Drug-Target Interaction Prediction

Petar Atanasovski 216052

February 2025

Abstract - Predicting drug-target interactions (DTIs) is key to accelerating drug discovery and repurposing. This work presents a hybrid approach that integrates real-time validation using the BindingDB IC50 database with a machine learning model for novel predictions. Leveraging SMILES-based Morgan fingerprints and FASTA-derived features, the system enables a React frontend for seamless PubChem and UniProt querying, while a Python backend handles preprocessing and inference.

I INTRODUCTION

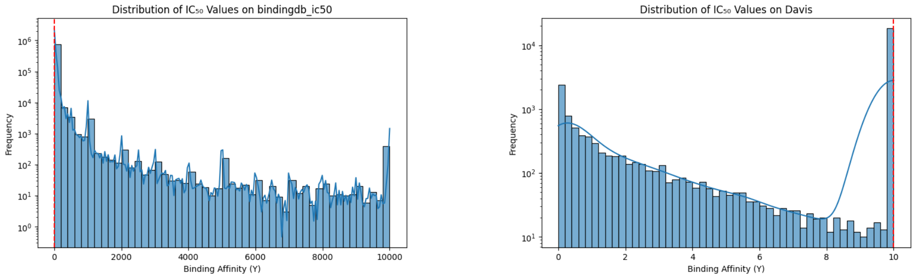
Identifying drug-target interactions (DTIs) is essential for developing new therapies and repurposing existing drugs. While experimental methods like high-throughput screening are highly reliable, they are costly and impractical for large-scale studies. Computational approaches, particularly those using cheminformatics and machine learning, provide a more scalable solution by predicting binding affinities based on molecular and protein features. However, many existing tools struggle to balance real-time validation with the ability to predict novel interactions.

To address this, I propose a hybrid system that combines instant API-driven queries to BindingDB for known interactions with a machine learning model trained on SMILES (chemical structures) and FASTA (protein sequences) to predict unseen pairs. With an intuitive interface, a robust preprocessing pipeline, and interpretable models, this framework aims to speed up early-stage drug discovery while tackling the cold-start problem in sparse biological data.

II RELATED WORK

Recent advancements in computational drug discovery have sparked interest among both computer scientists and biologists, laying the foundation for this project.

DeepDTA [1] introduced deep learning for DTI prediction by leveraging SMILES and protein sequences, demonstrating the potential of neural networks for feature fusion. However, unlike our approach, DeepDTA does not incorporate real-time database validation, highlighting the novelty of our hybrid framework. MoleculeNet [2] established key benchmarks in molecular machine learning, validating Morgan fingerprints as robust drug representations—an approach that aligns with our preprocessing pipeline and chem-



(a) BindingDB IC50 Distribution

(b) Davis Dataset Distribution

Figure 1: Comparison of Binding Affinity Distributions

informatics standards.

BindingDB [3] serves as a foundational resource for experimentally measured IC50 values, which I integrate in real-time via PyTDC to enhance the reliability of known interactions. Additionally, leveraging available technologies and free APIs significantly reduces workload while maximizing efficiency. For example, this project utilizes the PubChem API's [4] autocomplete functionality for streamlined drug input, showcasing the impact of community-driven cheminformatics tools on user accessibility.

Yamanishi et al. [5] pioneered pharmacological network models using protein domains and ligand substructures, inspiring our emphasis on feature-based machine learning over traditional similarity-based methods.

III DATABASE AND DATA PRE-PROCESSING

In this section, I will discuss the databases and data preprocessing steps employed in my project, focusing on the challenges and strategies I implemented to optimize my model's performance.

The first dataset I considered was obtained from the BindingDB IC50 database [Fig. 1.a], which contains approximately one million rows. While this large dataset offers an extensive range of interactions, it posed significant challenges due to its scale and the computational resources required for processing and model training. Moreover, the dataset exhibits a considerable range of IC50 values, including extremely large values, which risk skewing the learning process. A model trained on this data would likely default to predicting these extreme values rather than learning meaningful patterns related to binding affinity, which is a crucial aspect of drug-target interaction prediction.

