Dự đoán hiệu quả và có thể giải thích tương tác protein-ligand sử dụng mô hình tăng cường gradient và AI có thể giải thích

TÓM TẮT

Dự đoán khả năng liên kết của các phân tử nhỏ với các mục tiêu protein là một bước quan trọng trong quá trình khám phá thuốc hiện đại, mở ra tiềm năng đẩy nhanh việc xác định các liệu pháp điều trị hiệu quả đồng thời giảm chi phí thí nghiệm. Trong nghiên cứu này, chúng tôi sử dụng bộ dữ liệu BELKA, một thư viện hóa học mã hóa bằng DNA (DEL) quy mô lớn, để huấn luyện các mô hình học máy nhằm dự đoán khả năng liên kết. Bằng cách áp dụng XGBoost, một thuật toán gradient boosting dựa trên cây, cùng với các bước tiền xử lý và thiết kế đặc trưng chuyên sâu, chúng tôi đã phát triển các mô hình dự đoán cho ba mục tiêu protein: BRD4, HSA, và sEH. Các mô hình này thể hiện năng lực dự đoán mạnh mẽ, đồng thời cho phép giải thích kết quả thông qua phân tích SHAP nhằm xác định các đặc trưng phân tử quan trọng quyết định khả năng liên kết. Đánh giá trên bộ dữ liệu kiểm tra BELKA cho thấy những thách thức trong việc khái quát hóa, cung cấp những hiểu biết quý giá về sự phức tạp của mô hình dự đoán trong khám phá thuốc. Nghiên cứu này nhấn mạnh tiềm năng của học máy trong việc thúc đẩy quá trình khám phá thuốc bằng máy tính, cho phép khám phá không gian hóa học hiệu quả hơn để tìm kiếm các liệu pháp điều trị tiềm năng.

Từ khoá: *Khám phá thuốc, học máy, AI có thể giải thích.*

Efficient interpretable prediction of protein-ligand interactions using gradient boosting models and explainable AI

ABSTRACT

The prediction of small molecule binding affinity to protein targets is a critical step in modern drug discovery, offering the potential to accelerate the identification of effective therapeutics while reducing experimental costs. In this study, we employ the BELKA dataset, a large-scale DNA-encoded chemical library (DEL), to train machine learning models for binding affinity prediction. Using XGBoost, a tree-based gradient boosting algorithm, and extensive preprocessing and feature engineering, we develop predictive models for three protein targets: BRD4, HSA, and sEH. The models demonstrate strong predictive capabilities, with interpretability achieved through SHAP analysis to identify molecular features driving binding predictions. Evaluation of the BELKA test dataset reveals challenges in generalization, providing valuable insights into the complexities of predictive modelling in drug discovery. This work highlights the promise of machine learning in advancing computational drug discovery by enabling efficient exploration of the chemical space for potential therapeutics.

Keywords: *Drug discovery, machine learning, explainable artificial intelligence.*

1. INTRODUCTION

The development of machine learning (ML) models to predict the binding affinity of small molecules to specific protein targets holds transformative potential for drug discovery. Predicting these interactions is central to identifying new, effective drug candidates, as small molecule drugs interact with cellular protein machinery to influence disease-associated biological processes.

Traditionally, screening and testing small molecules for binding affinity to protein targets involve labour-intensive and costly physical experiments, which severely limits the speed and scope of drug discovery efforts.[1,2] The search space for small molecule drugs is estimated to encompass approximately 10^{60} chemical compounds, which is impractical to physically screen.[3]

With the pharmaceutical landscape evolving, the integration of ML-based predictive models offers a promising alternative to these conventional approaches, enabling efficient exploration of the vast chemical space for potential therapeutics. Traditional high-throughput screening (HTS) technologies can assess libraries of small molecules against protein targets, but they are often restricted to collections of tens of thousands to a few million compounds.[4] In response to this limitation, DNA-encoded chemical libraries (DELs) have emerged as a more scalable solution.[5] DELs use unique DNA barcodes to tag each molecule, allowing the pooling of millions of compounds and simplifying the identification of binders through DNA sequencing. This method has substantially expanded the feasible scale of chemical libraries and presents an attractive foundation for computational models aimed at binding affinity prediction.

Advances in ML architectures and feature representation techniques, such as Simplified Molecular Input Line Entry System (SMILES) and graph-based molecular representations, have made it possible to capture complex chemical properties and interactions computationally.[6] SMILES, as a string-based molecular representation, encodes atom connectivity and stereochemistry, facilitating ML models' application in molecular property prediction, drug discovery, and materials design.

Hence, in this work, we explore the application of a tree-based gradient boosting approach, specifically XGBoost, for predicting binding affinity.[7] In addition to model development, a tree-based Explainable AI (XAI) method is integrated to interpret model behaviour, enhancing transparency and interpretability in the prediction of molecular binding. The findings from this study aim to contribute to the broader field of computational drug discovery, leveraging ML to identify promising drug candidates with high precision and potentially reduce the costs associated with traditional drug development methods. By enabling more efficient exploration of chemical space, this work aspires to pave the way toward discovering new lifesaving

therapeutics for complex diseases. Conclusively, in this study, we make the following contributions:

- **Dataset Utilization:** We leverage the BELKA dataset[8], a large-scale DNA-encoded chemical library, providing a comprehensive resource for binding affinity modelling.
- **Predictive Modelling:** We employ the XGBoost model optimized with advanced preprocessing and feature reduction techniques to predict binding affinities for three biologically significant protein targets: BRD4, HSA, and sEH.
- **Interpretability:** Through XAI analysis, we enhance the interpretability of the models, offering molecular-level insights into the features influencing binding predictions.
- **Benchmarking:** We evaluate our methodology on the BELKA dataset, highlighting the challenges of generalization for unseen cases.

2. RELATED WORK

2.1. Drug Discovery and Protein-Target Interactions

The pharmaceutical field relies heavily on understanding and predicting protein-target interactions, as these molecular interactions are critical in developing effective drugs. Small molecule drugs are typically designed to modulate specific protein targets linked to disease mechanisms. Protein-ligand binding is fundamental to this process, as the ability of a drug candidate to bind to a specific protein target determines its efficacy and safety.

