# A Machine Learning Framework for Predicting Natural Product-Protein Interactions

**Jiabo Li** LIJIABO@SHU.EDU.CN
*School of Computer Engineering and Science, Shanghai University, Shanghai, China*

**Shijie Gai** 20241513023@SSPU.EDU.CN
*School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, China*

**Wenfeng Shen** WFSHEN@SHU.EDU.CN
*School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, China*

**Zhou Lei**[*] LEIZ@SHU.EDU.CN
*School of Computer Engineering and Science, Shanghai University, Shanghai, China*
[*] *Corresponding author*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

Natural products (NPs) are valuable resources for drug development, but accurately predicting their interactions with protein targets remains challenging due to the limitations of existing methods, which primarily rely on either ligand-based approaches or hybrid feature-based methods that require protein pocket data. To address these limitations, we developed a Y-shaped machine learning framework that integrates NP structural data with protein sequence information. We constructed a comprehensive NP-protein interaction dataset and extracted features from NPs, including Atom Sequence Path (ASP), PubChem, and Extended Connectivity Fingerprints (ECFP), as well as protein features such as Amino Acid Composition (AAC), Conjoint Triad (CTriad), and Dipeptide Composition (DPC). Six machine learning models—Random Forest (RF), AdaBoost, XGBoost, K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), and Logistic Regression (LR)—were trained and evaluated. Experimental results demonstrated that NP-derived PubChem features and protein-derived DPC features were the most effective, with XGBoost achieving the best performance among all models. Our study provides an efficient and generalizable framework for NP-protein interaction prediction, significantly advancing the potential for drug discovery.

**Keywords:** Drug discovery, Compound-protein interaction, Machine learning, Natural product

## 1. Introduction

The development of new drugs is a complex process, spanning from the discovery of lead compounds to the final production of a drug. It requires substantial and continuous investments of time and resources to ensure both safety and efficacy (Lavecchia and Di Giovanni, 2013). Compounds derived from natural products (NPs) play a crucial role in drug development and serve as a valuable resource for drug discovery (Atanasov et al., 2021). Over the past few decades, approximately 50% of new chemical entities (NCEs) have been inspired by NPs or their derivatives, with antitumor drugs accounting for 74% of these (Liang et al., 2022). Most drugs exert their therapeutic effects by interacting with specific molecular targets in the body, such as enzymes, nuclear receptors, G-protein-coupled receptors, or ion channels (Landry and Gies, 2008). Identifying the interactions between NPs and their targets is essential for accelerating drug discovery (Chen et al., 2016). Since

© 2025 J. Li, S. Gai, W. Shen & Z. Lei.

about 95% of drug targets are proteins, this study focuses on investigating the interactions between NPs and human proteins (NPI) (Wang and Kurgan, 2019).

Although traditional experimental methods can accurately determine NPIs, they are often time-consuming, costly, and difficult to scale. With advances in computational technology, researchers have developed various computational approaches to predict compound-protein interactions (CPI) and accelerate drug discovery. Several studies have successfully constructed machine learning and deep learning models for CPI prediction (Wu et al., 2024; Pei et al., 2023). However, these models typically do not distinguish between NPs and synthetic compounds. Liang et al. (2022) pointed out significant differences between NPs and synthetic compounds, such as chiral carbon content, ring system diversity, element composition, functional group variety, and molecular properties. These differences highlight the need for predictive models specifically designed for NPI.

However, existing studies on NPs are either ligand-based or hybrid feature-based (Liang et al., 2022; Yang et al., 2024) . Ligand-based methods assume that similar compounds exhibit similar properties, but they lose effectiveness when only a few ligands are known for a given target. Hybrid feature-based methods, on the other hand, rely on protein pocket data, which limits their scope and generalizability when such data is unavailable. In contrast, the Y-shaped architecture, which uses separate branches for compound and protein encoding, captures compound structural feature and protein sequence information to predict their interactions (Ye et al., 2021; Song et al., 2023). This approach has shown promise in overcoming these limitations of both ligand-based and hybrid feature-based methods.

This study has three main objectives: 1) to build a comprehensive dataset of interactions between NPs and proteins, including compound SMILES representations, protein amino acid sequences, and their binding activities; 2) to develop multiple supervised learning classifiers for predicting NPI; 3) to evaluate various compound and protein features to identify the most effective features or combinations for distinguishing NPI. By constructing and optimizing NPI prediction models, we aim to improve prediction accuracy and expand the model's applicability.

