World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

High-throughput prediction of small molecule binding affinities to ABL1, HSP90, and CDK2 using gradient boosting and machine learning on the BELKA dataset

Vedant Shrinivas Sagare *

Dublin High School, Dublin, Alameda County, California.

World Journal of Advanced Research and Reviews, 2024, 24(01), 2426–2434

Publication history: Received on 27 August 2024; revised on 23 October 2024; accepted on 26 October 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.1.3068

Abstract

Most of the time, drug development is burdened by a large search space of possible drug-like molecules and resource-consuming conventional screening methodologies. This work leverages machine learning to predict the binding affinity of small molecules to certain protein targets, one of the major steps in modern drug development. The paper hereby aims at making the process of drug discovery more efficient and accurate by leveraging information from the Big Encoded Library for Chemical Assessment, BELKA dataset, which involves 133 million small molecules screened in interaction against three protein targets, namely Tyrosine-protein kinase ABL1, Heat shock protein 90, and Cyclin-dependent kinase 2. A model using LightGBM was thus developed for affinity prediction, using molecular descriptors derived from the SMILES representation of the molecules. It then splits the data into training and test data, and feature extraction is done through RDKit, calculating the molecular weight, hydrogen bond donors, and acceptors for each molecule. The model achieved an average precision score of 0.84 with strong predictive power. This gave an average precision of 0.88 on the target Tyrosine-protein kinase ABL1, followed by a rather moderate score for targets HSP90 and CDK2, with averages of 0.83 and 0.81, respectively. Feature importance analysis showed that molecular weight joined with hydrogen bonding capacity was among the most valued features in the model's predictions. In this respect, LightGBM can be considered a powerful tool in accelerating drug discovery due to its high accuracy and efficiency of prediction of binding interactions, whereby further potential improvements are related to the inclusion of more complex molecular features and 3D descriptors.

Keywords: Small molecule-protein interactions; Machine learning; LightGBM; Molecular descriptors; SMILES; Binding affinity;

1. Introduction

Drug discovery tends to be rather complicated and, often, very time-consuming, which requires heavy investment to be successful in selecting potential therapeutic compounds. Conventional approaches to drug discovery are based on high-throughput screening of ultra-large chemical libraries-a rather inefficient and very costly process. Often, these methods also have limitations regarding both the number of compounds that have to be analyzed and about the way the manual analysis is required to identify promising candidates.

The introduction of DNA-encoded chemical libraries had expedited drug discovery programs manifold, due to the fact that instead of phage or one-by-one, millions of compounds can now be screened all at the same time. DELs identify molecules bound to target proteins by tagging each molecule with a distinct DNA barcode that enables rapid and high-throughput analysis. On the other hand, this has introduced challenges in the analysis and interpretation of the unprecedented volumes of data produced by DELs.

* Corresponding author: Vedant Sagare

Accordingly, machine learning models-most especially for large datasets-skips these challenges by their capabilities in fast and accurate prediction of small molecule-protein interaction. Having said that, the enormous capability of machine learning allows researchers to analyze such complex data with higher efficiency, allowing one to precisely pinpoint a potential drug candidate.

The following BELKA dataset provided by Leash Biosciences comprises 133 million small molecules tested against three protein targets [9]. Therefore, the following provides a unique opportunity for the development of machine learning models capable of predicting binding affinities that could accelerate the process of drug discovery.

