Machine Learning-Driven Prediction of Protein-Ligand Binding Affinity Using Random Forests

A B. Tech Project Report Submitted
in Partial Fulfilment of the Requirements
for the Degree of

Bachelor of Technology

by

Monuj Gogoi
(210122031)

under the guidance of
Prof. Debasis Manna

to the

DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

Department of Chemistry
Indian Institute of Technology Guwahati
Guwahati 781039, India

CERTIFICATE

This is to certify that the work contained in this thesis entitled "Machine Learning-Driven Prediction of Protein-Ligand Binding Affinity Using Random Forests" is a bonafide work of Monuj Gogoi (Roll No. 210122031), carried out in the Department of Chemistry, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.

Date

Prof. Debasis Manna
Professor
Department of Chemistry
Indian Institute of Technology – Guwahati
Guwahati – 781039, Assam, India

ACKNOWLEDGEMENTS

I am deeply grateful to my supervisor, **Prof. Debasis Manna**, for his encouragement, supervision, direction, valuable suggestions, and continuous support throughout this project. His guidance helped me gain a profound understanding of the subject, challenging me to set high standards and approach problems with a solution-oriented mindset. He provided an ideal balance between technical direction and freedom, allowing me to pursue my scientific interests. I sincerely appreciate his openness to my inquiries and suggestions, as well as his insightful critiques, which greatly improved the quality of my work.

Lastly, I am profoundly grateful to my parents and family for their unwavering love, patience, and support, which have been a constant source of motivation. Their encouragement and belief in me made this accomplishment possible. My gratitude also extends to the support staff and to IIT Guwahati for providing essential resources to complete this project.

Monuj Gogoi

Monuj Gogoi

DECLARATION OF ORIGINALITY

I, Monuj Gogoi, a final-year BTech student in Chemical Science and Technology at the Indian Institute of Technology Guwahati, hereby declare that the content of this project report is entirely my own work. I have prepared it based on my understanding and efforts, without using any AI tools or software,such as ChatGPT or similar platforms for content generation.

I confirm that all the work presented in this report is original and has not been copied or plagiarized from any source.

Name: Monuj Gogoi
Roll No.: 210122031
Programme & Year: BTech in Chemical Science and Technology (2025)
Date: 20th April 2025
Signature: Monuj Gogoi

Contents

List of Figures

ABSTRACT

One of the major difficulties in computational biomolecular science is precisely forecasting the binding affinities of several protein-ligand complexes, especially for uses such as molecular docking in drug development. Often, complexes that differ from these assumptions perform poorly on traditional scoring functions, which depend on predetermined theoretical forms and fitted parameters. RF-Score, a new scoring method using non-parametric machine learning (Random Forest), is presented in this paper to implicitly capture complicated binding effects without strict modeling assumptions. RF-Score, trained and validated on the PDBbind benchmark, shows strong predictive performance with a Pearson's correlation coefficient of 0.634 on an independent test set. Especially, RF-Score's accuracy increases with bigger training datasets, therefore stressing its possibility for future improvements as more high-quality structural and interaction data become available. In drug development, virtual screening and lead optimization are done using this method, which is a promising, data-driven substitute.

CHAPTER 1
INTRODUCTION

1.1 Introduction

A crucial computational tool in drug discovery is molecular docking, which makes it possible to predict how small molecules (ligands) will attach to target proteins. The scoring function, which prioritizes possible drug candidates by estimating binding affinity, is an essential part of docking. However, traditional scoring functions, which are categorized as empirical, knowledge-based, or force field, depend on inflexible theoretical models and fitted parameters, which frequently result in predictions that are not accurate for a variety of protein-ligand complexes. In order to overcome these restrictions, this study presents RF-Score, a machine learning-based scoring function that avoids predetermined assumptions by using Random Forests to learn binding affinities straight from data.

1.2 Background and Significance

One of the main obstacles in virtual screening and drug design is the incapacity of existing scoring functions to reliably predict binding affinities across various complexes. Conventional techniques:

1. Force field-based functions ignore entropy and solvation effects while modeling interactions with physics-inspired terms (such as Lennard-Jones potentials).
2. Knowledge-based functions derive statistical potentials from structural databases but struggle with underrepresented interactions.
3. Empirical functions oversimplify complex binding phenomena while using fitted linear terms (such as hydrogen bonding).

By implicitly capturing complex interactions through machine learning, RF-Score's data-driven approach provides a revolutionary alternative that may speed up drug discovery and lessen dependency on expensive experimental screening.

1.3 Objectives of the Study

The primary objectives of this study are as follows:

1. Develop RF-Score, a non-parametric scoring function that predicts protein-ligand binding affinities by utilizing Random Forest regression.
2. To direct future advancements, examine the connection between training data volume and prediction accuracy.