## 2. Methods

In a machine learning task, features and models are the two key components that largely determine the overall success of the project. Feature extraction is especially critical, as it directly influences the model performance. Effective feature selection requires a deep understanding of the data, combined with careful consideration of the specific application scenario and the characteristics of the algorithm employed. In this study, we combined NP and protein features to develop machine learning models for predicting interactions between NPs and human proteins. As shown in Fig. 1, we selected one type of compound feature and one type of protein feature in each case, concatenating the two feature sets and feeding them into the model to learn and predict whether an interaction exists. Six algorithms were applied: Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), XGBoost, and Adaptive Boosting (AdaBoost). A comprehensive evaluation of each model's performance was conducted, facilitating the identification of the most effective algorithm for this predictive task.

### 2.1. Molecular Representations

In this study, for machine learning tasks, we selected three types of compound descriptors based on different generation methods: Extended Connectivity Fingerprints with a diameter of 4 (ECFP4)

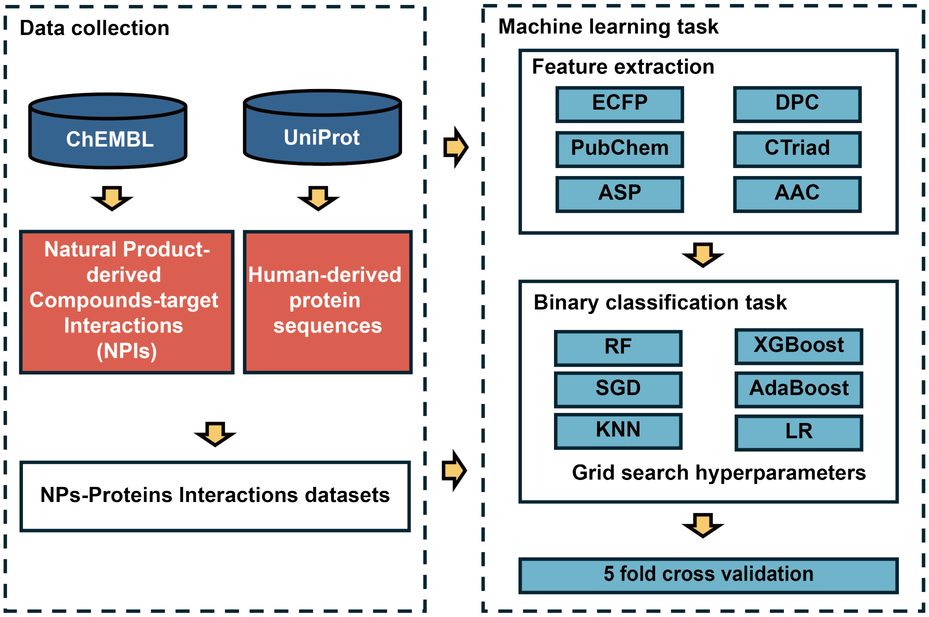
Data collection
ChEMBL
UniProt
Natural Product-derived Compounds-target Interactions (NPIs)
Human-derived protein sequences
NPs-Proteins Interactions datasets
Machine learning task
Feature extraction
ECFP
DPC
PubChem
CTriad
ASP
AAC
Binary classification task
RF
XGBoost
SGD
AdaBoost
KNN
LR
Grid search hyperparameters
5 fold cross validation

Figure 1: Research framework

(Rogers and Hahn, 2010), PubChem (Kim et al., 2021), and All Shortest Paths (ASP) (Hinselmann et al., 2011). The rationale for choosing these descriptors is as follows: ECFP4 is widely recognized as the benchmark fingerprint for representing drug-like compounds; ASP shows great potential in predicting biological activity; and PubChem, a substructure-based descriptor, performs better in handling the chemical space of NPs compared to the commonly used MACCS fingerprints (Boldini et al., 2024). Boldini et al. (2024). integrated RDKit, CDK (Willighagen et al., 2017), and jCompoundMapper (Hinselmann et al., 2011), and our NP features were primarily computed using their tools.

### 2.2. Protein Representations

For proteins, we selected three widely used sequence-based descriptors: Amino Acid Composition (AAC), Conjoint Triad (CTriad), and Dipeptide Composition (DPC) These descriptors were chosen for their popularity and unique ability to capture distinct aspects of protein sequences.