Traditional drug discovery methodologies, such as high-throughput screening (HTS), involve synthesizing large libraries of small molecules and testing their affinity with the protein targets. However, HTS is costly, time-intensive, and limited in scope due to physical constraints, allowing only a fraction of potential drug-like compounds to be examined. Innovations, such as DNA-encoded chemical libraries (DELs)[5], have addressed some of these limitations by enabling more extensive exploration of chemical space. In DELs, small molecules are tagged with unique DNA barcodes, allowing millions of compounds to be screened in a pooled format. As such, DELs offer a scalable and efficient alternative to traditional HTS. Advances in molecular biology and DNA sequencing have further accelerated DEL technology, facilitating its adoption in both academia and industry.

2.2. SMILES and Molecular Representations

SMILES is one of the most widely adopted formats for encoding chemical structures in computational chemistry.[6] SMILES strings represent molecular structures in a linear form, capturing atoms, bonds, and stereochemistry in a machine-readable format. This notation has become essential for ML applications in drug discovery due to its simplicity and the ease with which it can be integrated into computational pipelines. SMILES can also be converted to other representations, such as 3D structures and molecular graphs, allowing flexibility in model input formats.

Alternative molecular representations, such as molecular fingerprints and molecular graphs, offer distinct advantages. Molecular fingerprints encode the presence or absence of substructures, providing a high-dimensional, fixed-length vector representation suited for various ML tasks.[9–12] Meanwhile, molecular graphs represent the connectivity of atoms in the molecule, capturing spatial information that can be valuable for models like graph neural networks (GNNs).[13–16] Recent studies suggest that combining multiple representations, such as SMILES with molecular graphs, can enhance predictive accuracy by leveraging diverse information formats.

2.3. Machine Learning in Molecular Binding Prediction

ML has become essential to molecular binding prediction, with recent models achieving high performance by leveraging large datasets and sophisticated algorithms. ML models, especially deep learning (DL) frameworks, can capture complex relationships in chemical and biological data, allowing them to predict molecular properties with increasing accuracy.

Traditional ML methods, such as quantitative structure-activity relationship (QSAR) models, relied on engineered molecular descriptors to predict binding affinity. Still, recent ML approaches enable the use of raw chemical representations such as SMILES and molecular graphs, reducing the need for extensive feature engineering.[17–19] Convolutional neural networks (CNNs)[20], graph neural networks (GNNs)[21], and recurrent neural networks (RNNs)[22] have been widely used to encode molecular structures.

In addition to DL approaches, gradient-boosting algorithms like XGBoost have gained recognition for their efficacy in molecular property prediction. XGBoost suits tasks involving structured, high-dimensional data, such

as molecular fingerprints. By leveraging an ensemble of decision trees, XGBoost iteratively refines predictions, minimizing error while maintaining interpretability. Unlike deep learning models, XGBoost offers a computationally efficient alternative that is well-suited for datasets with tabular or fingerprint-based representations. Recent studies have shown that integrating molecular representations, such as Extended-Connectivity Fingerprints (ECFPs)[23] with XGBoost, yields highly accurate binding affinity predictions while retaining transparency. These models are particularly valuable in scenarios where interpretability is crucial, such as drug discovery pipelines. Additionally, XGBoost's robustness to overfitting, especially when combined with appropriate feature selection and regularization, makes it a strong candidate for handling imbalanced datasets often encountered in molecular binding tasks.

2.4. Explainability in ML for Drug Discovery

As ML models become increasingly complex, understanding the decision-making process within these models is critical for their adoption in sensitive fields like drug discovery. XAI methods aim to make the behaviour of complex ML models more interpretable by providing insights into how input features influence predictions. In drug discovery, XAI can offer insights into which molecular features contribute most significantly to binding affinity, helping chemists understand and validate model predictions.[24,25]

Tree-based models, such as XGBoost, offer interpretability advantages due to their structured decision paths. Techniques like SHAP (SHapley Additive exPlanations)[26] values and LIME (Local Interpretable Model-agnostic Explanations)[27] are often applied to these models, enabling the decomposition of predictions into contributions from individual features. For example, SHAP values, derived from cooperative game theory, were especially used to quantify each feature's influence on the prediction. These explanation methods not only facilitate model interpretation but also foster trust in ML predictions, an essential factor for the integration of AI into pharmaceutical workflows.

3. DATASET

The BELKA dataset used in this study comprises training and test samples that detail the interactions between various small molecules and three protein targets: bromodomain-containing protein 4 (BRD4), soluble epoxide hydrolase (EPHX2/sEH), and human serum albumin (ALB/HSA).[8]

3.1. Dataset Targets

The BELKA dataset encompasses three distinct protein targets: BRD4, EPHX2/sEH, and ALB/HSA. Each target represents a unique class of biomolecular interactions, selected to provide a diverse benchmarking ground for modelling small molecule-protein interactions. These targets were carefully chosen for their biological significance and existing therapeutic relevance. Their acquisition and preparation followed rigorous protocols to ensure data fidelity and reproducibility.

3.1.1. BRD4

Bromodomain-containing protein 4 is a pivotal member of the BET protein family, involved in recognizing acetylated lysines on histone tails.[28] BRD4 has emerged as a prominent therapeutic target in oncology, with inhibitors designed to disrupt its role in transcriptional regulation, particularly in cancer proliferation pathways. Recombinant BRD4 was acquired through baculovirus expression in insect cells to preserve post-translational modifications critical for its bromodomain function. Protein purity and structural integrity were validated through size-exclusion chromatography and binding assays with known BRD4 inhibitors. These quality-control measures ensured that the BRD4 used in DEL screenings retained its native binding characteristics, enabling high-confidence small molecule-protein interaction studies.

3.1.2. BPHX2/sEH

Soluble epoxide hydrolase is an enzyme involved in metabolizing lipid epoxides, converting them into diols through hydrolysis.[29] This enzymatic activity has been implicated in numerous physiological and pathological processes, including inflammation, pain, and cardiovascular diseases. Recombinant human EPHX2 was expressed in Escherichia coli and purified via affinity chromatography. Its activity was verified using substrate-based fluorescence assays to confirm functional integrity before integration into DEL screening assays. By selecting sEH as a target, the BELKA dataset facilitates the evaluation of ligand binding in the context of enzymatic specificity and inhibition.