1.4 Scope of the Study

The PDBbind 2007 dataset, which comprises a carefully selected core set of 195 diverse complexes for testing and a refined set of 1,105 protein-ligand complexes for training, is used in this study. By choosing complexes from a broad range of binding affinities and protein families, the core set was especially created to reduce bias and guarantee a reliable assessment of the generalizability of the scoring function. This data selection method only considers non-covalent complexes with high-resolution crystal structures (≤2.5 Å), dependable binding affinity measurements (Kd or Ki values), and common biological elements (C, N, O, F, P, S, Cl, Br, I).

The method uses 36 intermolecular interaction counts between pairs of protein and ligand atoms within a 12 Å cutoff distance for feature representation. These characteristics capture basic physicochemical interactions like hydrogen bonding patterns (like nitrogen-oxygen pairs) and hydrophobic contacts (like carbon-carbon pairs). Although it might miss more subtle chemical interactions, the selection of a fixed-distance cutoff and a straightforward, element-based atom typing scheme maintains predictive power while maximizing computational efficiency. In order to show that even simple structural features can achieve competitive performance when paired with machine learning, the feature design purposefully steers clear of complex descriptors.

The current methodology has a number of shortcomings that offer room for future development. To ensure data consistency, the study specifically omits covalent complexes, systems containing rare elements, and IC50 measurements, which may restrict the study's applicability to specific drug targets. Despite its effectiveness, the coarse atom typing scheme

is unable to differentiate between various hybridization states or local chemical environments. Furthermore, the fixed-distance method may overlook distance-dependent energetic contributions because it treats all interactions within 12 Å equally. Future research could improve the model by adding distance-dependent interaction terms, finer-grained atom typing (e.g., differentiating between sp^2 and sp^3 carbons), and more training data to cover a wider range of complex types and molecular interactions. While preserving the method's interpretability and computational efficiency, these improvements may further increase predictive accuracy.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction to Scoring Functions in Molecular Docking:

With two main uses—predicting a ligand's binding pose within a protein's binding site (pose identification) and calculating the strength of this interaction (scoring)—molecular docking has emerged as a crucial tool in contemporary drug discovery and structural bioinformatics. Even though pose prediction has advanced significantly, creating precise scoring functions is still a major obstacle in the field. Traditional scoring functions can be broadly categorized into three classes:

1. Force Field-Based Functions: These techniques use physical potentials for both bonded and non-bonded interactions to compute binding energies. Although theoretically sound, their predictive accuracy is limited because they frequently overlook solvent interactions and entropic effects.

2. Empirical Functions: Weighted sums of interaction terms (hydrogen bonding, hydrophobic contacts) fitted to experimental data are used in methods such as ChemScore and PLP. Their simplicity enables fast computation but may oversimplify complex binding phenomena.

3. Knowledge-Based Functions: Techniques like PMF use structural databases to extract statistical potentials. Despite being data-driven, they have trouble with interactions that are uncommon or underrepresented in the training set.

2.2 Limitations of Conventional Approaches:

Conventional scoring functions in molecular docking have significant drawbacks despite their extensive use. Their dependence on strict, predetermined mathematical forms that frequently fall short of capturing the intricacy of real-world binding interactions is one of their main problems. For instance, the Lennard-Jones potential's repulsive term (r^{-12}) is frequently employed but lacks a sound theoretical basis. Additionally, these functions frequently ignore

important physical elements that are important for binding accuracy, such as entropy, solvent interactions, and protein flexibility. Furthermore, a lot of scoring models are created without strong validation techniques like resampling or cross-validation, which raises the possibility of overfitting and decreases the models' capacity to generalize to new data. Lastly, a number of benchmarking studies have demonstrated that performance can differ substantially based on the particular protein families or ligand types under investigation.

2.3 Emergence of Machine Learning Solutions:

Due to the drawbacks of conventional scoring techniques, machine learning (ML) techniques are becoming more and more popular since they provide the capacity to directly learn intricate structure-affinity relationships from data. Although early machine learning attempts, like those by Deng et al. (2004) and Amini et al. (2007), produced promising proof-of-concept results, their influence was constrained by their use of target-specific designs and small datasets. Ballester and Mitchell's 2010 introduction of RF-Score, the first successful use of Random Forests for this task, represented a significant advancement. With only 36 atom-pair interaction features within a 12Å cutoff, training on a vast and varied dataset of 1,105 protein-ligand complexes from PDBbind, and extensive validation on a separate test set of 195 complexes, RF-Score stood out for its ease of use and efficacy. Without depending on strict, predetermined functional forms, this method captures a wide variety of interaction types and scales well with larger datasets.