To address these limitations, I explored another dataset [Fig. 1.b] derived from the Davis dataset, which is relatively smaller with only 25,000 rows. The key difference with this dataset is that the examples provided are primarily related to affinities that are less than or equal to 10. This smaller, more focused dataset limits the scope of the model, effectively training it only on interactions with specific affinity ranges. Although this restriction could potentially reduce model accuracy in predicting in-

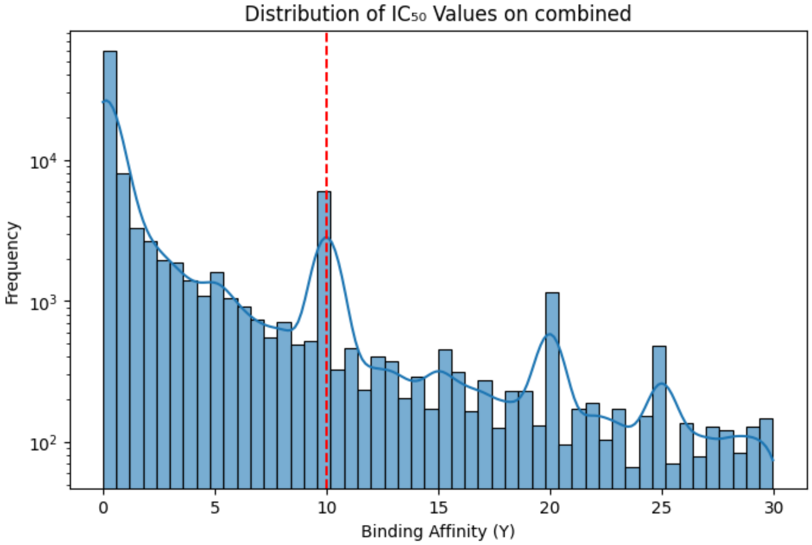


Figure 2: Distribution of Custom Dataset's Binding

teractions beyond the narrow affinity range, it allows the model to become specialized in a particular segment of data. However, this dataset still presented limitations, as it lacked diversity and did not represent the broader variability found in real-world drug-target interactions.

To combine the best aspects of both datasets, I created a new hybrid database [Fig. 2]. This new dataset integrates the extensive coverage of the BindingDB IC50 data with the more controlled affinity range from the Davis dataset. By focusing on a carefully curated selection of rows, I aimed to ensure that the model would be trained on a diverse set of interactions that better reflect the complexities of real-world drug-target binding affinities.

The target measure is IC50, or the half-maximal inhibitory concentration, is a critical measure in drug-target interaction studies, representing the concentration of a drug needed to inhibit a specific biological or biochemical function by 50%. In the context of this project, IC50 is used as the target variable (denoted as ”Y”), and it serves as a proxy for binding affinity. The lower the IC50 value, the stronger the binding affinity between the drug and its target, meaning the drug is more effective at lower concentrations. Higher IC50 values suggest weaker binding affinity, requiring higher drug concentrations to achieve the same inhibitory effect. This metric is central to assessing the potency of potential drug candidates.

The preprocessing phase was also critical to ensuring the effectiveness of the hybrid database. I employed a range of techniques, including handling missing data and transforming the raw features into usable inputs for machine learning. For the molecular data, I used SMILES (Simplified

Molecular Input Line Entry System) to represent chemical structures, converting these into Morgan fingerprints, which are highly effective for capturing molecular properties. Similarly, for the protein sequences, I used FASTA format and extracted amino acid features, enabling the model to learn from both the chemical structure and the protein sequence. These preprocessing steps were crucial for converting raw data into a form that could be efficiently processed by machine learning models.

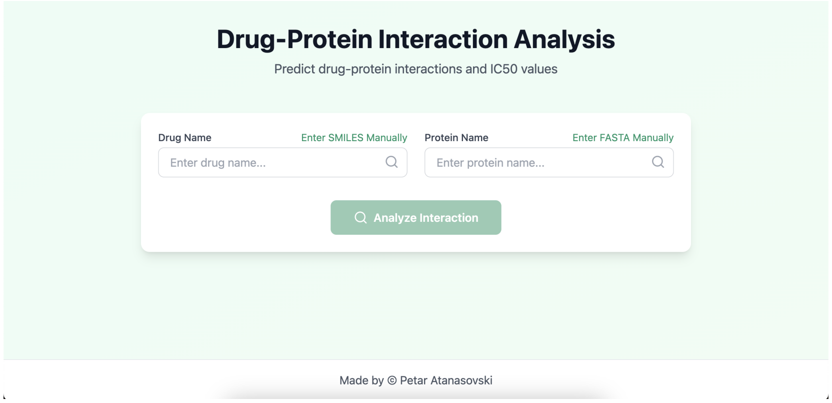
The dataset was further refined by ensuring that there were no missing values or inconsistencies, as shown in the output where missing data was confirmed to be absent. This level of cleanliness is essential to avoid introducing noise that could undermine the learning process. Once the data was preprocessed and cleaned, I split it into training and testing sets, ensuring that the model could be effectively evaluated on unseen data.