3.1.3. ALB/HSA

Human serum albumin, the most abundant plasma protein, plays a key role in drug pharmacokinetics by binding and transporting a wide range of endogenous and exogenous compounds.[30] For this dataset, HSA was isolated from human plasma and subjected to additional purification to remove

potential impurities. Its binding activity was assessed through equilibrium dialysis and competitive ligand-binding assays to confirm its ability to interact with small molecules[31]. Using HSA in the DEL screening enables exploring protein-small molecule interactions that influence drug bioavailability and distribution.

3.2. Dataset Acquisition

The raw readout acquisition process is visualized in Figure 1. The primary library, AMA014, is a triazine-based shree-cycle library designed to resemble DEL-A. An additional orthogonal DEL, termed *kinase0 (kin0)*, was designed to mimic kinase inhibitor chemistry.

The screening methodology involved combining the DEL with the target protein, isolating DEL/target complexes, eluting the bound DEL through heat application, and repeating the selection with the fresh target protein. This iterative process, conducted over three rounds for AMA014, aimed to enrich high-affinity binders. Each selection series for AMA014 was performed in triplicate to assess reproducibility. In contrast, the smaller kinase0 library underwent a single selection round, performed in duplicate with a single negative control. Post-selection, the eluted DELs were subjected to sequencing to quantify binding events. The dataset includes both binary binding labels and raw sequencing counts, facilitating diverse analyses, including evaluating hit-calling methods and experimental design parameters. The raw dataset encompasses approximately 4.25 billion physical measurements, with compressed data totalling around 600 GB.

All protein targets underwent rigorous selection and preparation to maintain high experimental reproducibility. For each target, protein binding assays were conducted to confirm the enrichment of small molecule binders across multiple rounds of DEL screening. The screening workflow included initial binding assays with the target protein, iterative selection and amplification of enriched libraries, and sequencing to quantify binding events. These protocols were designed to capture high-affinity interactions and a broad spectrum of molecular binders, ensuring a comprehensive dataset for benchmarking predictive models.

3.3. Dataset Description

Each row in the dataset encapsulates the chemical composition and binding characteristics of a small molecule with a specific protein target, providing a structured basis for learning binding patterns across different protein targets and molecular configurations.

The training dataset $\mathcal{D}_{train}$ (as shown in Table 1) includes molecular structures represented by SMILES strings, with each sample specifying four chemical building blocks, a complete molecular structure, the protein target, and a binary binding label (1 for binding, 0 for no binding) as the target variable. The test dataset $\mathcal{D}_{test}$ follows a similar structure without the binding label, providing the molecular structure and target protein only. Each column in the dataset is described as follows:

- **id**: A unique identifier for each record. Every unique combination of small molecule features is represented by three consecutive rows, each corresponding to a specific protein target: BRD4, HSA, or sEH. This structure allows for direct comparisons of binding affinity predictions across the three protein targets for the same molecular structure.
- **buildingblock1_smiles**: A SMILES string representing the first building block of the molecule. This component forms part of the molecular structure and contributes specific chemical properties to the final molecule.
- **buildingblock2_smiles:** A SMILES string for the second building block. Together with the first and third building blocks, it helps define the molecule's structure and potential binding characteristics.
- **buildingblock3_smiles:** A SMILES string representing the third building block of the molecule, completing the combination of foundational elements used to form the final molecule.
- **molecule_smiles:** A SMILES string for the entire molecule, constructed from the building blocks and representing the complete molecular structure, including atoms, bonds, and stereochemistry. This column is a primary input for machine learning models to predict binding affinity based on the molecule's overall chemical structure.
- **protein_name:** The name of the protein target for each molecule, which can be one of three values—BRD4, HSA, or sEH. Each protein target has a specific biological significance and is used to determine the binding affinity of the molecule to a particular protein. For each unique molecule, the dataset includes rows for all three proteins to allow cross-target comparisons.

- **binding_label** ($\mathcal{D}_{train}$ only): A binary label indicating whether the molecule binds to the specified protein target. A value of '1' signifies that the molecule binds to the target, while '0' indicates no binding. This label is used as the target variable y.

For instance, the 2D representation of a molecule in the BELKA dataset is demonstrated in Figure 1.

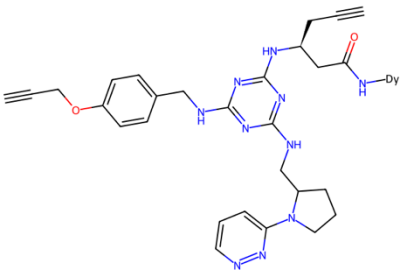

Figure 1. The 2D representation of a BELKA molecule (C#CCOc1ccc(CNc2nc(NCC3CCCN3c3cccnn3)nc(N[C@@H](CC#C)CC(=O)N[Dy])n2)cc1).

Table 1. Training dataset excerpts for 3 targets: BRD4, HSA, and sEH.

id	buildingblock1_smiles	buildingblock2_smiles	buildingblock3_smiles	molecule_smiles	protein_name	binds
0	C#CC[C@@H](CC(=O)O)NC(=O)OCC1c2ccccc2-c2ccccc21	C#CCOc1ccc(CN)cc1.Cl	Br.Br.NCC1CCCN1c1cccnn1	C#CCOc1ccc(CNc2nc(NCC3CCCN3c3cccnn3)nc(N[C@@H](CC#C)CC(=O)N[Dy])n2)cc1	BRD4	0
1	C#CC[C@@H](CC(=O)O)NC(=O)OCC1c2ccccc2-c2ccccc21	C#CCOc1ccc(CN)cc1.Cl	Br.Br.NCC1CCCN1c1cccnn1	C#CCOc1ccc(CNc2nc(NCC3CCCN3c3cccnn3)nc(N[C@@H](CC#C)CC(=O)N[Dy])n2)cc1	HSA	0
2	C#CC[C@@H](CC(=O)O)NC(=O)OCC1c2ccccc2-c2ccccc21	C#CCOc1ccc(CN)cc1.Cl	Br.Br.NCC1CCCN1c1cccnn1	C#CCOc1ccc(CNc2nc(NCC3CCCN3c3cccnn3)nc(N[C@@H](CC#C)CC(=O)N[Dy])n2)cc1	sEH	0