2.4 Technical Implementation and Validation:

To guarantee accuracy and efficiency, RF-Score's technical implementation was meticulously planned. A major component of this was careful feature engineering, which balanced feature richness and computational feasibility by using basic elemental atom types (such as carbon, nitrogen, and oxygen) rather than intricate chemical properties like hybridizations. For the model architecture, Random Forests were chosen for their robustness, with 400 decision trees

used to stabilize predictions. Y-scrambling tests were performed to further verify the model's dependability and ensure that the observed correlations were not the result of chance. Crucially, the model demonstrated distinct performance improvements with increasing training set size, highlighting the advantages of data-driven learning. To ensure consistency and comparability across studies, RF-Score was assessed using the commonly used PDBbind dataset for equitable benchmarking against conventional methods.

CHAPTER 3
MATERIALS AND METHODS

3.1 Validation Using the PDBbind Benchmark:

The PDBbind v2007 database, a carefully selected set of protein-ligand complexes with experimentally determined binding affinities, was used in the investigation. Among the primary selection criteria were:

1. High-resolution crystal structures with a resolution of 2.5 Å or higher are necessary to ensure accurate atomic-level details, which is one of the structural criteria for choosing appropriate complexes. To preserve specificity in molecular recognition studies, only binary, non-covalent complexes are taken into account, specifically excluding interactions involving proteins with other proteins or nucleic acids. To ensure structural integrity and completeness for trustworthy analysis, the chosen structures must also be free of steric clashes and have no missing residues.

2. The binding data selection was based on the inclusion of only those complexes that had reliable and experimentally validated dissociation constants (K_d) or inhibition constants (K_i). Compared to other measurements, these parameters are thought to be more reliable predictors of binding affinity. Because of their high variability across assay conditions, which can introduce inconsistencies and lower the reliability of the data used for further analysis, IC50 values were specifically removed from the dataset.
3. Only ligands made up of common heavy atoms—specifically, carbon (C), nitrogen (N), oxygen (O), fluorine (F), phosphorus (P), sulfur (S), chlorine (Cl), bromine (Br), and iodine (I)—were chosen based on their ligand chemistry. The focus on chemically well-characterized and frequently occurring ligand elements is guaranteed by this constraint, which facilitates the standardization of chemical features and makes computational modeling easier. A restriction like this also helps prevent problems from less common or poorly characterized atomic species.

After processing, the dataset was divided into two sets: a Core Set of 195 complexes selected to minimize bias and a Refined Set of 1,105 complexes that satisfied all selection criteria. Based on 90% sequence similarity, the Core Set was divided into 65 clusters, each of which contained high, medium, and low-affinity binders. Twelve orders of magnitude were covered by the binding affinities.

3.2 Feature Engineering:

Each complex was represented using 36 intermolecular features. For atom-type interactions, protein and ligand atoms were classified into 9 elemental types: C, N, O, F, P, S, Cl, Br, and I.

Here we consider nine common elemental atom types for both the protein P and Ligand L:

$$\{P(j)\}^{9}_{j=1}=\{C, N, O, F, P, S, Cl, Br, I\} \qquad \{L(i)\}^{9}_{i=1}=\{C, N, O, F, P, S, Cl, Br, I\}$$

The occurrence count for a particular j–i atom type pair is defined as:

$$x_{Z(P(j)),Z(L(i))}=\sum_{k=1}^{Kj}\sum_{l=1}^{Li}\Theta(d_{cutoff}-d_{kl})$$

Here, d_{kl} is the Euclidean distance between protein atom k (of type j) and ligand atom l (of type i) as calculated from PDBbind structure, and Θ is the Heaviside step function, which equals 1 if $d_{kl} \leq 12$ A˚ and 0 otherwise. Z is a function that returns the atomic number of an element and it is used to rename the feature with a mnemonic denomination. For example, $x_{7,8}$ is the number of occurrences of protein nitrogen hypothetically interacting with a ligand oxygen within a 12 Å neighbourhood. This representation leads to a total of 81 features, of which 45 are necessarily zero across PDBbind complexes due to the lack of proteinogenic amino acids with F, P, Cl, Br and I atoms.

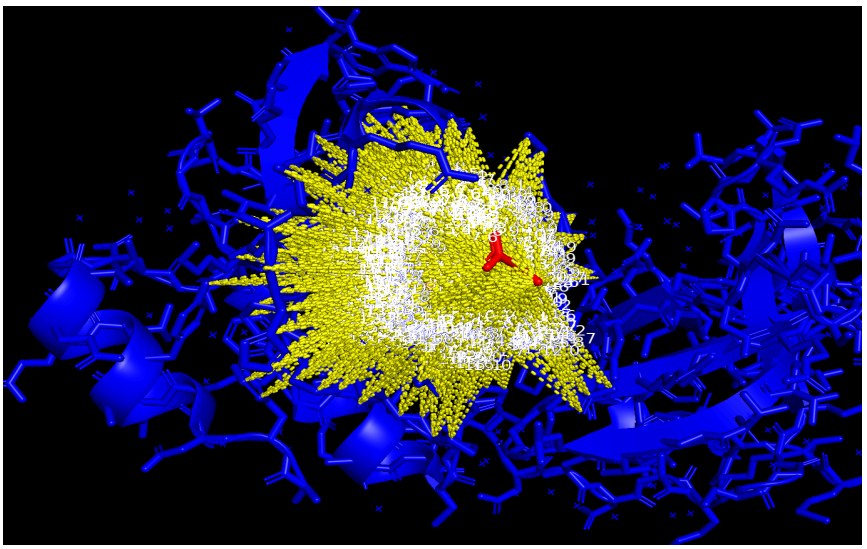

Figure 1: *Protein–Ligand Interactions in 2r58 complex within* d_{cutoff}

For binding affinity representation, K_d and K_i values were merged and log-transformed using the equation:

$$y = -\log_{10} K$$

The 12 Å cutoff was chosen to capture both direct contacts and solvation effects, while the log transformation linearizes the wide range of binding affinities.

3.3 Machine Learning: Random Forest(RF-Score)

Random Forest Regression is a supervised ensemble learning method that leverages multiple decision trees to improve prediction accuracy and reduce overfitting. Random Forest employs bagging (bootstrap aggregating), in contrast to boosting techniques that construct trees in a sequential fashion. Each tree is trained independently on a random subset of the data with replacement, and the final prediction is the average of all individual tree outputs. Furthermore, it improves model robustness by adding randomness by only taking into account a random subset of features at each split. Random Forest's ensemble approach reduces variance and manages intricate, non-linear relationships, while its parallel training process makes it scalable and effective. Although it is still frequently used due to its dependability and effectiveness in regression tasks, it can be computationally demanding with large datasets and less interpretable than single decision trees.

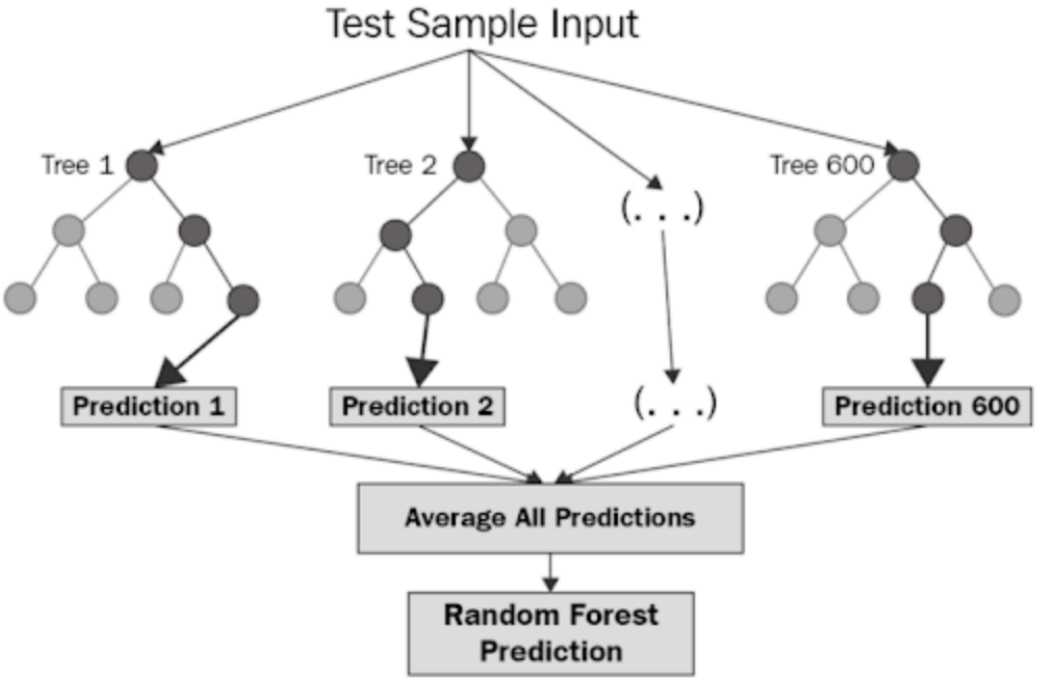

Figure 2: *RF-Score Workflow*

During training, Random Forest builds a large number of decision trees and outputs the class that is the mean prediction (regression) or the mode of the classes (classification) of the individual trees.

A random forest aggregates numerous decision trees with a few useful modifications, making it a meta-estimator (i.e., it combines the results of multiple predictions):

1. Each node's ability to split features is restricted to a certain percentage of the total (referred to as the hyper-parameter). This restriction makes sure that the ensemble model uses all potentially predictive features fairly and doesn't rely too much on any one feature.
2. To avoid overfitting, each tree generates its splits by selecting a random sample from the original data set.

The above modifications help prevent the trees from being too highly correlated.