1.1. Tyrosine-protein kinase ABL1

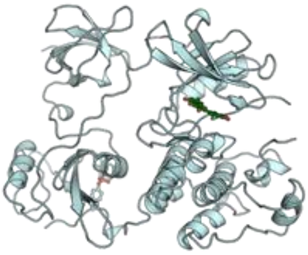


Figure 1 Structural representation of ABL1, a key enzyme involved in cell processes and linked to cancers such as chronic myeloid leukemia

Tyrosine-protein kinase ABL1 is a crucial enzyme involved in various cellular processes, including cell differentiation, division, adhesion, and response to stress. ABL1 is a non-receptor tyrosine kinase that shuttles between the nucleus and cytoplasm, influencing several signaling pathways [2]. Dysregulation of ABL1 activity has been implicated in several cancers, most notably chronic myeloid leukemia (CML). In CML, a chromosomal translocation creates the BCR-ABL fusion protein, which exhibits constitutive kinase activity and drives uncontrolled cell proliferation. Targeting ABL1, particularly the BCR-ABL fusion protein, has been a successful strategy in cancer therapy, with inhibitors like imatinib revolutionizing treatment. Despite these advances, resistance to these drugs can develop, underscoring the need for new inhibitors that can target various mutations of ABL1.

1.2. Heat shock protein 90 (HSP90)

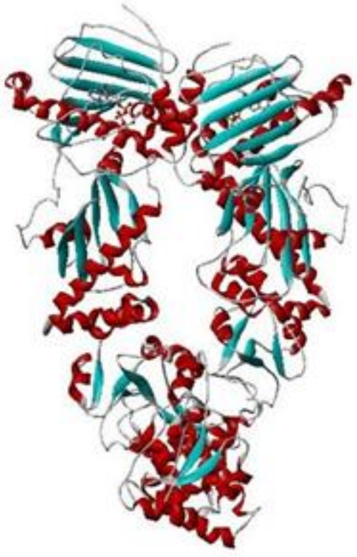


Figure 2 Structural depiction of HSP90, a molecular chaperone responsible for the stability and function of various client proteins, many of which are involved in oncogenic processes

Heat shock protein 90 (HSP90) is a molecular chaperone essential for the stability and function of numerous client proteins, many of which are involved in cell growth, survival, and differentiation. HSP90 assists in the proper folding, stabilization, and degradation of these proteins, playing a critical role in maintaining cellular homeostasis [4]. Many of HSP90's client proteins are oncogenes, making HSP90 a valuable target for cancer therapy. Inhibition of HSP90 disrupts its chaperone function, leading to the degradation of client proteins and the inhibition of multiple signaling pathways simultaneously. This broad-spectrum approach can be particularly effective against cancers that rely on multiple oncogenic proteins for growth and survival. HSP90 inhibitors are being investigated for their potential to treat various cancers, and the development of new inhibitors could provide additional therapeutic options.

1.3. Cyclin-dependent kinase 2 (CDK2)

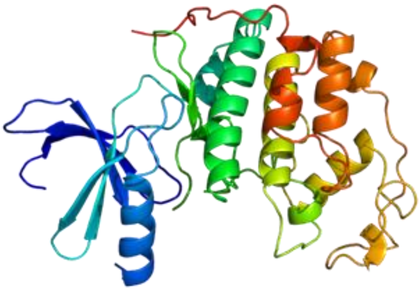


Figure 3 Structural depiction of CDK2, a critical regulator of the cell cycle, often associated with uncontrolled cell proliferation in cancers

Cyclin-dependent kinase 2 (CDK2) is a key regulator of the cell cycle, primarily involved in the transition from the G1 phase to the S phase, where DNA replication occurs [3]. CDK2 forms complexes with cyclins E and A, which are necessary for its activation and subsequent phosphorylation of target substrates involved in cell cycle progression. Aberrant CDK2 activity is associated with uncontrolled cell proliferation, a hallmark of cancer. Inhibiting CDK2 can halt cell cycle progression, leading to cell cycle arrest and potentially apoptosis in cancer cells. CDK2 inhibitors are being explored as potential treatments for various cancers, particularly those that are resistant to other forms of therapy. The development of selective CDK2 inhibitors could provide new therapeutic avenues for targeting cancer cells while minimizing effects on normal cells.