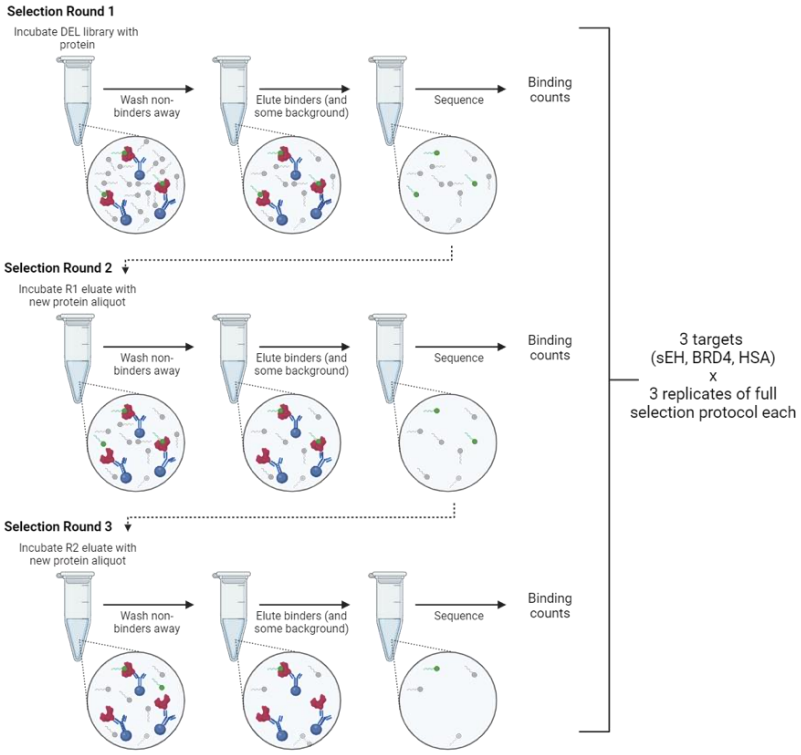

Figure 2. The methodology of recording the raw readouts for 3 targets (sEH, BRD4, HSA)[8].

4. METHODOLOGY

4.1. Data Preprocessing

In the data preprocessing phase, as demonstrated in **Figure 3**, the dataset was processed in increments of 10^4 rows to manage memory efficiently, given its large size. Each molecule in the dataset, represented by SMILES strings, was processed to create ECFPs, a commonly used molecular representation in cheminformatics. The SMILES strings for each molecule were converted into RDKit molecular objects, and ECFPs were generated with a radius of 2 and a fingerprint size of 2048 bits. The ECFPs were transformed into sparse matrix format to optimize memory usage, and additional bit information for each fingerprint was captured to enhance interpretability.

To reduce the dimensionality and improve computational efficiency, each of the building blocks (as shown in Section 3), namely *buildingblock1_smiles*, *buildingblock2_smiles* and *buildingblock3_smiles* were mapped to unique integer identifiers, with dictionaries created for each set of SMILES strings.

For instance, building blocks in *buildingblock1_smiles* were mapped to integer values in *blocks_dict_1*, while a shared dictionary, *blocks_dict_23*, was created for *buildingblock2_smiles* and *buildingblock3_smiles* due to the overlap between these blocks. These mappings were then saved, allowing for efficient lookup and reuse. To address the class imbalance, particularly given the scarcity of positive binding cases, the dataset was downsampled for non-binding entries, retaining all rows where the binding was detected and sampling a subset of non-binding cases. This balanced dataset provided an optimal size and improved training stability.

Processed data, including the sparse ECFP matrices and integer-encoded building blocks, were saved into a training balanced set $\mathcal{E}_{train}$ and a test set $\mathcal{E}_{test}$, in compressed formats for efficient storage and retrieval. This preprocessing pipeline allowed for structured and memory-efficient representation of the dataset, supporting effective model development for binding prediction across the three protein targets.

4.2. Model Implementation

In this section, we described our implementation of a multi-step model training and evaluation process to predict the binding affinity of small molecules to three protein targets: BRD4, HSA, and sEH. This approach involved model selection, feature reduction, model training and evaluation.

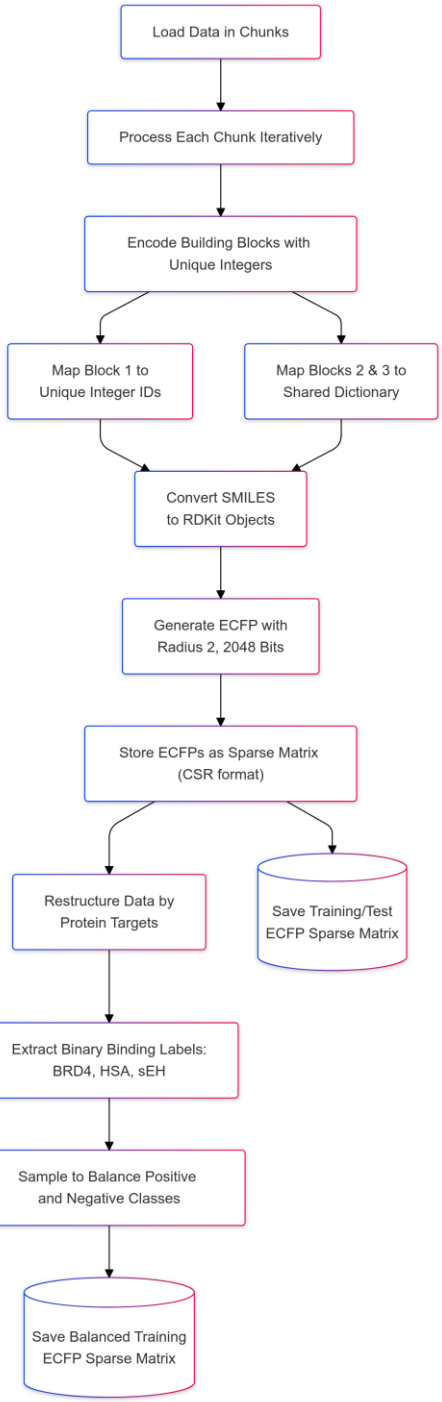

Figure 3. The pipeline of preprocessing dataset.

