Improving interpretable machine-learning models for predicting protein–ligand binding affinity

Ansh Meshram CS22B051[1]

1 Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

Abstract

Protein–ligand binding affinity prediction is a fundamental task in structure-based drug discovery. While deep learning models have demonstrated state-of-the-art predictive performance, they often demand extensive computational resources and large volumes of high-quality data—challenges that can limit their practical applicability and interpretability. This project focuses on developing and evaluating interpretable machine learning models that maintain competitive predictive accuracy while providing clear insights into the underlying decision-making process.

The study explores a suite of models grounded in Intermolecular Contact Profiles (IMCPs) [1]and enhances them with environmental descriptors that capture the local atomic context of both protein and ligand atoms. To evaluate the effectiveness of our proposed models, we compare their performance against several state-of-the-art techniques, including those that incorporate advanced descriptors such as Morgan fingerprints [2] and embeddings derived from the Evolutionary Scale Modeling (ESM) [3] protein language model. We also benchmark our results against established models including Pafnucy , GraphDTA and DataDTA . These comparisons provide a clear context for assessing the predictive accuracy and practical relevance of our interpretable models.

This work demonstrates the potential of combining feature-driven modeling with biochemically meaningful descriptors to produce models that are not only accurate but also transparent and scalable. By preserving interpretability, these models offer valuable biological insights, supporting rational molecular design and facilitating their integration into early-stage drug discovery pipelines.

Introduction

Protein–ligand binding affinity prediction is a foundational challenge in structure-based drug discovery, where accurate estimation of interaction strength between a target protein and a small-molecule ligand can dramatically accelerate lead identification and optimization processes. Traditional high-throughput screening methods, while effective, are costly and time-consuming, often requiring years and substantial financial investment to yield viable drug candidates. Three-dimensional structural information gathered from biological targets are a prominent component of modern medicinal chemistry. Binding affinity prediction (BAP) is one of the key problems in structure-based drug design (SBDD). BAP is generally oriented toward putative protein-ligand complexes, and aims to create a bridge between the structures of these complexes and the binding affinities in them.

Various descriptors like Intermolecular Contacts (IMC) [4] , Intermolecular Contact Profile (IMCP) [1], Intermolecular Contact in Distance Bins (IMCiDB) [5], Intermolecular Contact Profiles in Distance Bins (IMCPiDB) [1] Extended connectivity interaction features (ECIF) [6], Structural protein-ligand interaction fingerprints (SPLIFs) [7]and various models like Random Foreset , Gradient Boost Decision Treee, Covulation Neural Networks to more classical methods like Molecular Docking methods [8] have been employed. But they are either to complex or they lack sufficient binding affinity accuracy.

Recent advances in deep learning—leveraging graph-based molecular representations, attention mechanisms, and multiscale embeddings—have achieved remarkable predictive accuracy for binding affinities, enabling virtual screening at unprecedented scale. However, these models demand extensive computational resources and large, high-quality datasets, and their "black-box" nature limits biological interpretability and hinders adoption in regulated settings.

In this work, we explore a suite of models that extend the traditional IMCP [1] descriptor by incorporating the chemical environment of interacting atoms, inspired by approaches such as ECIF [6]. By enriching IMCPs with chemically meaningful features—such as local atomic density, connectivity patterns, and atom-type context—we aim to develop models that are not only more interpretable and chemically grounded, but also achieve improved predictive accuracy. We also explored the application of Principal Component Analysis (PCA) as a dimensionality reduction technique to mitigate the effects of high-dimensional feature spaces. Our goal is to construct models that strike a balance between performance and interpretability, offering a viable alternative to state-of-the-art deep learning methods in protein–ligand binding affinity prediction

Materials and methods

Data Acquisition and Preparation

1. **Datasets:** We utilized the PDBbind v2020 refined set and CASF-2016 dataset:
   - *PDBbind v2020 refined* – contains high-quality protein–ligand complexes with experimentally measured binding affinities.
   - *CASF-2016 dataset* – a standard benchmark for evaluating scoring functions, focusing on assessing their predictive accuracy independently of docking processes

2. **Ligand Conversion:** All ligand mol2 files were converted to PDB format using UCSF Chimera to ensure consistent parsing of coordinates.