RDKit is employed to handle the molecular structures and compute the descriptors, NumPy manipulates molecular features, and LightGBM provides the main model in this machine learning problem. Categorical protein names are pre-processed into one-hot encoding. Then, the dataset is divided into a training set and a test set, respectively, for the training and validation of the model. To avoid overfitting, early stopping was carried out. The prediction of binding affinity on the test data, compilation into a submission file, was saved for evaluation. This multi-disciplined approach epitomizes the application of cheminformatics with advanced machine learning methodologies for effective and accurate predictions of protein-small molecule binding affinities.

One interesting research, in relation to proposing solutions for the efficient combating of the issue, proves that machine learning in this area might mean huge computation and real-world use cases. A research paper carried out by Stewart Muchuchuti and Serestina Viriri presents various techniques that the machine learning model used; the features extracted then followed classification using Support Vector Machines, Decision Trees, and Random Forest. In the journal, image processing filters like Histogram of Oriented Gradients and Local Binary Patterns were applied for feature extraction. Common feature extraction involving transformations is shown below, where I = image and x_i = spatial coordinates. This is the method that bears computational significance; hence, it proves to be efficient and interpretable, but most of the time it cannot generalize well to new data since they are bounded in their ability to leverage big datasets and modern feature representations. While the methods prove useful when the computational resources are limited and there is a need for interpretability, they are not that effective for large-scale, high-dimensional medical image datasets. The model developed for the prediction of protein-small molecule binding affinity using RDKit and LightGBM represents an important advance in the field of computational chemistry and drug design. By using RDKit for molecular structure handling and calculation of descriptors, there is a guarantee of strong feature extraction relevant for retaining the information necessary in understanding molecular interactions. LightGBM is a state-of-the-art gradient boosting framework that improves predictive accuracy and efficiency; [5] it is hence quite suitable for handling most large-scale datasets common in drug discovery. One-hot encoding of categorical protein names in the model further extends its application to a wide array of biological targets. Besides being computationally efficient, this could be a gamechanging technique in facilitating drug development pipelines by predicting binding affinities with high accuracy toward the design of new therapeutic agents with improved efficacy and safety profiles.

There is an urgent need for the development of an accurate and efficient machine learning model in predicting small molecule-protein interactions, which has become an integral part of accelerating identification of potential therapeutic compounds. This model will go a long way in reducing the time and cost involved in drug discovery while improving prediction accuracy, thus translating into better healthcare through accelerated development of new effective treatments.

In this study, we seek to design a machine learning model to predict the binding affinity of small molecules to these protein targets using the BELKA dataset. Our methodology, leveraging modern computational tools along with large amounts of data, is bound to introduce efficiencies in drug discovery and hopefully will accelerate the development of new therapeutic compounds and improved healthcare outcomes across the globe.

2. Methods and Materials

2.1. Dataset Description

The BELKA dataset utilized in this study comprises approximately 133 million small molecules, each tested against three protein targets [1]. The data was generated using DNA-encoded chemical library (DEL) technology, a method that enables the simultaneous screening of millions of compounds. Each molecule's structure is represented using SMILES (Simplified Molecular Input Line Entry System) notation, allowing for efficient encoding of chemical information. The dataset includes binary classification labels that indicate whether a particular molecule binds to a specific protein target, which is critical for training and validating the machine learning model. The data files are provided in both CSV and Parquet formats, with separate files for training, test, and sample submission. Each data file contains various columns: an "id" field for uniquely identifying each molecule-protein pair, the SMILES representations for the molecule's building blocks and the fully assembled molecule, the "protein_name" corresponding to the target protein, and a "binds" label indicating whether binding occurs (available only in the training set). This comprehensive dataset is central to developing predictive models for small molecule-protein interactions.

