Liu; Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods (2016)
- Le Guilloux, Schmidtke, and Tuffery; Fpocket: An open source platform for ligand pocket detection(2009)
- DiMasi, Grabowski and Hansen; nnovation in the pharmaceutical industry: New estimates of R & D costs (2016)
- John H. Holland. Genetic Algorithms. (1960)
- Is rotation forest the best classifier for problems with continuous features? A. Bagnall, M. Flynn, J. Large, J. Line, A. Bostrom, and G. Cawley (2020)

Appendix

- 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%.