Drug Discovery Through Deep Learning and 3D Protein-Ligand Modeling

Qamil Mirza [1] Aditi Jain [1] Aarav Shah [1] Kelly Hui [1] Milena Novakovic [1] Alan Cheng [2] Matthew Adrian [2] Yunsie Chung [2]

[1]University of California Berkeley

[2]Merck

Introduction

Accurate prediction of half-maximal inhibitory concentration (IC_{50}) is essential for therapeutic development. EGFR and CYP2D6 are clinically important targets in cancer progression and drug metabolism, respectively. While experimental IC_{50} measurements are costly and time-consuming, computational models often lack accuracy due to limited use of protein-ligand context. Chemprop, a graph-based message passing neural network, offers strong performance on molecular property prediction, but does not account for binding interactions. To address this, we integrate Protein-Ligand Interaction Fingerprints (PLIFs) generated using PLIP into Chemprop. This work evaluates whether incorporating PLIF-derived features can improve IC_{50} prediction accuracy for EGFR and CYP2D6 inhibitors.

Dataset Overview

EGFR

The Epidermal Growth Factor Receptor (EGFR) is a well-established target in cancer therapy. EGFR plays a critical role in cell growth, migration, proliferation, and apoptotic resistance in various tumor types. Many anti-cancer drugs aim to inhibit EGFR by binding to its kinase domain, thereby preventing aberrant signaling that drives tumor progression. The data is sourced from BindingDB (curated from literature, PubChem, patents/WIPO, and ChEMBL), containing ~9,500 molecules.

CYP2D6

The Cytochrome P450 2D6 (CYP2D6) is a critical enzyme that is responsible for metabolizing commonly prescribed medications, including antidepressants, analgesics, and anticancer drugs. Primarily expressed in the liver and the brain, CYP2D6 modulates functional groups from drugs via processes such as hydroxylation and demethylation. The data is sourced from BindingDB (curated from literature, PubChem, patents/WIPO, and ChEMBL), containing ~5,000 molecules.

Methodology

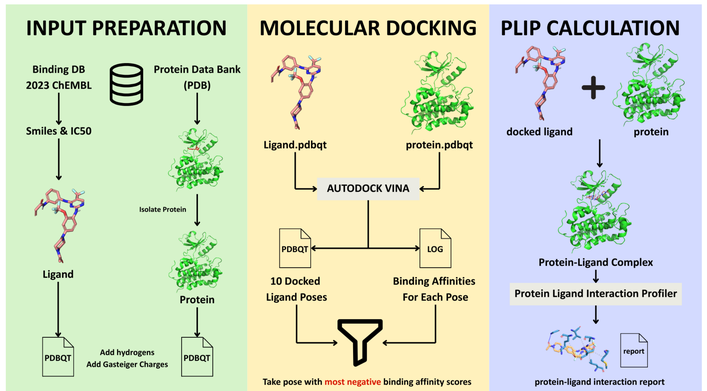


Figure 1. Overview of the molecular docking and interaction profiling workflow. Ligands and proteins are prepared from public databases, converted to PDBQT format, and docked using AutoDock Vina. The top-scoring pose is analyzed using PLIP to extract interaction features for downstream modeling.

Interaction Fingerprint Extraction and Feature Engineering

- Parsed `report.xml` from PLIP to extract:
  - **Atom-level** features: added to ligand atoms.
  - **Graph-level** features: summarize interaction counts/distances.
- Reduced 108D PLIF feature set using:
  - **Unsupervised:** PCA, PLS regression
  - **Supervised:** Random Forest importance, F-regression
- Generated compact PLIFs (3D, 10D) for:
  - Improved model performance
  - Enhanced interpretability
- Applied transformations to emphasize short-range effects:
  - $1/d$
  - $1/d^2$

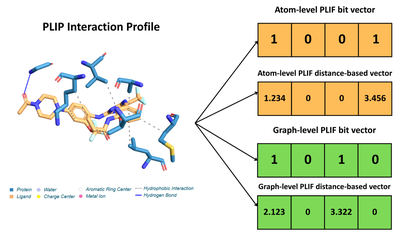
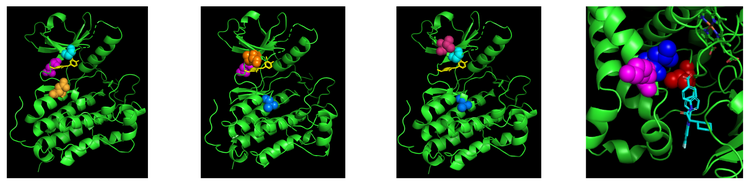


Figure 2. Atom-level and Graph-level Feature Extraction From PLIP Interaction Profile

Key Residues Involved In Binding



EGFR – Random Forest | EGFR – PLS Regression | EGFR – F Regression | CYP – Random Forest

Figure 3. Visualization of protein structures with highlighted important residues identified by labeled method.

Residues selected by our feature methods align with biologically important EGFR regions:

- **Ala840 (blue):** Loop involved in recognizing phosphorylated peptides.
- **Leu718, Val717, Gly719 (orange/red/cyan):** Key residues in the flexible P-loop; modulate ligand affinity by adjusting around the binding site.
- **Asp800 (light orange):** Near kinase hinge; interacts with ligands extending into this region.
- **Gln791 (purple):** Crucial hinge residue forming backbone interactions with ligands.

These overlaps suggest our model learns chemically and biologically meaningful patterns.