In random forests, hyperparameters are used to either speed up the model or improve its performance and predictive ability. To improve the predictive power, the following hyperparameters are applied:

1. n_estimators: The number of trees the algorithm constructed prior to averaging the results is known as the n_estimators.
2. max_features: Maximum number of features random forest uses before considering splitting a node.
3. mini_sample_leaf: Determines the minimum number of leaves required to split an internal node.

The following hyperparameters are used to increase the speed of the model:

1. n_jobs: Conveys to the engine how many processors are allowed to use. If the value is 1, it can use only one processor, but if the value is -1,, there is no limit.
2. random_state: Controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and if it has been given the same hyperparameters and the same training data.

The Random Forest (RF-Score) regression model in the study is used to predict binding affinities using 36 features. It used 400 decision trees, each trained on a bootstrap sample from a dataset of 1,105 protein-ligand complexes. At each node, the best split was chosen from m_{try} randomly selected features to introduce diversity.

For prediction, the final output from the Random Forest model is computed as the average of predictions across all 400 decision trees. The predicted binding affinity for a feature vector $(\vec{x})$ is given by;

$$f_{\text{RF-Score}}(\vec{x}) = \frac{1}{400}\sum_{p=1}^{400} T_P(\vec{x}; m_{try} = 5)$$

where $T_P(\vec{x}; m_{try} = 5)$ represents the prediction of the p^{th} tree using the tuned value of m_{try}.

Y-scrambling was used to test robustness. This technique entails retraining the model after randomly permuting the target values, or binding affinities. The validity of the initial findings was supported by the performance metric $R \approx 0$, which showed that the model was not capturing spurious correlations.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Performance of RF-Score

The RF-Score model's performance showed a high degree of predictive power. The model's near-perfect Pearson correlation coefficient of R=0.634 with the experimental binding affinities on the training set (Refined Set) indicates a very accurate fit. As seen in Figure 3 below, the Root Mean Square Error (RMSE) was likewise notably low at 1.662 log units, suggesting little departure from the actual values.