IV CREATING A MODEL

The core of this project lies in developing a predictive model that bridges chemical and biological data to estimate drug-target binding affinity (IC50). After evaluating several algorithms, I utilized a Random Forest Regressor model to predict drug-target interaction affinities. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of individual trees. This approach is effective for handling large datasets with complex, non-linear relationships, as it can learn from diverse feature sets and capture intricate patterns without overfitting easily. By setting n_estimators=100, I created a forest of 100 trees, allowing the model to benefit from averaging, which reduces variance and improves generalization. Additionally, setting random_state=42 ensures reproducibility of the model's results.

While Random Forest is a strong contender for this task, it's important to consider other models as well. For instance, Gradient Boosting Machines (GBM), such as XGBoost or LightGBM, have proven highly effective for regression tasks in various domains, offering faster training times and better performance with tuned hyperparameters. Another possibility is Support Vector Regression (SVR), which works well in high-dimensional spaces, though it can be computationally expensive for large datasets. Neural Networks, particularly deep learning models, could also be explored, but they require significantly more data and computational power.

Random Forest was chosen for its ease of implementation, robustness, and capacity to perform well with the available data, but future iterations may include exploration of more complex models for further improvement. While larger datasets might enable more complex models, this approach prioritizes transparency and accessibility. The pipeline achieves reliable performance on modest hardware, making it viable for researchers without high-performance computing resources. Future iterations could explore hybrid models or domain adaptation techniques, but the current methodology strikes a pragmatic balance between biological insight and computational pragmatism.



Drug-Protein Interaction Analysis

Predict drug-protein interactions and IC50 values

Made by © Petar Atanasovski

Figure 3: User Friendly App

V EXPERIMENTS, RESULTS, DISCUSSION

The predictive model built using a Random Forest Regressor offers valuable insights into the potential interactions between drugs and proteins. While the results obtained from this model are satisfactory for the scope of the project, the true potential of drug-target interaction prediction lies in the ability to scale up and leverage more advanced techniques, such as deep learning models, on larger datasets. By utilizing a neural network with more computational power, we could tap into the full depth of available biological data, leading to more accurate and generalized predictions.

In addition to the model itself, this project incorporates both a frontend and backend application [Fig. 3] that facilitate direct user interaction and testing. The frontend, built using React, provides a user-friendly interface where users can enter drug names and protein names either manually or by pasting SMILES and FASTA sequences. This ease of interaction ensures that users can test and evaluate the predictions made by the model in real time. The backend, implemented in Python, handles data preprocessing, database search and model prediction. The separation between frontend and backend allows for a clean and modular design, ensuring that both parts of the system can be iteratively improved without affecting the other.

One of the strengths of this approach is the ability to run real-time analyses, offering immediate feedback to users about the potential interactions between a drug and a protein. This direct testing mechanism is crucial for researchers and pharmaceutical companies looking to accelerate drug discovery by enabling quick testing of various drug-protein combinations.

The results from the model were promising within the scope of this project. The predictions were able to capture the general trends in drug-target interactions and IC50 values, though some limitations were present due to the size of the dataset and the complexity of the interactions involved. For example, while the model performed well on known interactions with smaller IC50 values, it struggled to predict interactions with extreme IC50 values that were underrepresented in the data.

With larger datasets and more powerful models, such as neural networks, the system has great potential to make more accurate and robust predictions. Neural networks can learn from vast amounts of data and generalize better across different drug-target pairs, which would be beneficial for improving predictions in a real-world context.

This project has demonstrated the feasibility and utility of integrating machine learning models with real-time validation systems to predict drug-target interactions. The potential for further improvements, particularly by leveraging larger and more complex datasets, is enormous. As computational resources grow, this approach could be expanded to address a broader range of interactions and more precise predictions, further accelerating the drug discovery process.

VI DOCUMENTATION, SOURCE CODE, AND PRESENTATION

The source code for this project is available at the following GitHub Repository. The README file serves as the documentation, providing an overview of the project, installation instructions, and guidance on how to use the model and the application. This repository includes all the necessary code for both the frontend and backend, allowing others to replicate, contribute, or further develop the system. Additionally, a presentation outlining the key components of the project and its findings is available for further reference.

References

[1] Hakime Öztürk, Arzucan Özgür and Elif Ozkirimli, "DeepDTA: deep drug-target binding affinity prediction", 2018

[2] Zhenqin Wu, Bharath Ramsundar et. al., "MoleculeNet: a benchmark for molecular machine learning", 2018

[3] Michael K Gilson, Tiqing Liu et. al., "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology", 2026

[4] Sunghwan Kim, Jie Chen, Tiejun Cheng et. al., "PubChem 2023 update", 2023

[5] Yamanishi et al., "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces", 2008