3. **Dataset Consolidation:** The refined set and CASF-2016 dataset were combined into a unified `data/` directory for downstream processing.

4. **File Structure:** Final data was organized as follows:

```
PDBbind/
   data/        % protein{ligand complex subfolders
       {complex\_id}/
           protein.pdb
           ligand.mol2
   index/
        % train/test split CSVs mapping complex IDs to -log(K\_d)
       train.csv
       test.csv
model notebooks
```

Feature Engineering

Features in model were inspired by the IMCP framework and variations thereof, as well as advanced embeddings:

1. **Interaction Distance Binning:** For each protein–ligand complex, atom pairs within a maximum interaction cutoff of 12 Å were grouped into distance bins. These are the different types of grouping that we used:

   $$\text{Bins} = \{(0, 12)\}, \{(0, 6), (6, 12)\}, \{(0, 4), (4, 8), (8, 12)\}.$$

   The 12 Å cutoff was adopted from prior work demonstrating optimal performance (IMCP, [1]).

2. **Local Chemical Environment (env):** For each protein atom, env is defined as the number of neighboring atoms within a 2 Å radius. We additionally consider its reciprocal 1/env as a feature for some models.

3. **Feature Sets:**

   - *Standard IMC:* Counts of interacting atom pairs per bin.
   - *Standard IMCP:* Pair counts and average interatomic distance per bin.
   - *Weighted IMCP (WIMCP):* Distance features weighted by 1/env of the protein atom.
   - *Intermolecular Contact Environmental Profiles (IMCEP):* Adds the mean 1/env of protein atom of each type of interacting pair in a bin.
   - *Intermolecular Contact Profiles in Environmental Bins (IMCPiEB):* Bins each pair by whether 1/env of the protein atom is above or below the global mean for its interaction type in a bin.
   - *Intermolecular Contact Environmental Profiles 2 (IMCEP2):* Includes the mean env of the ligand atom for each interacting atom type in addition to IMCEP features.
   - *ESM:* Concatenation of protein embeddings from the `esm2_t6_8M_UR50D` model and Morgan fingerprints of ligands.

4. **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to selected feature matrices to mitigate the curse of dimensionality and improve model generalization. Components were retained to explain 95% of variance.

Model Training

We evaluated multiple regression algorithms on each feature representation and bin scheme:

- **Algorithms:**
  - Random Forest Regressor (650 trees, fixed random seed)
  - Gradient Boosting Regressor (700 estimators)
  - Support Vector Regression (polynomial kernel, degree 6)
  - XGBoost Regressor (max depth 10, learning rate 0.1)
  - Ordinary Least Squares Linear Regression
- **Cross-Validation:** 10-fold cross-validation was performed on the training set (90% of data) to assess stability.
- **Final Evaluation:** The best model from cross-validation was tested on the held-out 10% test set.

Performance Metrics

Each model was trained to predict the binding affinity in terms of the negative logarithm of the binding affinity, denoted as $-\log(K_d)$. The predictive performance was evaluated using the following metrics:

- **Pearson Correlation Coefficient (r):** Assesses the linear correlation between the predicted values (y_{pred}) and the experimental values ($y_{\text{true}} = -\log(K_d)$), indicating how well the model preserves the relative ranking of binding affinities.
- **Root Mean Square Error (RMSE):** Measures the square root of the average squared differences between the predicted and true values. It reflects the average magnitude of prediction error in the same scale as the target variable.

Results and Discussion

Feature Correlation and Selection

We observed that the count of interacting atom pairs and their average distance are highly correlated with binding affinity. Additionally, the ligand environment and the inverse of the protein environment (1/env_protein) exhibit stronger correlations than their respective inverses. Excluding the nitrogen atom from the environmental feature calculations further improves results.