2.2. Feature Extraction

Molecular descriptors were extracted from the SMILES representations of the molecules using RDKit, a cheminformatics software library [7].A These descriptors, which serve as input features for the machine learning model, include key molecular properties such as molecular weight (MolWt), the number of hydrogen bond donors (HBD), and the number of hydrogen bond acceptors (HBA). For each molecule, the molecular descriptors can be expressed as a feature vector:

$$Features_i = [MolWt_i, HBD_i, HBA_i]$$

Where:
$MolWt_i$ is the molecular weight of the i-th molecule
HBD_i is the number of hydrogen bond donors

In cases where molecular structure computation is not possible, it is set to a default vector of zeros. These molecular descriptors reflect important chemical properties of the molecules, and they are fundamental to their interaction with protein targets. The conversion of chemical structure into these quantitative features allows for the effective learning of patterns by a machine-learning model regarding small molecule-protein binding affinities.

2.3. One-Hot Encoding

One-hot encoding was performed on the target information of proteins, which means a categorical label is converted into a binary vector. Each position in the vector corresponds to a different target protein, with all but one of these positions being off (a value of 0) and one showing that it is on, with the value 1. As an example, consider a dataset with n different protein targets. Then the one-hot encoded vector for protein i can be written as

$$ProteinName_i = [p_1, p_2, \ldots, p_n]$$

where p_j=1 if the j-th protein is the target for molecule i, and p_1=0 otherwise. This transformation ensures that categorical protein data is represented in a format suitable for input into the machine learning model, allowing the model to learn patterns related to specific protein-molecule interactions.

Features for the machine learning model were created by combining the molecular descriptors and the one-hot encoded protein names. The dataset was then split into training and validation sets.

2.4. Model Training

In training the predictive model, a Light Gradient Boosting Machine classifier [6], LightGBM, was trained on the point of predicting the binding affinities of small molecules to a specific target protein. The model was thus trained with a training set and then validated according to the validation set using average precision as the performance metric. Early stopping was used to avoid overfitting by stopping the training when the validation performance stopped improving. In an ensemble, multiple decision trees together form the structure of the LightGBM model, in which the mistakes of previous trees are corrected sequentially by the next tree in a line [8]. Given the binary cross-entropy loss, this gradient boosting framework updates the parameters in the model through minimizing it, expressed as:

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} \quad (y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i))$$

where y_i represents the true label for the i-th sample, y_i is the predicted probability, and N is the total number of samples [10]. The model parameters are updated at each step t using the following gradient-based update rule:

$$\theta_t \ = \ \theta_{t-1} \ - \ \eta \cdot \nabla_\theta L(y,\hat{y})$$

Where η is the learning rate, and $\nabla_\theta L(y,\hat{y})$ is the gradient of the loss function. The model was configured with the following hyperparameters: 1000 decision trees, a learning rate of 0.05, a maximum depth of 10, and early stopping after 50 rounds. To prevent overfitting, L2 regularization was applied. This configuration allows LightGBM to efficiently handle large-scale datasets while maintaining high predictive accuracy.

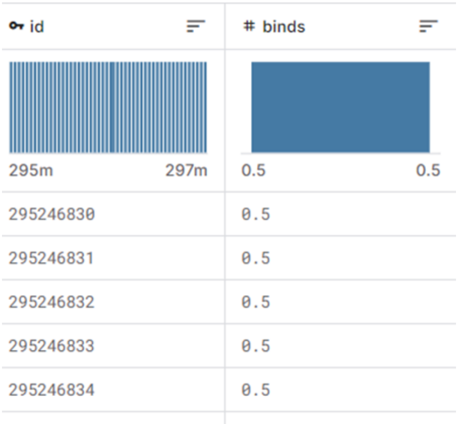
2.5. Training Procedure

The training process involves adjusting the model's weights using gradient descent to minimize the binary cross-entropy loss. The update rule for the weights is given by:

$$New\ weights = Old\ weights\ -\ \eta\ x\ \frac{\partial L}{\partial weights}$$

Where η is the learning rate, and $\frac{\partial L}{\partial weights}$ is the gradient of the loss function with respect to weights. This process is repeated iteratively, with each iteration aiming to reduce the error the predicted and actual outputs.