AAC represents the proportion of each standard amino acid in the protein sequence, yielding a 20-dimensional feature vector. The CTriad descriptor classifies the 20 natural amino acids based on their dipoles and side chain volumes. DPC quantifies the frequency of amino acid pairs in each protein sequence.

## 3. Experimental

### 3.1. Datasets

Our data primarily comes from the ChEMBL and UniProt databases. ChEMBL integrates the rich NP information from the COCONUT database and includes many experimental data for compounds and proteins, while the UniProt database provides comprehensive protein information. We downloaded the ChEMBL 34 version of the SQLite data and used SQL queries to filter NPs from the COCONUT database with an NP-likeness score of $\geq 0.6$. Then, we extracted the activity data (pChEMBL values) for these NPs related to human protein targets. Next, we retrieved the amino acid sequences of these human proteins from UniProt. Ultimately, we obtained a total of 19,630 data entries. Each entry includes the following information: the ChEMBL ID of the NPs, SMILES representation, UniProt ID, protein sequence, and pChEMBL value. Based on previous studies, we considered pChEMBL values $\leq 4.5$ as non-interacting and values $\geq 5.5$ as interacting, to construct a binary classification dataset. In the end, we obtained 10,497 data entries, involving 1,190 proteins and 4,061 NPs.

### 3.2. Experimental Details

The experiments were conducted using the scikit-learn framework and executed on a system equipped with an Intel Xeon Platinum 8358P CPU. For each model, grid search was performed to determine the model's hyperparameters. Model performance was evaluated using five-fold cross-validation with standard metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). This approach ensured a rigorous and reliable assessment of our proposed method.

## 4. Results

### 4.1. Evaluation of Different Feature Combinations

Given that no single feature consistently outperforms others across all aspects, we conducted a series of comparative experiments using our dataset to evaluate the impact of different handcrafted features on the performance of six machine learning models. The results are presented in Fig. 2, showing the mean AUC of 54 models obtained through 5-fold cross-validation. It is demonstrated in Fig. 2 that feature combinations significantly influenced performance. Ensemble learning models, such as XGBoost, RF and AdaBoost, exhibited relatively stable results, while other models showed greater variability. Among all feature and model combinations, the XGBoost model using PubChem and DPC features achieved the best performance, with an AUC of 0.8636. In contrast, the SGD model using ASP and AAC features showed the poorest performance, achieving an AUC of only 0.637. Combinations of PubChem with CTriad or PubChem with DPC consistently performed well, whereas ASP and AAC combinations underperformed across all models.

To further explore the contributions of compound and protein features, we separately analyzed their impact on model performance. For example, we assessed XGBoost with PubChem features by averaging its performance across three combinations: PubChem + DPC, PubChem + CTriad, and PubChem + AAC. This average reflects the overall predictive capability of XGBoost using PubChem features. As shown in Fig. 3, PubChem features outperformed the other two compound features, consistently achieving the highest AUC scores across all models. This indicates that PubChem descriptors effectively capture relevant information for the predictive task. Ensemble models like XGBoost and RF showed exceptional performance when utilizing PubChem features, highlighting their robustness in handling high-quality molecular representations. In contrast, ASP features exhibited significantly lower AUC scores compared to PubChem and ECFP features. This performance gap was especially noticeable in simple linear models, such as LR and SGD, where the AUC