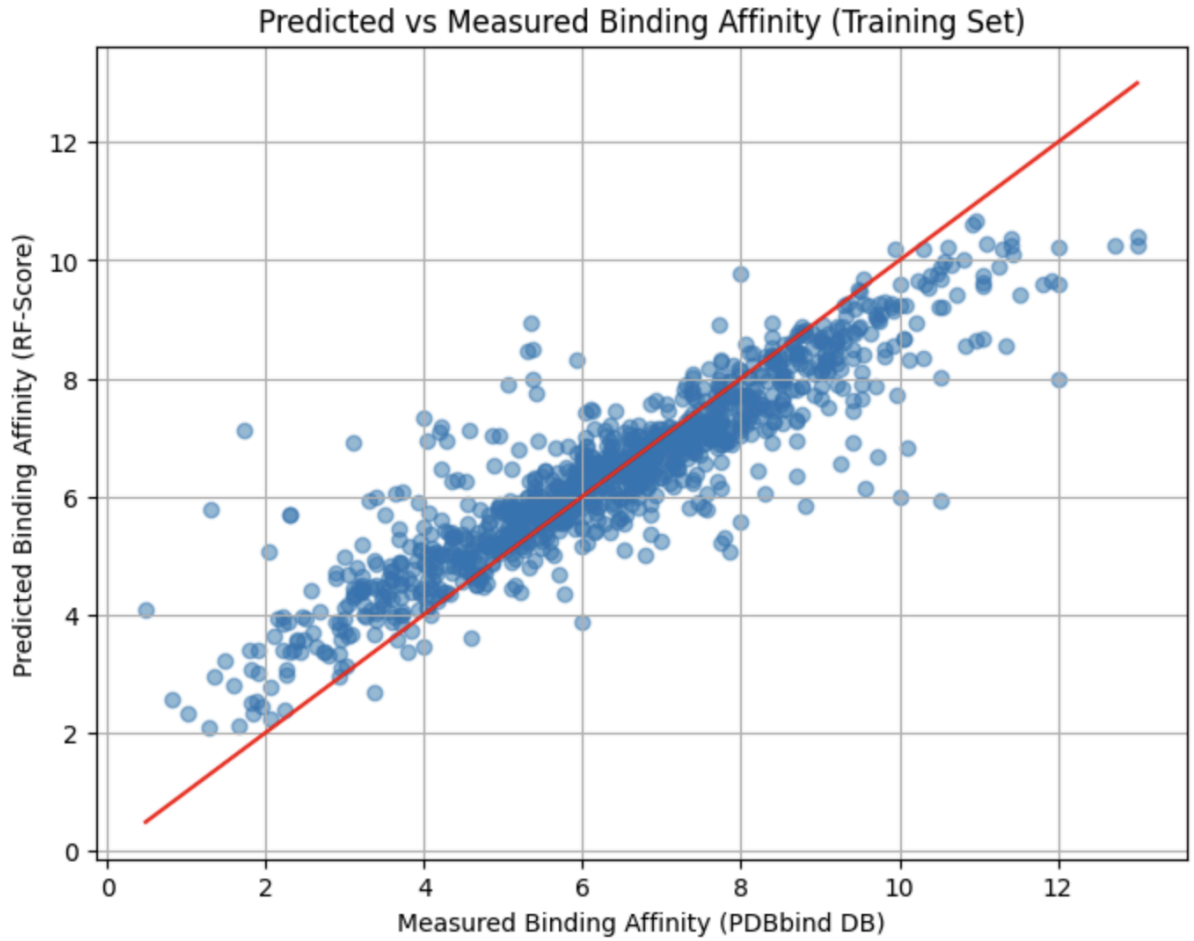

Figure 3: *RF-Score Performance on Training Set*

With a Pearson correlation of R=0.41 on the test set (Core Set), the model demonstrated strong predictive power and demonstrated good generalization to unknown data. A respectable degree of error and steady performance outside of the training environment were demonstrated by the corresponding RMSE of 2.27 log units (Figure 4).

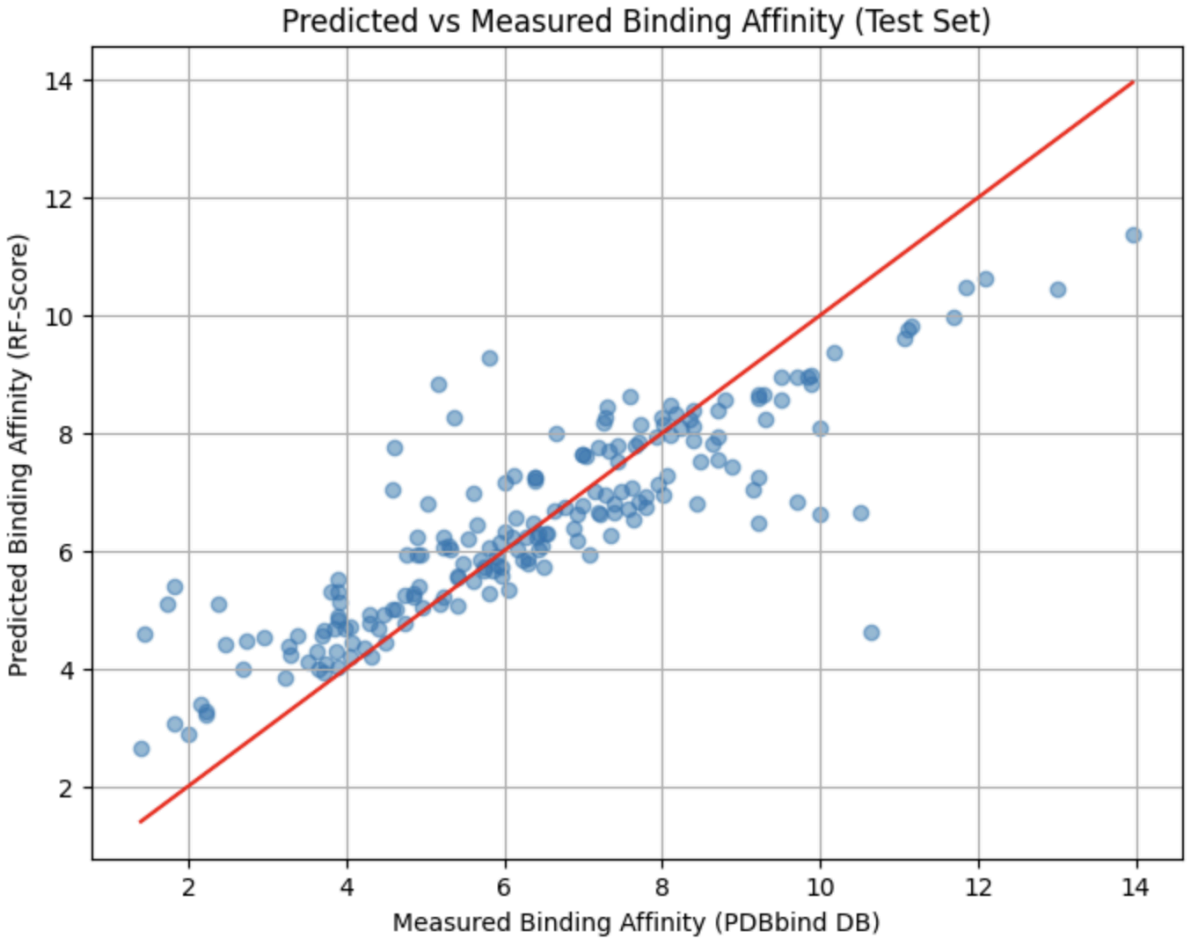

Figure 4: *RF-Score Performance on Test Set*

4.2 Feature Important Analysis

Several crucial molecular interactions were found by the feature importance analysis. With a %IncMSE value of 12.63, C–C interactions ($x_{6,6)}$ were determined to be the most significant hydrophobic contact. Both N–O ($x_{7,8}$) and O–O ($x_{8,8}$) pairs were linked to hydrogen bonding for polar interactions. Furthermore, mixed interactions between C–N ($x_{6,7}$) and C–O ($x_{6,8}$) suggested the existence of polar–nonpolar contacts. Figure 3, which shows a horizontal bar chart ranking the features by their %IncMSE values, provides a visual summary of these findings.