4.2.1. Training Setup and Data Partitioning

To control randomization across the training process, we initialized a fixed seed as 42, allowing for reproducibility in sampling and shuffling steps. The dataset $\mathcal{E}_{train}$ was then split into a training set $\mathcal{A}_{train}$ (90%) and a validation set $\mathcal{A}_{val}$ (10%) based on a shuffled index of samples. This partitioning enabled model tuning on the training set while using the $\mathcal{A}_{val}$ to assess model generalization and prevent overfitting.

4.2.2. Feature Reduction by Variance Threshold

The initial input data contained high-dimensional molecular fingerprints generated from ECFPs. To reduce dimensionality and enhance model performance, we applied a variance threshold to the ECFP feature matrix. Features with variance θ below 0.005 were removed, as low-variance features contribute minimally to distinguishing between classes. This filtering reduced computational complexity and mitigated overfitting by retaining only the most informative features.

Figure 4. The pipeline of model training, validation and inference.

4.2.3. Model Training and Inference

In this step, we employed XGBoost, a gradient boosting algorithm, to train separate binary classification models for each protein target: BRD4, HSA, and sEH.

For each model, we configured the objective function as binary logistic regression with a learning rate r of 0.2, and the evaluation metric as the average precision score (AP). Early stopping was applied with a patience p of 100 rounds to prevent overfitting, and the model was allowed up to 4000 iterations for convergence. To handle imbalanced data, we computed a scale positive weight for each target, defined as $\omega = \frac{N_{neg}}{N_{pos}}$, where N_{neg} and N_{pos} represent the counts of non-binding and binding samples, respectively.

4.3. Model Explanation with XAI

To understand the contribution of specific molecular features to each model's predictions, we applied an interpretability method, namely SHAP.

We utilized the SHAP TreeExplainer for XGBoost, which computes Shapley values efficiently in tree-based models. SHAP summary plots and bar charts were generated to visualize the global importance of features in predicting binding affinity for each target.

5. RESULTS

5.1. Model Performance

To evaluate the effectiveness of our predictive models on the binding affinity classification task, we conducted a comprehensive performance assessment across the three protein targets: BRD4, HSA, and sEH. The metrics used for evaluation included accuracy, mean average precision (MAP), recall, and the area under the precision-recall curve (AUCPR), which are presented in Table 1. Results were reported separately for the training set $\mathcal{A}_{train}$ and the validation set $\mathcal{A}_{val}$ to provide insights into training stability and generalization.

5.1.1. Evaluation Metrics

Accuracy measures the overall correctness of predictions and is defined as the ratio of correctly classified samples (both positive and negative) to the total number of samples. Mathematically, accuracy is expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. While accuracy provides a general assessment of the model's classification capability, it can be less informative in imbalanced datasets, as it may overemphasize the correct classification of the majority class.

Recall (Sensitivity) measures the proportion of actual positive cases correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall focuses on the model's ability to capture true binders, which is critical in applications where missing positive cases (e.g., potential drug candidates) could have significant consequences. A high recall ensures that the model effectively identifies most true binding interactions.

Mean Average Precision (MAP) evaluates the ranking quality of predictions, particularly the precision of positive cases across various thresholds. It is calculated as the mean of the Average Precision (AP) scores over all classes, where AP combines precision and recall into a single metric that emphasizes the ranking order of positive predictions. MAP is computed as:

$$\mathbf{MAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}\left(y_{\mathrm{true}}^{i}, y_{\mathrm{pred}}^{i}\right)$$

where $\boldsymbol{R_k}$ is the recall at rank $\boldsymbol{k}$ and $\boldsymbol{P_k}$ is the precision at rank $\boldsymbol{k}$. MAP is particularly valuable for imbalanced datasets, as it prioritizes the correct ranking of true positives, making it sensitive to the quality of predictions for the minority class.

Area Under the Precision-Recall Curve (AUCPR) quantifies the trade-off between precision and recall across all decision thresholds. Unlike the Receiver Operating Characteristic (ROC) curve, the Precision-Recall curve is more informative in imbalanced datasets, as it emphasizes the model's ability to correctly classify the positive class. AUCPR is calculated as the area under the curve formed by plotting precision against recall at varying thresholds. A higher AUCPR indicates a better balance between precision and recall, reflecting the model's ability to maintain high sensitivity (recall) without compromising specificity (precision).

Table 2. The model performance on the training and validation set with accuracy, mean average precision (MAP), recall and area under the precision-recall curve (AUCPR).

	BRD4	**HSA**	**sEH**	**BRD4**	**HSA**	**sEH**
	Accuracy			**MAP**		
$\mathcal{A}_{train}$	0.9637	0.9164	0.9798	0.5708	0.3341	0.7913
$\mathcal{A}_{val}$	0.9583	0.9082	0.9767	0.5364	0.3006	0.7754
	Recall			**AUCPR**		
$\mathcal{A}_{train}$	0.9910	0.9543	0.9979	0.9098	0.6751	0.9773
$\mathcal{A}_{val}$	0.9275	0.8467	0.9778	0.8663	0.6076	0.9629

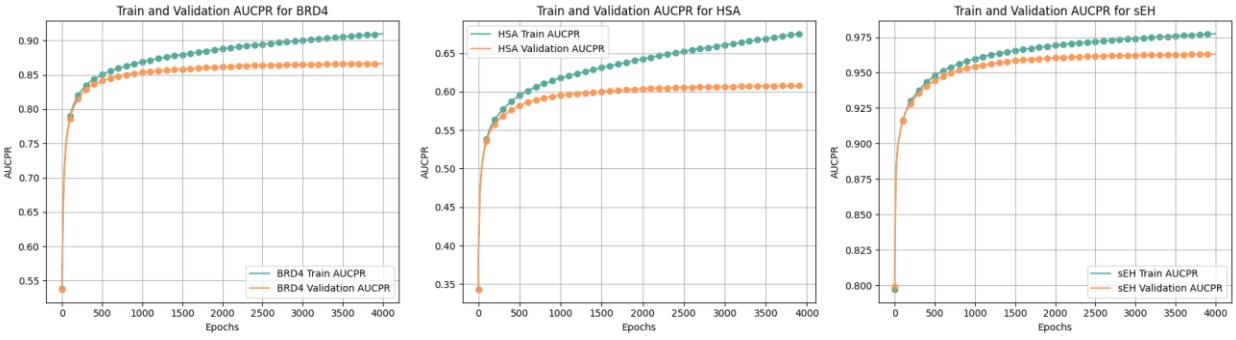

Figure 5. The area under the precision-recall curve (AUCPR) visualization of the model on the training (in green) and validation (in orange) set.