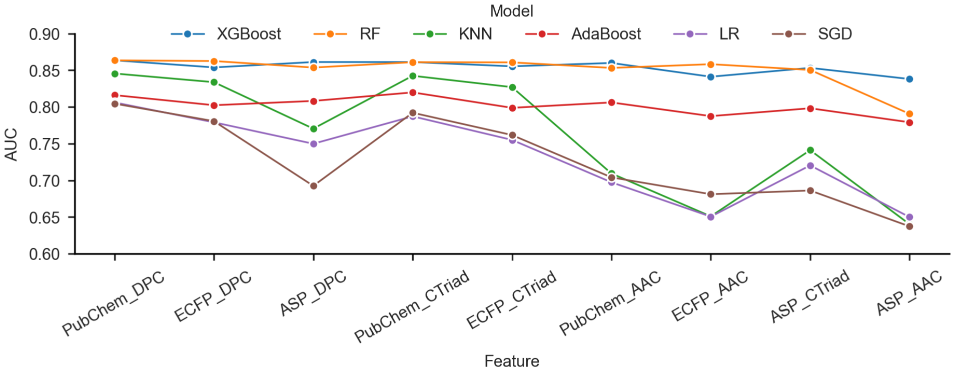

Figure 2: Performance of various models across different feature types

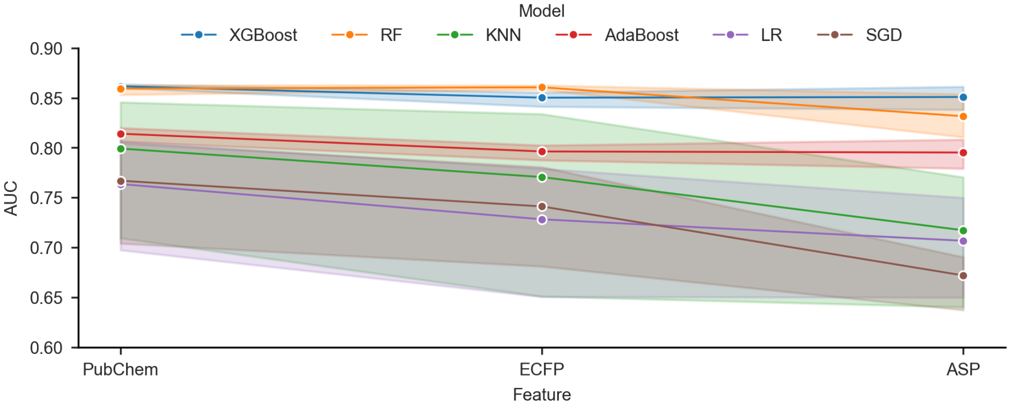

Figure 3: The impact of different NP features on model performance

values were close to random prediction levels. This suggests that ASP features may lack sufficient information for this task or be challenging for the models to learn effectively.

Fig. 4 highlights the superior performance of DPC features across all models. Even simpler models, such as LR and SGD, performed relatively well with DPC features, achieving AUC values around 0.8, demonstrating the broad applicability of DPC across different model types. In contrast, AAC features exhibited the poorest performance across all models. Even the ensemble models, such as XGBoost and RF, showed lower AUC values with AAC features compared to CTriad and DPC features. This suggests that AAC features may lack sufficient information or structure to effectively support predictive tasks.

### 4.2. Model Performance Comparison

The experimental results are presented in Table 1, and the features and hyperparameters used for the models are shown in Table 2. Among all the machine learning models, ensemble classifiers (RF, XGBoost, and AdaBoost) generally outperformed individual classifiers. The best-performing