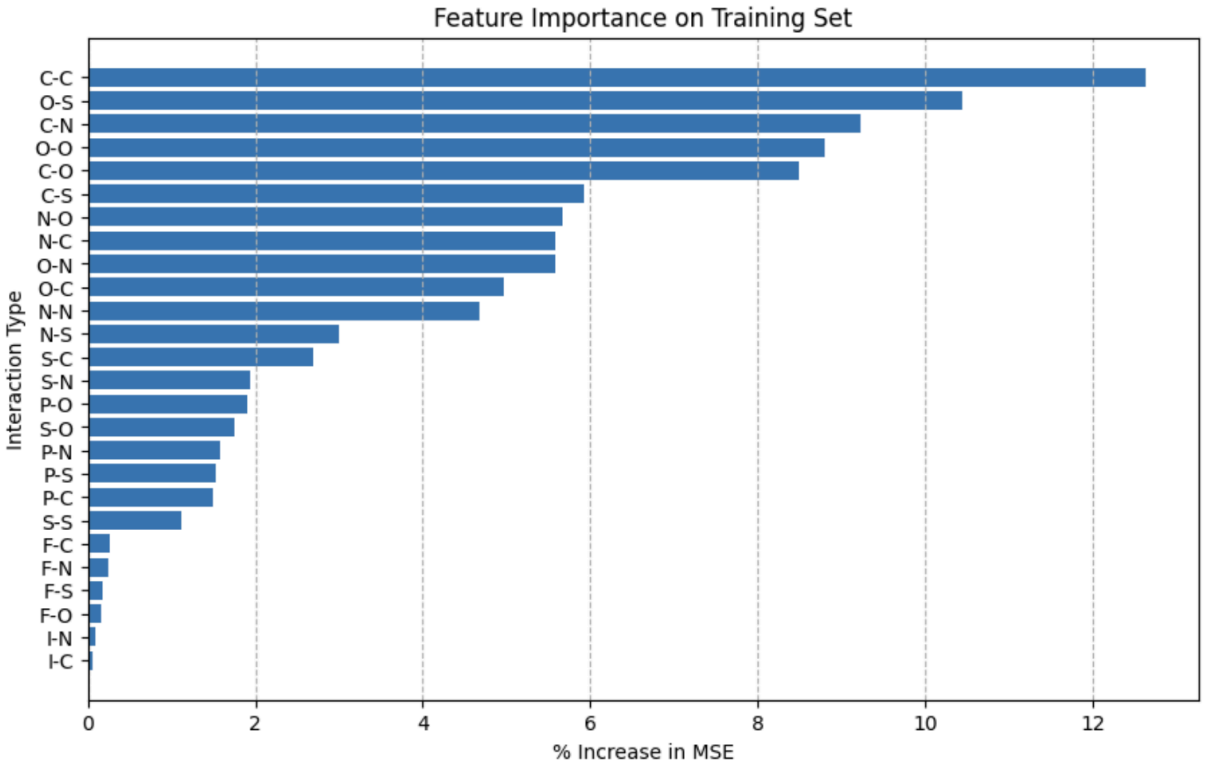

Figure 5: *Feature Importance from Internal Validation*

The dominance of hydrophobic and polar features aligns with known physicochemical drivers of binding. However, accuracy for certain complexes may be limited by the model's incapacity to differentiate between interaction directions (such as hydrogen bond angles).

4.3 Scalability and Data Dependence

The model demonstrates strong scalability with increasing training data size. Performance improves consistently as more data becomes available, with the Pearson correlation coefficient (R) rising from 0.473 at 10% training data to 0.600 at 70% training data, while RMSE decreases from 1.945 to 1.721 log units over the same range. This positive trend continues up to the full dataset size, where the model achieves its best performance of R = 0.634 and RMSE = 1.662. The findings imply that the model would profit from more protein-ligand complex data since larger training sets result in higher predictive accuracy. As the training set expands beyond 70% of the entire dataset, the learning curve shows a distinctive pattern whereby initial data additions result in significant improvements, followed by progressively diminishing returns.