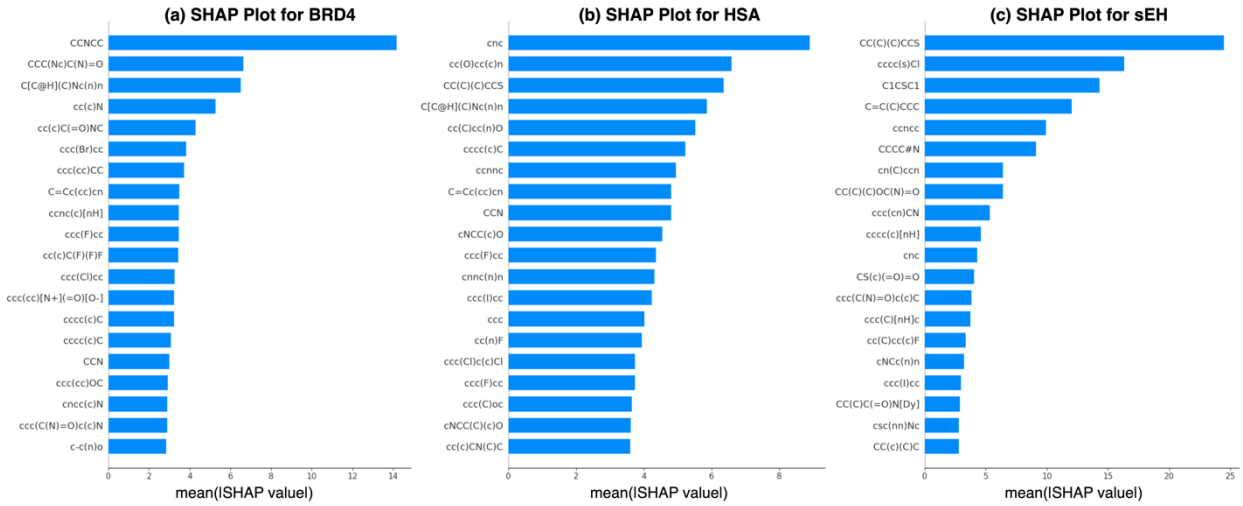

Figure 6. The global explanation (top influencing features) for the model's decisions in predicting the binding affinity for 3 targets: BRD4, HSA and sEH.

These four metrics were chosen to provide a comprehensive evaluation of the models, capturing their overall classification accuracy, ranking quality, sensitivity to true positives, and the precision-recall trade-off. By analyzing these metrics, we can gain deeper insights into the strengths and limitations of the models for each protein target, enabling targeted improvements in future iterations.

5.1.2. Performance Evaluation

For BRD4, the model achieved a high accuracy of 0.9637 on $\mathcal{A}_{train}$ and 0.9583 on $\mathcal{A}_{val}$ indicating minimal overfitting and strong predictive performance. However, the MAP values, which assess the ranking quality of positive predictions, were relatively modest at 0.5708 for $\mathcal{A}_{train}$ and 0.5364 for $\mathcal{A}_{val}$. This reflects the inherent difficulty of ranking positive binders for BRD4. Despite this, the recall values were consistently high, reaching 0.9910 for $\mathcal{A}_{train}$ and 0.9275 for $\mathcal{A}_{val}$, demonstrating the model's capability to identify a significant proportion of true binders. The AUCPR scores, 0.9098 for $\mathcal{A}_{train}$ and 0.8663 for $\mathcal{A}_{val}$, further confirming the model's effectiveness in differentiating binders from non-binders.

For HSA, the model exhibited slightly lower accuracy than BRD4, with values of 0.9164 for $\mathcal{A}_{train}$ and 0.9082 for $\mathcal{A}_{val}$. The MAP scores for HSA, 0.3341 for $\mathcal{A}_{train}$ and 0.3006 for $\mathcal{A}_{val}$, were the lowest among the three targets, indicating challenges in ranking true binders effectively. Nevertheless, the recall metrics for HSA were robust, achieving 0.9543 on $\mathcal{A}_{train}$ and 0.8467 on $\mathcal{A}_{val}$. The AUCPR values, 0.6751 for $\mathcal{A}_{train}$ and 0.6076 for $\mathcal{A}_{val}$, suggest the model's reasonable ability to identify binding patterns, though there is room for improvement in precision-recall balance.

The model's performance on sEH was the strongest overall. The accuracy reached 0.9798 for $\mathcal{A}_{train}$ and 0.9767 for $\mathcal{A}_{val}$, showcasing exceptional classification accuracy. Similarly, the MAP scores, 0.7913 for $\mathcal{A}_{train}$ and 0.7754 for $\mathcal{A}_{val}$, were significantly higher than those for BRD4 and HSA, indicating superior ranking performance. Recall values were near perfect at 0.9979 for $\mathcal{A}_{train}$ and 0.9778 for $\mathcal{A}_{val}$, further emphasizing the model's sensitivity in detecting true binding events. The AUCPR scores, 0.9773 for $\mathcal{A}_{train}$ and 0.9629 for $\mathcal{A}_{val}$, reinforce the robustness of the model for sEH, highlighting its capability to effectively separate binders from non-binders with high confidence.