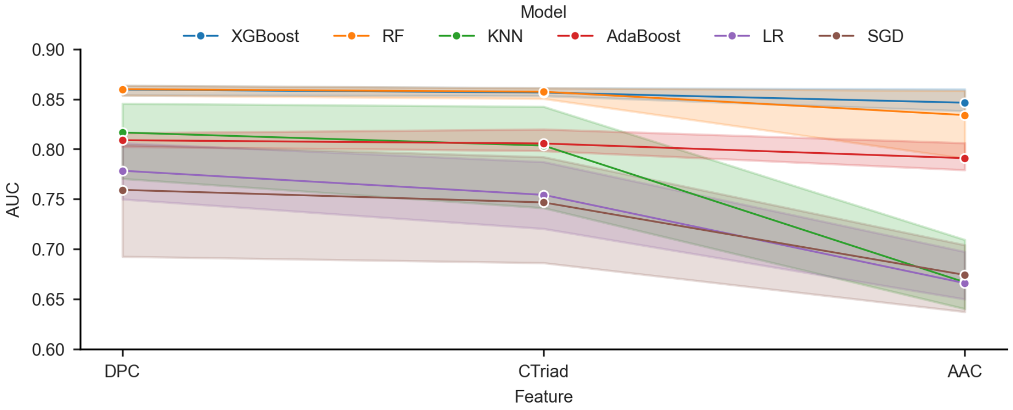

Figure 4: The impact of different protein features on model performance

Table 1: Model performance comparison

| Methods | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| AdaBoost | 0.7615±0.0188 | 0.7991±0.0295 | 0.8421±0.0200 | 0.8199±0.0223 | 0.8198±0.0186 |
| KNN | 0.7859±0.0162 | 0.8337±0.0227 | 0.8347±0.0210 | 0.8341±0.0195 | 0.8456±0.0119 |
| LR | 0.7569±0.0247 | 0.7932±0.0324 | 0.8422±0.0224 | 0.8169±0.0270 | 0.8061±0.0255 |
| RF | 0.8010±0.0153 | 0.8335±0.0277 | 0.8642±0.0137 | 0.8484±0.0188 | 0.8635±0.0069 |
| SGD | 0.7596±0.0253 | 0.7842±0.0374 | 0.8662±0.0171 | 0.8228±0.0255 | 0.8045±0.0252 |
| XGBoost | 0.8006±0.0179 | 0.8316±0.0298 | 0.8661±0.0119 | 0.8484±0.0207 | 0.8636±0.0108 |

model was XGBoost, with an average AUC of 0.8636. Ensemble classifiers like RF, XGBoost, and AdaBoost combine multiple learners, allowing them to capture nonlinear relationships and complex patterns more effectively, while also exhibiting stronger resistance to noise, resulting in superior performance.

Under the optimal feature combination, AdaBoost and KNN demonstrated comparable performance, with KNN slightly outperforming AdaBoost. However, as shown in Fig. 2, KNN's performance is highly sensitive to feature combinations, whereas AdaBoost exhibits greater robustness. KNN relies on distance metrics between nearest neighbors, making it sensitive to data noise and prone to the ”curse of dimensionality,” where distance measurements between samples lose meaning in high-dimensional spaces, explaining its poor performance.

LR and SGD showed the poorest performance on the dataset. LR assumes a linear relationship between features and the target variable, which limits its effectiveness when the data exhibits greater complexity. Similarly, SGD uses a stochastic approach to optimize the objective function but is susceptible to noisy gradients and saddle points, resulting in suboptimal performance.

### 4.3. Model Interpretation

Machine learning models are widely applied in drug screening, particularly when handling large-scale datasets with complex feature interactions. These models demonstrate strong predictive capa-

Table 2: Features and Hyperparameters for Each Model

| Methods | NP Features | Protein Features | Hyperparameters |
|---|---|---|---|
| AdaBoost | PubChem | CTriad | learning_rate=1.0, n_estimators=200 |
| KNN | PubChem | DPC | metric=manhattan, n_neighbors=9, weights=distance |
| LR | PubChem | DPC | C=0.01, solver=lbfgs |
| RF | PubChem | DPC | max_depth=None, min_samples_split=5, n_estimators=200 |
| SGD | PubChem | DPC | alpha=0.01, loss=log_loss, penalty=elasticnet |
| XGBoost | PubChem | DPC | learning_rate=0.1, max_depth=9, n_estimators=100 |

bilities; however, their ”black box” nature often makes it difficult for scientists to understand how specific predictions are made. To address this issue, we performed further interpretability analysis on the well-performing RF and XGBoost models. The interpretability of these models can be examined from two perspectives. First, we assessed the relationship between model predictions and input features to understand the contribution of each feature to the model's decision-making process. Second, we focused on the interpretability of manually curated features of NPs and proteins.