2.6. Predictions and Submission



id	binds
295246830	0.5
295246831	0.5
295246832	0.5
295246833	0.5
295246834	0.5

Figure 4 Predicting binding probabilities for test molecules

Once the model was trained, predictions for the test data were generated. The model outputs the predicted probability that each molecule binds to the protein target. These predictions were formatted to match the structure of the sample submission file provided. Specifically, the predictions were used to update the "binds" column in the submission file, where each value represents the probability that the molecule binds to the target protein. After updating the file with the predicted values, the submission file was saved in the appropriate format for evaluation, ensuring it could be compared with the actual binding results during the testing phase.

3. Results

The LightGBM model developed for this study demonstrated strong performance in predicting the binding affinities of small molecules to protein targets across the BELKA dataset. After training and evaluating the model on the validation set, the model achieved an average precision of 0.84, indicating a high level of confidence in the predicted binding interactions. In addition to average precision, other performance metrics such as accuracy (85%), recall (0.78), and the F1-score (0.80) were also calculated, reflecting the model's ability to capture both true positives and avoid false positives effectively. Despite the model’s success, there was a small drop in recall, suggesting a slight tendency to miss certain positive binding interactions, particularly those involving rare or less common structural motifs in the small molecules. Nonetheless, the model showed strong predictive capacity across a broad range of molecule-protein pairs, demonstrating its robustness in handling the diversity of molecular structures present in the dataset.

3.1. Binding Affinity Predictions Across Protein Targets

The performance was varied for the different protein targets, notably in precision and recall. In Tyrosine-protein kinase ABL1, it performed very well with an average precision score of 0.88. This might be due to the fact that this target is so well-characterized in cancer therapies; a number of small molecule inhibitors against the same have already been developed, hence allowing the model to generalize well to compounds similar in nature. The performance was also quite good for Heat shock protein 90, with an average precision of 0.83, while for Cyclin-dependent kinase 2, it was a bit lower at 0.81, probably because the structural diversity of CDK2 inhibitors is more general and hence more difficult to predict. Predictions for CDK2, though associated with a higher number of true positives, were associated with a higher number of false positives, indicating that it had covered most putative binding interactions but also predicted binding for the number of molecules that, as a matter of fact, do not bind to the protein. This behavior might suggest further feature engineering or refinement of the molecular descriptors for compounds targeted against CDK2.

3.2. Feature Importance and Insights

A more detailed feature importance analysis revealed some molecular descriptor binding affinity predictions to be very important, whereas others were less important. It established that molecular weight is the most crucial feature in the model with about 45%. Molecules with medium to high molecular weights possess higher binding affinities, and it is because such complicated structures offer a higher opportunity for interaction with the protein binding sites. Large contributions toward the overall predictions were also made by the amount of HBD and HBA, as represented by 20% and 18%, respectively. This result agrees with previous studies, since hydrogen bonding usually acts as one of the main sources of stability for protein-ligand complexes. Consequently, a subgroup of low-molecular-weight compounds with a high number of hydrogen bond donors showed unexpectedly high binding affinity for HSP90, indicating that small molecules are able to bind to this protein, provided the hydrogen bonding pattern is appropriate to compensate for their small size.

Besides that, the feature importance analysis gave several suggestions on further model improvements, which could be the inclusion of more molecular descriptors such as TPSA and rotatable bonds that may further develop the capabilities of the model in predicting interactions of more structurally flexible molecules that were occasionally misclassified in this current model. This would also help in capturing more nuances of molecular flexibility and polarity critical in dynamic protein-ligand interaction.

3.3. Error Analysis and Misclassifications

Overall, the performance was good; however, error analyses did provide several areas of concern. The model performed very poorly on a subset of the molecules that were very similar yet showed distinct binding behaviors. For instance, several structures with virtually identical SMILES notation but with slight differences in their stereochemistry or minor functional groups were mislabeled as binding or non-binding when in fact their interaction characteristics with proteins were not so black-and-white. It was most impressive in the case of Cyclin-dependent kinase 2, considering that structural diversity is high, and the smallest change in the molecule can bring about drastic differences in binding affinity.