In addition to evaluating model performance on the $\mathcal{A}_{train}$ and $\mathcal{A}_{val}$ sets, the final models were assessed on the BELKA test dataset $\mathcal{E}_{test}$, as part of a Kaggle competition. Since the test labels for each target were not made available, the evaluation relied solely on the predictions' final test scores. On the public test set, the model achieved an accuracy of 0.2042, while the private test set yielded a slightly lower accuracy of 0.1843.

Finally, these results demonstrate that the boosting models, when combined with efficient data preprocessing and dimensionality reduction, can achieve reliable predictions across diverse protein targets. The differences in MAP and AUCPR scores among the targets underscore the varying complexities of binding prediction, with sEH being the most tractable and HSA presenting the greatest challenges. The results on the test

datasets highlight the challenges posed by the BELKA dataset, particularly the difficulty in achieving generalizable predictions across unseen data. The gap between validation and test performance underscores the potential for further enhancements in model robustness and generalization.

5.2. Model Intepretability

To provide insights into the decision-making process of the predictive models, we employed SHAP to quantify the contribution of individual molecular features to the model's output. SHAP explanations are particularly valuable in understanding which molecular substructures, represented as SMILES fragments, had the most significant impact on the binding affinity prediction for each protein target: BRD4, HSA, and sEH. The SHAP summary plots for the three targets are presented in Figure 6, with the x-axis representing the mean absolute SHAP value, indicative of the average magnitude of a feature's impact on the model's predictions.

BRD4 (Figure 6a): the most influential molecular feature was the fragment "CCNCC," which exhibited the highest mean SHAP value, highlighting its strong association with binding predictions. Other significant contributors included fragments with nitrogen and aromatic substructures such as "CCC(N)C(N)=O" and "C(c)(H)C(Nc=In)," suggesting that these groups may play a key role in interacting with BRD4's bromodomains. Notably, the diversity of impactful features underscores the model's ability to capture complex molecular patterns that influence binding specificity.

HSA (Figure 6b): SHAP analysis revealed "CC(C)(C)CCS" as the most impactful feature. This fragment aligns with HSA's known affinity for hydrophobic and bulky molecular groups, which are critical for its role as a drug carrier protein. Additional significant features included "ccccc1CCl" and "C1CSC1," suggesting a preference for aromatic and cyclic substructures. These insights provide a molecular-level understanding of the interactions influencing the binding of small molecules to HSA.

sEH (Figure 6c): the SHAP summary plot demonstrated that the fragment "CC(C)(C)CCS" had the largest average impact on model predictions, followed by "ccccc1CCl" and "C=C(C)C(CC)." These features are consistent with known hydrophobic binding pockets in sEH, highlighting the model's ability to identify molecular characteristics critical for binding affinity. Notably, the sEH model exhibited a larger range of SHAP values than the other targets, reflecting a higher sensitivity to specific molecular fragments.

The SHAP analysis across all three targets highlights the models' reliance on chemically meaningful features, providing interpretability and transparency in their predictions. These findings not only enhance confidence in the models but also offer valuable insights for the rational design of small molecules with desired binding properties. Future efforts could involve leveraging these SHAP-derived insights for feature engineering or guiding experimental validation to further refine predictive performance.

6. DISCUSSION

This study demonstrates the potential and challenges of ML in molecular binding prediction. The XGBoost models achieved high performance on training and validation datasets, particularly for sEH, which benefited from its consistent molecular binding patterns. However, BRD4 and HSA presented unique challenges due to more diverse binding chemistries, resulting in slightly lower scores. SHAP analysis revealed chemically meaningful features, providing valuable insights into the molecular determinants of binding and guiding potential drug design efforts. The evaluation of the BELKA test dataset highlights a notable performance drop, with public and private test scores of 0.2042 and 0.1843, respectively. This gap underscores the inherent difficulty of generalizing predictive models to unseen data in large, diverse chemical spaces. It also highlights the importance of robust feature selection, additional data augmentation, and more generalized learning methods to bridge the gap between validation and test performance.

While tree-based models like XGBoost are interpretable and effective for structured data, the reliance on binary binding labels rather than continuous affinity scores limits their ability to capture nuanced interactions. Future work could integrate graph-based molecular representations or hybrid approaches combining DL with traditional ML to improve prediction accuracy and generalizability.[14,16,32,33] Additionally, leveraging semi-supervised learning or transfer learning could further enhance model robustness in unseen data scenarios.[34–36]

7. CONCLUSION

This study highlights the potential of machine learning to revolutionize drug discovery by predicting small molecule binding affinities with

high efficiency. Using the BELKA dataset, we demonstrated the capability of XGBoost models to achieve strong predictive performance while providing interpretability through SHAP analysis. However, challenges in generalization, particularly on unseen test datasets, reveal areas for methodological improvement. By combining robust predictive capabilities with interpretable outputs, this work advances computational approaches for drug discovery, enabling more efficient exploration of chemical space and paving the way for identifying novel therapeutics.