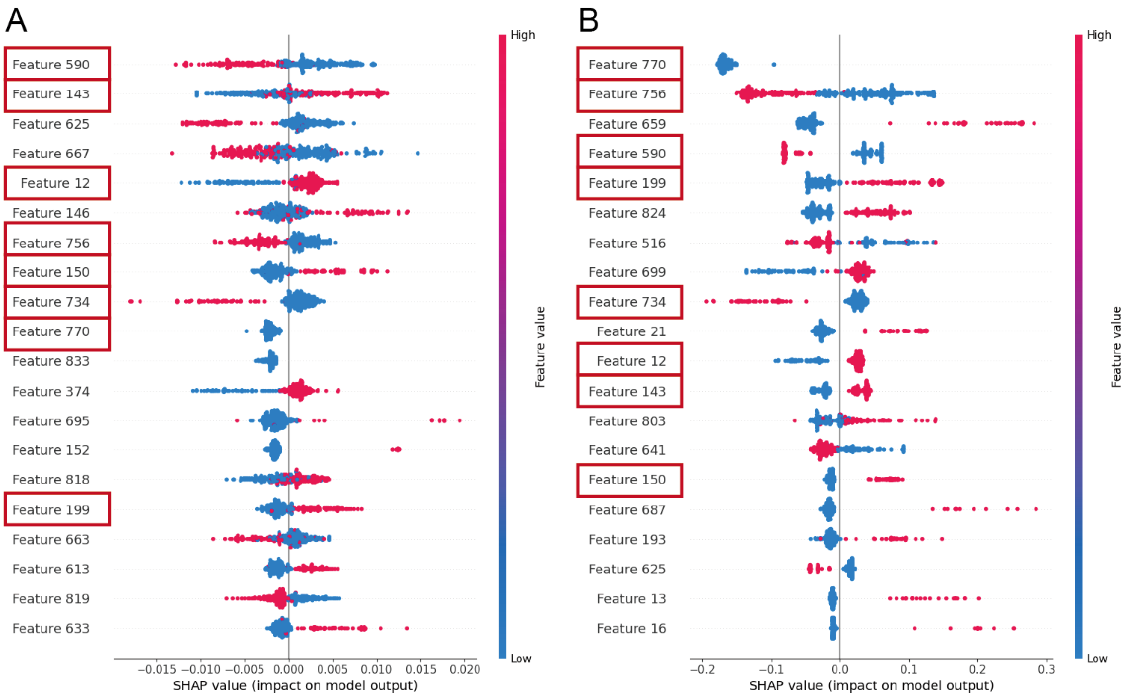

Figure 5: Global explanation of the model for the prediction results of all ligands of protein P08684. (A) SHAP values of the Random Forest model, showing the impact of each feature on the model's predictions. (B) SHAP values of the XGBoost model, illustrating the contribution of each feature to the model's output. Positive SHAP values indicate features that increase the likelihood of interaction, while negative values indicate features that decrease it.

In this study, we used PubChem fingerprints as feature representations for NPs. These fingerprints are structural descriptors based on molecular substructures: if a specific substructure is present in each NP, the corresponding feature is assigned a value of 1; otherwise, it is set to 0. This binary representation is not only simple but also highly interpretable in analyzing the relationship between molecular structures and biological activity. In contrast, protein features were encoded using the DPC method, which represents protein sequences based on dipeptide combinations. While this method provides sequence information, it lacks details about protein 3D structures or binding sites. Therefore, in our model interpretation process, we focused more on the features of NPs to extract biologically meaningful insights into protein-ligand interactions. To better understand the role of NP features, we fixed the protein features during interpretation and analyzed only the contributions of ligand (NP) features to model predictions. For this analysis, we employed the SHAP (SHapley Additive exPlanations) method, which provides a unified and theoretically grounded approach to explain individual feature contributions. By leveraging Shapley values from cooperative game theory, SHAP ensures fair and consistent attribution of feature importance, making it highly effective for interpreting complex machine learning models. For example, in a case study of the P08684 protein, we used the SHAP method to interpret the predicted interactions between P08684 and its 328 associated NPs. Since the protein features were kept constant, the SHAP analysis reflected only the contributions of the ligand features, as shown in Fig. 5. The results indicate that the important features identified by both models (RF and XGBoost) show substantial overlap. Furthermore, the impact of these overlapping features on model predictions (whether positive or negative) is generally consistent. For instance, in the case of protein P08684, when ligand feature 590 has a high value, the model is more likely to classify the ligand as a negative interactor. Given that PubChem fingerprints use binary encoding (0 or 1), this observation suggests that the presence of the substructure corresponding to feature 590 in a NP may inhibit its interaction with P08684. The full description of PubChem fingerprints can be found at: https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt.

Through this analysis method, we can identify the key factors driving the model's decisions, thereby improving its transparency. At the same time, we can identify specific substructures that may either enhance or inhibit the interaction between NPs and proteins. These insights can be used to guide the optimization of lead compounds—for example, by avoiding substructures that negatively impact target binding or prioritizing molecular scaffolds that facilitate binding. Ultimately, this approach helps researchers understand the key molecular features of NPs, offering valuable guidance for drug screening, molecular optimization, and future experimental design.

## 5. Conclusion

This study developed a machine learning-based framework to predict NPI by integrating structural data from NPs with sequence features of proteins. We first constructed a comprehensive dataset of NPI and evaluated various feature combinations, identifying PubChem features for NPs and DPC features for proteins as the most effective for this task. Among the six machine learning models tested, XGBoost achieved the highest accuracy and robustness. Moreover, our approach extends the applicability of NPI prediction to target proteins with limited or no known ligand or binding pocket information. These findings provide a valuable methodological foundation for computer-aided drug discovery and highlight the potential of machine learning in identifying novel drug targets, offering new opportunities for accelerating drug development.