Test-Set Results

While our eXtreme Gradient Boosting (XGB)-based IMCPEB2 model narrowly outperforms earlier variants, it still lags behind several deep 3D architectures, underscoring the challenge of closing the remaining accuracy gap.

Cross-Validation Results

Although slight variations are observed, the improvements across enhanced feature sets lie within one standard deviation of the baseline IMCP model, indicating that—while environmental descriptors show promise—statistical significance remains marginal.

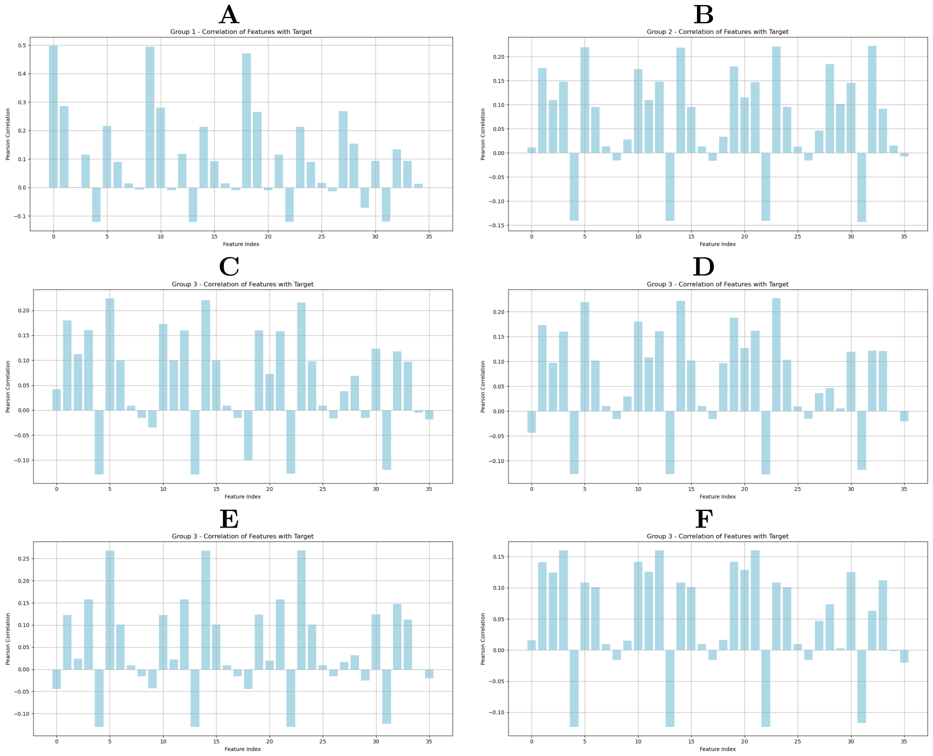

Fig 1. Correlation with binding affinity. A: Correlation of count of atom-pair B: Correlation of count of average distance of atom-pair C: Correlation of env of protein atom in atom-pair D: Correlation of inverse of env of protein atom in the atom-pair E: Correlation of env of ligand atom in atom-pair F: Correlation of inverse of env of ligand atom in the atom-pair.

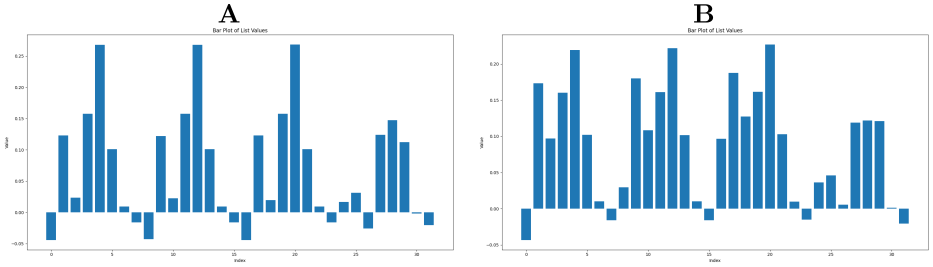

Fig 2. Improved Correlations A: Improved Correlation of ligand env B: Improved Correlation of inverse of protien env.