REFERENCES

1. P. Arora, M. Behera, S. A. Saraf, R. Shukla. *Leveraging Artificial Intelligence for Synergies in Drug Discovery: From Computers to Clinics*, *Current Pharmaceutical Design*, **2024**, (*30*), 2187–2205.
2. K. Sharma, P. Manchikanti. Artificial Intelligence in Drug Development and Healthcare—Nature and Scope. *Artificial Intelligence in Drug Development: Patenting and Regulatory Aspects*, Springer, 2024, 1–33.
3. J.-L. Reymond, R. Van Deursen, L. C. Blum, L. Ruddigkeit. *Chemical space as a source for new drugs*, *MedChemComm*, **2010**, (*1*), 30–38.
4. J. P. Landry, Y. Fei, X. Zhu. *High throughput, label-free screening small molecule compound libraries for protein-ligands using combination of small molecule microarrays and a special ellipsometry-based optical scanner*, *International drug discovery*, **2011**, 8.
5. A. Gironda-Martínez, E. J. Donckele, F. Samain, D. Neri. *DNA-encoded chemical libraries: a comprehensive review with succesful stories and future challenges*, , *ACS Pharmacology & Translational Science*, **2021**, (*4*), 1265–1279.
6. D. Weininger. *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*, , *Journal of chemical information and computer sciences*, **1988**, (*28*), 31–36.
7. T. Chen, C. Guestrin. Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, **2016**, 785–794.
8. A. Blevins, I. K. Quigley, B. J. Halverson, N. Wilkinson, R. S. Levin, A. Pulapaka, W. Reade, A. Howard. NeurIPS 2024 - Predict New Medicines with BELKA, **2024**.
9. S. F. Baygi, D. K. Barupal. *IDSL_MINT: a deep learning framework to predict molecular fingerprints from mass spectra*, *Journal of Cheminformatics*, **2024**, (*16*), 8.
10. D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni, S. A. Sieber. *Effectiveness of molecular fingerprints for exploring the chemical space of natural products*, *Journal of Cheminformatics*, **2024**, (*16*), 35.
11. D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams. *Convolutional networks on graphs for learning molecular fingerprints*, *Advances in neural information processing systems*, **2015**, (*28*).
12. J. Wang, L. Zhang, J. Sun, X. Yang, W. Wu, W. Chen, Q. Zhao. *Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints*, *Methods*, **2024**, (*221*), 18–26.
13. S. Goldman, J. Li, C. W. Coley. *Generating molecular fragmentation graphs with autoregressive neural networks*, *Analytical Chemistry*, **2024**, (*96*), 3419–3428.
14. S. Kim, J. Woo, W. Y. Kim. *Diffusion-based generative AI for exploring transition states from 2D molecular graphs*, *Nature Communications*, **2024**, (*15*), 341.
15. K.-D. Luong, A. K. Singh. *Fragment-based pretraining and finetuning on molecular graphs*, , *Advances in Neural Information Processing Systems*, **2024**, (*36*).
16. A. Morehead, J. Cheng. *Geometry-complete perceptron networks for 3d molecular graphs*, , *Bioinformatics*, **2024**, (*40*), btae087.
17. A. Z. Dudek, T. Arodz, J. Gálvez. *Computational methods in developing quantitative structure-activity relationships (QSAR): a review*, *Combinatorial chemistry & high throughput screening*, **2006**, (*9*), 213–228.
18. T. A. Soares, A. Nunes-Alves, A. Mazzolari, F. Ruggiu, G.-W. Wei, K. Merz. The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) studies propelled by the surge of machine learning methods, **2022**, (*62*), 5317–5320.

19. M. Honma. *An assessment of mutagenicity of chemical substances by (quantitative) structure–activity relationship*, *Genes and Environment*, **2020**, (*42*), 23.

20. Y. LeCun, Y. Bengio, others. *Convolutional networks for images, speech, and time series*, *The handbook of brain theory and neural networks*, **1995**, (*3361*), 1995.

21. F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini. *The graph neural network model*, *IEEE transactions on neural networks*, **2008**, (*20*), 61–80.

22. D. E. Rumelhart, G. E. Hinton, R. J. Williams. *Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986*, *Biometrika*, **1986**, (*71*), 6.

23. D. Rogers, M. Hahn. *Extended-connectivity fingerprints*, *Journal of chemical information and modeling*, **2010**, (*50*), 742–754.

24. R. Alizadehsani, S. S. Oyelere, S. Hussain, S. K. Jagatheesaperumal, R. R. Calixto, M. Rahouti, M. Roshanzamir, V. H. C. De Albuquerque. *Explainable artificial intelligence for drug discovery and development-a comprehensive survey*, *IEEE Access*, **2024**.

25. K. K. Kırboğa, S. Abbasi, E. U. Küçüksille. *Explainability and white box in drug discovery*, *Chemical Biology & Drug Design*, **2023**, (*102*), 217–233.

26. S. M. Lundberg, S.-I. Lee. A unified approach to interpreting model predictions, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, **2017**, 4768–4777.

27. M. T. Ribeiro, S. Singh, C. Guestrin. ‘ Why should i trust you?’ Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, **2016**, 1135–1144.

28. B. Padmanabhan, S. Mathur, R. Manjula, S. Tripathi. *Bromodomain and extra-terminal (BET) family proteins: New therapeutic targets in major diseases*, *Journal of biosciences*, **2016**, (*41*), 295–311.

29. J. W. Newman, C. Morisseau, B. D. Hammock. *Epoxide hydrolases: their roles and interactions with lipid metabolism*, *Progress in lipid research*, **2005**, (*44*), 1–51.

30. A. Varshney, P. Sen, E. Ahmad, M. Rehan, N. Subbarao, R. H. Khan. *Ligand binding strategies of human serum albumin: how can the cargo be utilized?*, *Chirality: The Pharmacological, Biological, and Chemical Consequences of Molecular Asymmetry*, **2010**, (*22*), 77–87.

31. Y. Sun, Z. Ji, X. Liang, G. Li, S. Yang, S. Wei, Y. Zhao, X. Hu, J. Fan. *Studies on the binding of rhaponticin with human serum albumin by molecular spectroscopy, modeling and equilibrium dialysis*, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **2012**, (*87*), 171–178.

32. F. Ekström Kelvinius, D. Georgiev, A. Toshev, J. Gasteiger. *Accelerating molecular graph neural networks via knowledge distillation*, *Advances in Neural Information Processing Systems*, **2024**, (*36*).

33. V. Fung, J. Zhang, E. Juarez, B. G. Sumpter. *Benchmarking graph neural networks for materials chemistry*, *npj Computational Materials*, **2021**, (*7*), 84.

34. Y. Cho, J. Shim, K. Seok. *Binding Affinity Prediction Using Self-supervised Learning and QP-Ensemble*, *QBS*, **2023**, (*42*), 57–64.

35. Y. Gu, X. Zhang, A. Xu, W. Chen, K. Liu, L. Wu, S. Mo, Y. Hu, M. Liu, Q. Luo. *Protein–ligand binding affinity prediction with edge awareness and supervised attention*, *Iscience*, **2023**, (*26*).

36. B. Tanoori, M. Zolghadri Jahromi, E. G. Mansoori. *Binding affinity prediction for binary drug–target interactions using semi-supervised transfer learning*, *Journal of Computer-Aided Molecular Design*, **2021**, (*35*), 883–900.