## Acknowledgments

The authors thank the team members for their guidance and appreciate the researchers who provided critical data and tools for this study, and thank the reviewers for their time and valuable suggestions.

## References

Atanas G. Atanasov, Sergey B. Zotchev, Verena M. Dirsch, International Natural Product Sciences Taskforce, and Claudiu T. Supuran. Natural products in drug discovery: advances and opportunities. *Nature Reviews. Drug Discovery*, 20(3):200–216, March 2021. doi: 10.1038/s41573-020-00114-z.

Davide Boldini, Davide Ballabio, Viviana Consonni, Roberto Todeschini, Francesca Grisoni, and Stephan A. Sieber. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *Journal of Cheminformatics*, 16(1):35, March 2024. doi: 10.1186/s13321-024-00830-3.

Xing Chen, Chenggang Clarence Yan, Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin, and Yongdong Zhang. Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics*, 17(4):696–712, 2016. Publisher: Oxford University Press.

Georg Hinselmann, Lars Rosenbaum, Andreas Jahn, Nikolas Fechner, and Andreas Zell. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. *Journal of Cheminformatics*, 3(1):3, December 2011. doi: 10.1186/1758-2946-3-3.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, and Bo Yu. PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395, 2021. Publisher: Oxford University Press.

Yves Landry and Jean-Pierre Gies. Drugs and their molecular targets: an updated overview. *Fundamental & Clinical Pharmacology*, 22(1):1–18, February 2008. doi: 10.1111/j.1472-8206.2007.00548.x.

Antonio Lavecchia and Carmen Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current medicinal chemistry*, 20(23):2839–2860, 2013. Publisher: Bentham Science Publishers.

Lu Liang, Ye Liu, Bo Kang, Ru Wang, Meng-Yu Sun, Qi Wu, Xiang-Fei Meng, and Jian-Ping Lin. Large-scale comparison of machine learning algorithms for target prediction of natural products. *Briefings in Bioinformatics*, 23(5):bbac359, September 2022. doi: 10.1093/bib/bbac359.

Qizhi Pei, Lijun Wu, Jinhua Zhu, Yingce Xia, Shufang Xie, Tao Qin, Haiguang Liu, Tie-Yan Liu, and Rui Yan. Breaking the barriers of data scarcity in drug–target affinity prediction. *Briefings in Bioinformatics*, 24(6):bbad386, September 2023. doi: 10.1093/bib/bbad386.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. doi: 10.1021/ci100050t.

Nan Song, Ruihan Dong, Yuqian Pu, Ercheng Wang, Junhai Xu, and Fei Guo. PMF-CPI: assessing drug selectivity with a pretrained multi-functional model for compound–protein interactions. *Journal of Cheminformatics*, 15(1):97, October 2023. doi: 10.1186/s13321-023-00767-z.

Chen Wang and Lukasz Kurgan. Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Briefings in bioinformatics*, 20(6):2066–2087, 2019. Publisher: Oxford University Press.

Egon L. Willighagen, John W. Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliazkova, Stefan Kuhn, Tomáš Pluskal, Miquel Rojas-Chertó, Ola Spjuth, Gilleain Torrance, Chris T. Evelo, Rajarshi Guha, and Christoph Steinbeck. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9(1):33, December 2017. doi: 10.1186/s13321-017-0220-4.

Hongjie Wu, Junkai Liu, Tengsheng Jiang, Quan Zou, Shujie Qi, Zhiming Cui, Prayag Tiwari, and Yijie Ding. AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Networks*, 169:623–636, January 2024. doi: 10.1016/j.neunet.2023.11.018.

Ruoqi Yang, Lili Zhang, Fanyou Bu, Fuqiang Sun, and Bin Cheng. AI-based prediction of protein–ligand binding affinity and discovery of potential natural product inhibitors against ERK2. *BMC Chemistry*, 18(1):108, June 2024. doi: 10.1186/s13065-024-01219-x.

Qing Ye, Chang-Yu Hsieh, Ziyi Yang, Yu Kang, Jiming Chen, Dongsheng Cao, Shibo He, and Tingjun Hou. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. *Nature Communications*, 12(1):6775, November 2021. doi: 10.1038/s41467-021-27137-3.