PCA-Based Models

Dimensionality reduction via PCA led to degraded test performance across all variants, indicating that principal components alone do not capture the most predictive subspaces of the original environmental and structural descriptors.

Table 1. Test-Set Performance and State-of-the-Art Comparison.

Model	Test Pearson	Test RMSE
IMCP	0.762	1.328
IMC	0.760	1.326
WIMCP	0.763	1.309
IMCEP	0.767	1.315
IMCPiEB	0.775	1.286
IMCEP2	0.770	1.297
Pafnucy	0.774	1.129
GraphDTA	0.660	1.193
DataDTA	0.800	1.012

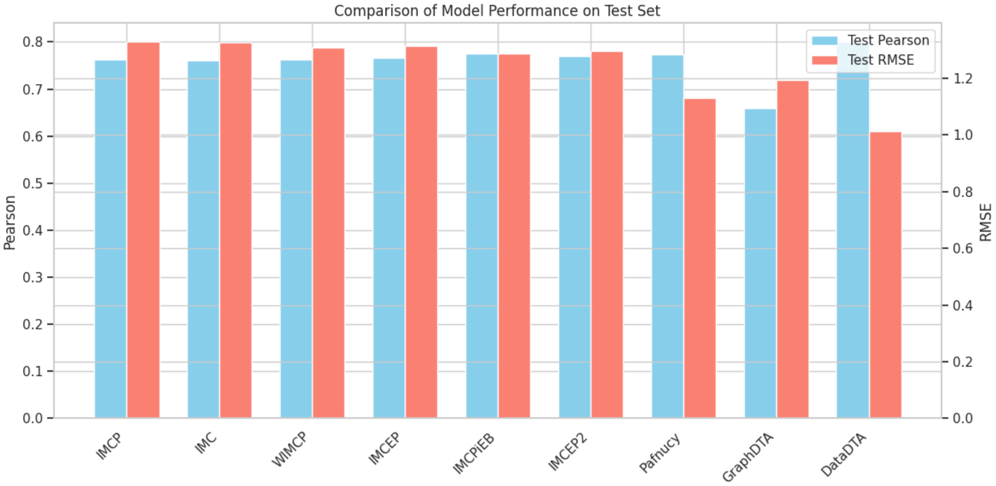

Fig 3. Comparison of all the models While our models may not exhibit the highest predictive accuracy, their performance is comparable, suggesting consistent behavior across different modeling approaches.

Conclusion

Our systematic evaluation shows that simple atomic-pair counts, their mean distances, and tailored environmental descriptors capture significant variance in binding affinity, with ligand environment and the inverse of protein environment emerging as especially informative. On the test set, our XGB-based IMCPiEB model achieved a Pearson correlation of 0.775 and RMSE of 1.286, closely rivaling methods like Pafnucy and outperforming earlier IMCP variants. Cross-validation confirmed these gains, though largely within baseline variability, indicating steady but incremental improvements. While dimensionality reduction via PCA provided limited benefit, preserving the original descriptor granularity proved crucial. Although our models do not yet surpass leading 3D deep-learning architectures, they offer interpretable, competitive alternatives grounded in chemically intuitive features. Moving forward, integrating richer spatial information and advanced model architectures should help bridge the remaining accuracy gap.

Acknowledgments

I would like to express my sincere gratitude to Professor Manikanandan Narayana for providing valuable guidance in the statistical validation of my models. I also