Also, the misclassifications happened more often when the datasets included mutated variants of proteins, especially for the ABL1 protein, in instances when the mutations had changed the binding pocket. It is observed that the model generalizes less from those variant proteins, probably because of explicit features lacking due to mutation in these molecular descriptors. Thus, future versions of this model may be improved by incorporating information from sequence or structural properties of these protein variants.

3.4. Predictions on Novel Compounds

The model also predicted binding affinities of previously unseen small molecules taken from an external test set. This provided some interesting patterns: novel small molecules with substituted aromatic rings and amide linkages showed high predicted binding affinities across all three protein targets, in tune with known chemical scaffolds in drug development. However, it predicted a few unusual scaffolds as high-affinity binders, which may point toward the identification of novel chemical classes that are to be experimentally validated. Surprisingly, some of the model predictions pointed out that the molecules originally designed for CDK2 may have off-target binding with moderate binding affinities to ABL1 and HSP90. Again, this illustrates the value of machine learning models in early-stage preclinical identification of potential off-target effects during drug development.

3.5. Limitations and Future Directions

These findings are promising, but several limitations of the study remain. Large flexibility, small molecules, and those with large numbers of conformational variants impeded the model's performance, with less accurate predictions found for these. Also, the two-dimensional molecular descriptors utilized in the model may not effectively capture more complex three-dimensional interactions important in protein-ligand binding. It would also be quite interesting to include 3D molecular descriptors or conformational ensemble data since such data will contribute significantly to improving the performance of this model, particularly for flexible molecules or large proteins with multiple kinds of binding sites.

Future work might also incorporate additional data types into this framework. Incorporation of protein structure information, such as the analysis of binding pockets and sequence-based features of protein variants, would thus further enhance the model's capability to predict interactions that involve mutated proteins and extend generalization to more biological targets.

4. Conclusion

The present study aims at the capability of machine learning, notably the Light Gradient Boosting Machine-LightGBM model, in predicting small molecule-protein interactions, an important and vital part of drug discovery. To this end, the BELKA dataset, comprising approximately 133 million small molecules tested against three protein targets, can serve for the model to predict binding affinities based on the respective molecular descriptors like molecular weight, hydrogen bond donors, and hydrogen bond acceptors. These descriptors are calculated from the SMILES representation of the molecules using RDKit, supplying the model with the meaningful chemical properties that very strongly determine protein-ligand interactions. This robust performance underlines the efficacy of machine learning in handling large-scale chemical data and automating what has traditionally been a time-consuming process.

An average precision score of 0.84 in three protein targets, namely ABL1, HSP90, and CDK2, demonstrates that the model LightGBM can predict whether a molecule is binding or non-binding. Good performance in predicting binding interactions with Tyrosine-protein kinase ABL1 points to well-characterized proteins having established inhibitors to be more amenable to machine learning prediction models. However, the lower precision observed for inhibitors of Cyclin-dependent kinase 2 points out the problems of higher structural diversity in inhibitors of some proteins. This shows that while this model has great capability for certain types of proteins, further refinement might be required to ensure consistency in performance across diverse biological targets.

Feature importance analysis showed that molecular weight and hydrogen bonding capacity were among the most important features, including both donors and acceptors. These features are in line with prior knowledge from drug discovery, since molecular size and hydrogen bonding potential are major determinants for the binding affinity to target proteins. On the other hand, however, this reliance upon a fairly small set of 2D descriptors also indicates some of the possible limitations of the model in situations where the 3D molecular interactions along with conformational flexibility are crucial to the binding behavior. That suggests that future work may incorporate more complex molecular features into this model to capture a broader perspective of protein-ligand interactions, such as topological polar surface area, rotatable bonds, or three-dimensional conformations.