Directed Message Passing Neural Network (DMPNN)

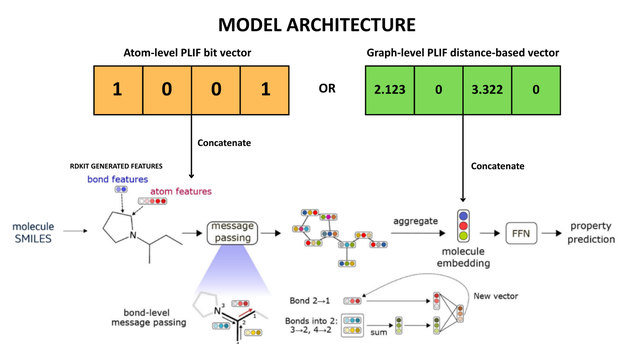
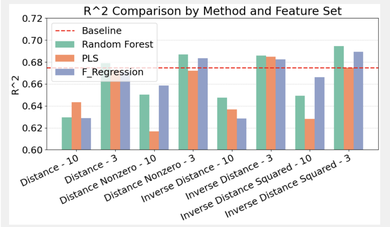


Figure 4. Integration of PLIP-Derived Interaction Features into Chemprop for Property Prediction. PLIP-derived interaction fingerprints (PLIFs), capturing hydrogen bonds, and hydrophobic contacts, are incorporated into Chemprop as graph-level features appended to the final molecular embedding or atom/level features added before message passing. These enriched inputs aim to improve prediction via a feedforward neural network.

Results and Discussion

Model Variant	EGFR R^2	EGFR r	EGFR MAE	EGFR RMSE	CYP2D6 R^2	CYP2D6 r	CYP2D6 MAE	CYP2D6 RMSE
Baseline	0.6747	**0.6785**	**0.5731**	**0.7817**	**0.3113**	**0.3428**	0.5261	**0.7178**
Atom-Level (1/Dist., H-Bond)	0.6152	0.6255	0.6308	0.8488	0.3101	0.3414	**0.5214**	0.7201
Atom-Level (Binary, Hydrophobic)	0.6061	0.6206	0.6324	0.8589	0.2476	0.2710	0.5849	0.7733
Atom-Level (1/Dist, Hydrophobic)	0.6066	0.6203	0.6344	0.8578	0.2540	0.2761	0.5691	0.7540
Graph-Level (All Res.)	0.5023	0.5194	0.7244	0.9664	0.1083	0.1187	0.6310	0.8340
Graph-Level (1/Dist, 3 Res.)	**0.6946**	0.6543	0.6038	0.8160	0.2578	0.2957	0.5499	0.7456

Table 1. Performance comparison across feature strategies for EGFR and CYP2D6



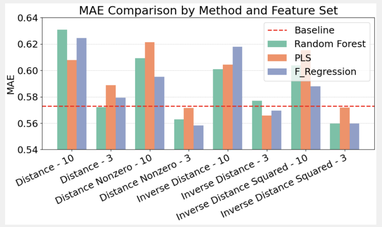


Figure 5. Comparison of dimensionality reduction strategies for the EGFR feature set. Random Forest (green), PLS (orange), and F-regression (purple) were used to select 3 or 10 features from 108-dimensional fingerprints, with variants transformed via inverse and inverse-squared distance. Bar plots show model performance in R^2 (left) and MAE (right) using Chemprop.

Adding atom-level interaction fingerprints degraded model performance across all settings. We attribute this to the **sparsity** of atom-level features, which introduced noise rather than meaningful signals.In contrast, **graph-level features** based on ligand-to-residue distances improved performance slightly when carefully selected.

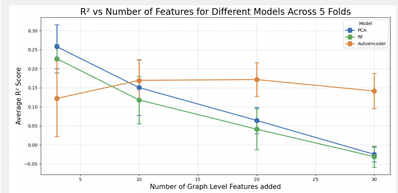


Figure 6. Comparison of graph-level features' impact on R^2 for CYP2D6 prediction. Adding features decreases the R^2 between experimental and predicted IC50 values for PCA and Random Forest, while the Autoencoder (trained on top 5 RF features) prediction remains stable with increasing latent dimensions.

Key Findings:

- Using distances to the top 3 most important residues (via Random Forest or F-regression) led to higher R^2 and lower MAE than using 10 or all 108 residues.
- Inverse ($1/d$) and inverse-squared ($1/d^2$) distance transformations further improved correlation with the target.
- Random Forest and F-regression outperformed PLS for feature selection.
- On the EGFR dataset, the best-performing setup (top 3 residues + $1/d^2$ + RF) achieved an R^2 of 0.6946, surpassing the baseline.

Future Directions

To enhance predictive performance and interpretability, future work will explore directly learning atom-level interaction features from 3D structures, replacing rule-based PLIP extraction. We will summarize residue spatial information using compact descriptors (e.g., mean distances, residue counts) instead of high-dimensional vectors. Additionally, we aim to integrate protein context by incorporating residues directly as graph nodes or edge features, potentially modifying Chemprop architecture.

References

[1] Bryant, P., *et al.*, Nat. Commun., **15**(1), 2024.

[2] He, H., *et al.*, NPJ Digit. Med., **8**(1), 2025.

[3] Liu, Y., *et al.*, J. Mol. Graph. Model., **116**, 2022.

[4] Ross, J., *et al.*, Nat. Mach. Intell., **4**(12), 2022.

[5] Swain, C., Molecular Interactions, `https://cambridgemedchemconsulting.com`, 2025.

[6] Gilmer, J., *et al.*, MPNN Explained, `https://paperswithcode.com/method/mpnn`, 2025.

[7] Patel, H., PLIFs with RDKit and PLIP, `https://blopig.com`, 2016.