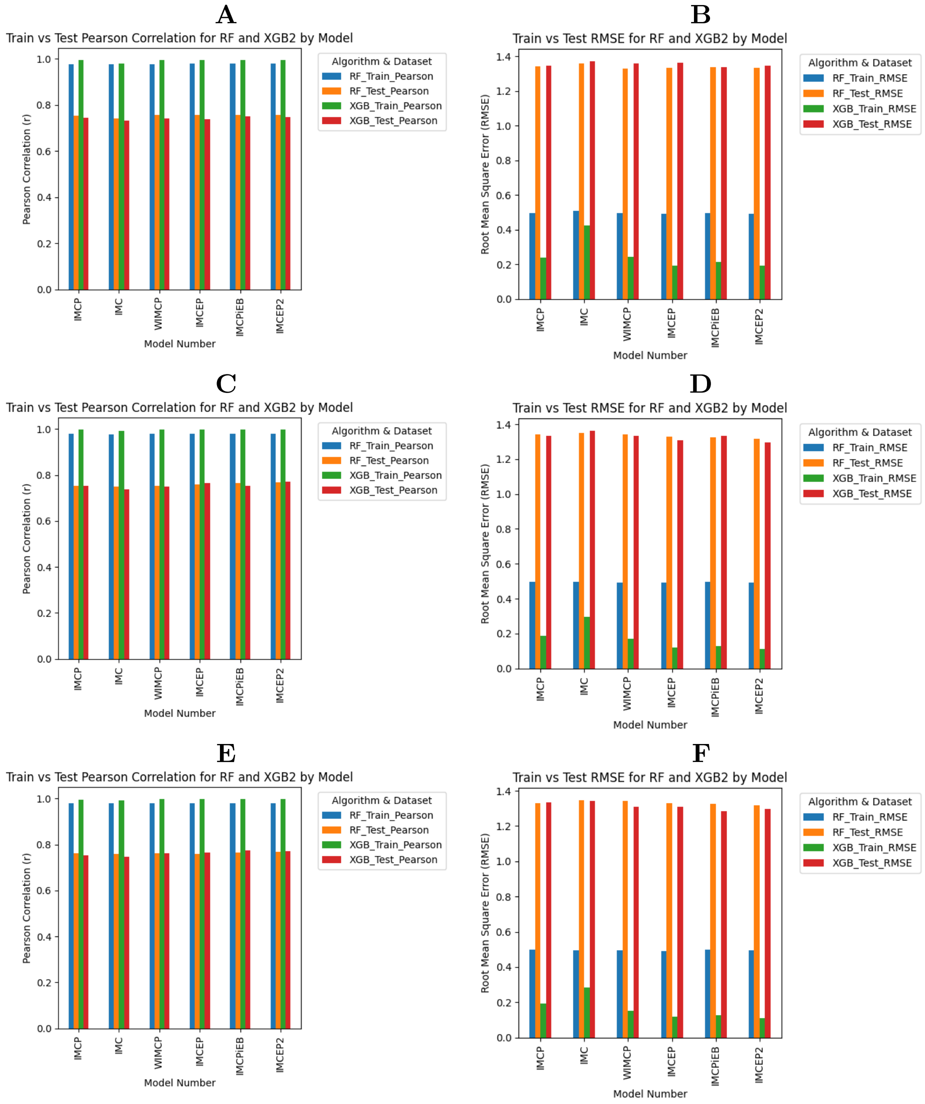

Fig 4. Result of new models A: Pearson correlation for different features in (0,12) bin B: RMSE for different features in (0,12) bin C: Pearson correlation for different features in (0,6),(6,12) bin D: RMSE for different features in (0,6),(6,12) bin E: Pearson correlation for different features in (0,4),(4,8),(8,12) bin F: RMSE for different features in (0,4),(4,8),(8,12) bin

acknowledge the use of external resources, including the UCEF Chimera tool, the PDBbind dataset, and the IMCP implementation. I am especially thankful to Debby D. Wang, the author of the IMCP paper, for making their code publicly available, which greatly supported my research.

Author Contribution

All work presented in this project, including conceptualization, data analysis, model development, evaluation, and documentation, was solely carried out by the author.

Table 2. Cross-Validation Performance (3 bins).