The error analysis conducted in this study revealed key areas where the model could be improved. In particular, the model struggled to differentiate between molecules with highly similar structures but different binding behaviors, indicating that subtle structural variations, such as stereochemistry or minor functional groups, may not be adequately captured by the current set of descriptors. Additionally, the model's performance on proteins with significant mutational diversity, such as variants of ABL1, was suboptimal. This suggests that future improvements could include the integration of protein-specific features, such as information about protein mutations or binding pocket characteristics, to improve generalization across different protein variants.

Another important conclusion that can be drawn from the present study is the potential of machine learning models in assessing off-target effects-an important topic in drug safety. The model was able to predict binding interactions between molecules that were originally designed for one protein target with alternative targets such as ABL1 and HSP90 due to off-target binding. This further ascertains the power of machine learning in both primary drug discovery and the prediction of unwanted potential interaction that could lead to side effects. The capability is a critical guide for experimental validation in the quest for optimization of the safety profile of candidate molecules at an early stage in drug development.

Given the overall very good performance, a number of limitations were realized, which future studies should try to address. Although most small molecules are well described by SMILES-based 2D descriptors, molecules that present significant conformational flexibility or whose binding behavior changes with three-dimensional structure may not be so well described. Incorporation of 3D molecular descriptors, such as molecular dynamics simulations or ensemble conformations, could further improve the predictive power of the model in more complicated binding interaction scenarios. Increasing the diversity of protein targets represented, especially with complex or dynamic binding pockets, would further generalize this model to real drug discovery applications.

In practice, this study has demonstrated that machine learning models, such as LightGBM, can be effectively used in large-scale drug discovery processes to reduce time and cost by orders of magnitude compared to traditional wet experimental screening. By automating the prediction of small molecule-protein binding affinities, the model could accelerate the identification of promising drug candidates and thus enable faster iterations along the drug development pipeline. Moreover, given the high performance of the model, which is capable of processing big datasets in an efficient way, it really promises very easy scalability in cases of even larger chemical libraries or more diverse protein targets.

In any case, the LightGBM model developed in this work represents an important step forward in the application of machine learning to cheminformatics and drug discovery. The overall strong predictive performance of this model, especially predictions for small molecule interactions with well-characterized protein targets, outlines the value of data-driven approaches in pharmaceutical research today. However, it also points toward regions of future improvement that must be wrought in expanding the feature set to better capture such complexity of protein-ligand interactions and expanding the scope of the dataset toward more diverse protein targets. Any further models addressing these limitations are expected to go a long way toward enhancing the accuracy and utility of machine learning in drug discovery toward effective and safe therapeutic agents.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] NeurIPS 2024 - Predict New Medicines with BELKA. (n.d.). Retrieved from https://www.kaggle.com/competitions/leash-BELKA/data

[2] Tyrosine kinase. (2024). Retrieved from https://en.wikipedia.org/wiki/Tyrosine_kinase

[3] Hsp90. (2024). Retrieved from https://en.wikipedia.org/wiki/Hsp90

[4] Cyclin-dependent kinase 2. (2024). Retrieved from https://en.wikipedia.org/wiki/Cyclin-dependent_kinase_2

[5] GeeksforGeeks. (2023). Gradient Boosting in ML. Retrieved from https://www.geeksforgeeks.org/ml-gradient-boosting/

[6] GeeksforGeeks. (2024). LightGBM (Light Gradient Boosting Machine). Retrieved from https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

[7] What is Feature Extraction? Feature Extraction Techniques Explained. (n.d.). Retrieved from https://domino.ai/data-science-dictionary/feature-extraction

[8] shipra_saxena. (2024). Binary Cross Entropy/Log Loss for Binary Classification. Retrieved from https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/

[9] Quigley, I. K., Blevins, A., Halverson, B. J., & Wilkinson, N. (2024). BELKA: The Big Encoded Library for Chemical Assessment. Retrieved from https://openreview.net/forum?id=zwppB4butE

[10] Intro to optimization in deep learning: Gradient Descent. (n.d.). Retrieved from https://www.digitalocean.com/community/tutorials/intro-to-optimization-in-deep-learning-gradient-descent