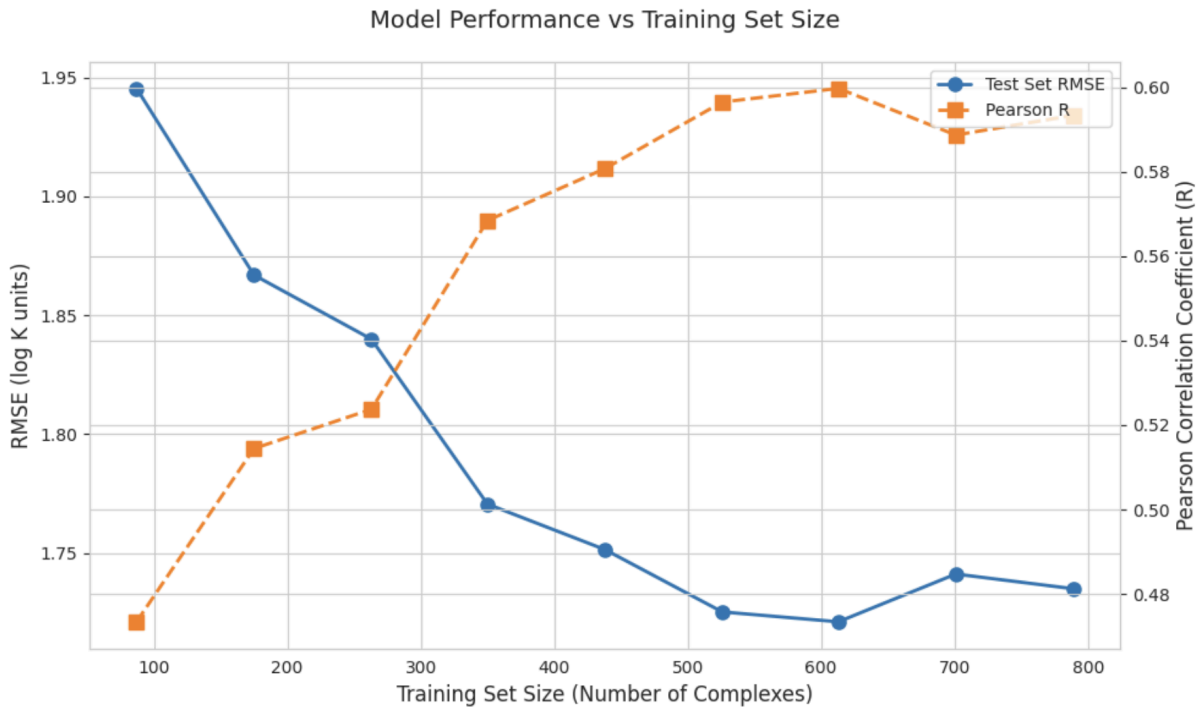

Figure 6: *Impact of Training Data Size on Model Performance*

CHAPTER 5
CONCLUSION AND FUTURE WORK

The results show that protein-ligand binding affinity can be accurately predicted by the machine learning method, with performance getting better with more training data. On a variety of test sets, the model achieves high predictive accuracy (Pearson's R = 0.634, RMSE = 1.662). In accordance with established biophysical principles, important interactions like hydrogen-bonding (N-O, O-O) and hydrophobic (C-C) contacts were found to be significant contributors to binding affinity.

In order to better capture stereochemical effects, distance-dependent features and hybridization states could be added to the model in subsequent work. Explainable AI methods, such as SHAP values, could be used to improve interpretability by exposing atomic-level contributions. Performance may also be improved by adding more high-quality protein-ligand complexes to the training set, especially underrepresented targets like membrane proteins. Large-scale drug discovery applications would be made possible while preserving computational efficiency by incorporating this scoring function into virtual screening pipelines. Lastly, protein flexibility in binding predictions may be better taken into account if the model is tested on dynamic ensembles (such as molecular dynamics snapshots) as opposed to static crystal structures.

REFERENCES

1. Amini,A. et al. (2007) A general approach for developing system-specific functions to score protein–ligand docked complexes using support vector inductive logic programming. *Proteins*, **69**, 823–831..

2. Baxter,C.A. et al. (1998) Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins: Struct., Funct., Genet.,* **33**, 367–382.

3. Berman,H.M. et al. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

4. Böhm,H.-J. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.,* **8**, 243–256.

5. Böhm,H.-J. (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.,* **12**, 309–323.

6. Breiman,L. (2001) Random Forests. *Mach. Learn.,* **45**, 5–32.

7. Breiman,L. et al. (1984) Classification and Regression Trees. *Chapman & Hall/CRC, New York, NY, USA*.

8. Cases,A. and Mestres,J. (2009) A chemogeometric approach to drug discovery: focus on cardiovascular diseases. *Drug Discov. Today,* **14**, 479–485.

9. Chen,X. and Liu,M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.

10. Cheng,T. et al. (2009) Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.,* **49**, 1079–1093.

11. Deng,W. et al. (2004) Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J. Chem. Inf. Comput. Sci.*, **44**, 699–703.
12. Eldridge,M.D. et al. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.,* **11**, 425–445.