Model	CV Pearson (mean ± std)	CV RMSE (mean ± std)
IMCP	0.723 ± 0.021	1.351 ± 0.032
IMC	0.725 ± 0.023	1.345 ± 0.039
WIMCP	0.715 ± 0.024	1.363 ± 0.039
IMCEP 9	0.728 ± 0.023	1.343 ± 0.034
IMCPiEB	0.712 ± 0.025	1.369 ± 0.031
IMCEP2	0.719 ± 0.021	1.355 ± 0.037

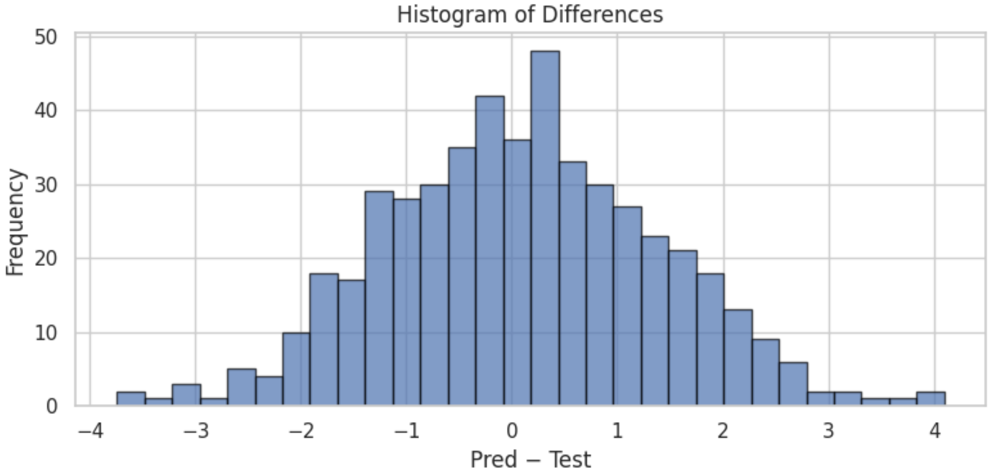

Fig 5. Difference betweeen predicted and true value The Figure shows that our model predicts binding affinity correctly for most of the complexes

9. GitHub Link

The GitHub repository containing the source code is attached herewith for reference. `https://github.com/drinferni/CompBio`

10. Progress After Presentation

I analyzed the correlation between environmental features of commonly occurring atoms in both ligands and proteins and their relationship with binding affinity. Through this analysis, I discovered that excluding nitrogen from the ligand environment calculations led to improved predictive performance. I computed Pearson correlation and RMSE values on the training data for various models, including SVM, linear regression, and XGBoost. Additionally, I explored the use of PCA to enhance model performance. Finally, I compared the models I developed with state-of-the-art deep learning approaches to evaluate their relative effectiveness.

11. Similar Work in Other Course

No similar work has been submitted in any other course. This project is entirely original and has been developed specifically for the current course, with no overlap in content or prior submissions.

The bibliography below will be generated from the references.bib file

Table 3. PCA-Derived Features on the Test Set (RandomForest9).

Model	Test Pearson	Test RMSE
IMCEP2 (PCA)	0.723	1.416
IMCEP (PCA)	0.736	1.389
IMCPiEB (PCA)	0.736	1.389

References

1. Wang D, Chan MT. Protein-ligand binding affinity prediction based on profiles of intermolecular contacts. Computational and Structural Biotechnology Journal. 2022;20:1088–1096. doi:10.1016/j.csbj.2022.02.004.

2. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. Journal of Chemical Documentation. 1965;5(2):107–113. doi:10.1021/c160017a018.

3. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences. 2021;118(15):e2016239118. doi:10.1073/pnas.2016239118.

4. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics. 2010;26(9):1169–1175. doi:10.1093/bioinformatics/btq112.

5. Zheng L, Fan J, Mu Y. OnionNet: A multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction. ACS Omega. 2019;4(14):15956–15965. doi:10.1021/acsomega.9b01997.

6. Sánchez-Cruz N, Medina-Franco JL, Mestres J, Barril X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. Bioinformatics. 2021;37(10):1376–1382. doi:10.1093/bioinformatics/btaa982.

7. Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. Journal of Chemical Information and Modeling. 2014;54(9):2555–2561. doi:10.1021/ci500319f.

8. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. Molecules. 2015;20(7):13384–13421. doi:10.3390/molecules200713384.