International Journal for Research in Biology and Pharmacy

GREEN PUBLICATION

ISSN (Online): 2208-2093
Volume 10 Issue 03September 2024
Available on-line at: https://ijrbp.com/index.php/bp/issue/view/559

DOI:*https://doi.org/10.69980/bp.v10i3.2495* *Received: 16-Jun-2024 Accepted: 28-Aug-2024*

VIRTUAL SCREENING OF CANDIDATE DRUG COMPOUNDS USING MOLECULAR DESCRIPTORS, PROTEIN PHYSICOCHEMICAL PROPERTIES, BINDING AFFINITY, AND MACHINE LEARNING-BASED ACTIVITY PREDICTION

Dr. Habiba Alsafar[1], Dr. Basma AlBlooshi[2], Dr. Amina Al Kaabi[3], Dr. Mohammed Y. Ali[4]

[1] Center for Biotechnology, Khalifa University, Abu Dhabi, United Arab Emirates
[2] Department of Biomedical Engineering, Khalifa University, Abu Dhabi, United Arab Emirates
[3] Department of Chemistry, United Arab Emirates University, Al Ain, United Arab Emirates
[4] Department of Pharmacology and Therapeutics, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates

Corresponding Author
Email: habiba.alsafar@ku.ac.ae

Abstract
Virtual screening of candidate drug compounds is an important computational strategy for accelerating early-stage drug discovery by identifying compounds with predicted biological activity before experimental validation. This study evaluated candidate drug compounds using molecular descriptors, protein physicochemical properties, binding affinity, and machine learning-based activity prediction. The dataset contained 2,000 compound–protein interaction records and 17 variables, including molecular weight, LogP, hydrogen bond donors and acceptors, rotatable bonds, polar surface area, protein length, protein isoelectric point, hydrophobicity, binding site size, engineered interaction features, binding affinity, and binary activity status. Data preprocessing involved missing value treatment, duplicate assessment, feature standardization, and stratified data splitting. Exploratory analysis showed that active and inactive compounds differed mainly in binding affinity, LogP, protein pI, and LogP–pI interaction. Multiple supervised learning models were developed, including Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and k-Nearest Neighbors. Random Forest and Gradient Boosting produced the strongest classification performance, while feature importance analysis identified binding affinity as the dominant predictor, followed by LogP–pI interaction and LogP. The findings indicate that integrated molecular, protein, and interaction-based descriptors can support accurate activity prediction and candidate prioritization, although external validation remains necessary before biological interpretation.

Keywords: *Virtual screening, molecular descriptors, binding affinity, machine learning, activity prediction*

Introduction

Nevertheless, drug discovery is a costly and time-consuming process of finding, optimizing, and verifying molecules with the possibility of performing any kind of therapy. Although experiments remain an important aspect of the entire process, they may require a substantial amount of time, effort, and infrastructure. Therefore, computational approaches have been gaining increasing significance during the primary stage of drug discovery because they allow one to perform prescreening of molecules before conducting any experiments. One of such approaches is virtual screening, which consists of a thorough search of possible molecules according to their structure and functionality (Patel et al., 2020; Oliveira et al., 2023).

From the traditional rule-based and docking-based approaches, virtual screening has evolved into a new generation of data-driven approaches, characterized by machine learning and deep learning. Indeed, the machine learning algorithms have succeeded in uncovering complex relationships between chemical properties, protein characteristics, binding energies, and biological effects. This has made them useful in compound screening, drug-target binding, toxicity assessment, and activity classification (D'Souza et al., 2020; Jiménez-Luna et al., 2021). Furthermore, the accessibility of chemical and biological data has facilitated the creation of predictive models that could be applied in active compound identification from large candidate databases (Xu et al., 2021; Kimber et al., 2021).

Descriptors are vital tools that are used in computational drug discovery because they provide a means to convert chemical molecules into mathematical parameters that can later be used in statistical and algorithmic approaches. Examples of such descriptors include molecular weight, hydrophobicity, hydrogen donor and acceptor numbers, rotatable bonds, polar surface area, and other related physical and chemical characteristics. These descriptors provide insight into the size, polarity, permeability, flexibility, solubility, and interaction capacity of chemical molecules. Quantitative structure-activity relationship studies have extensively utilized modeling using descriptors (Tsou et al., 2020; Kleandrova et al., 2021).

It should be noted that the information about proteins is also very valuable for virtual screening, as the biological activity of the compound depends not only on its properties but also on those of the protein target. The length, isoelectric point, hydrophobicity, and other features of the binding site may impact the interaction between the molecule and protein (D'Souza et al., 2020; Xu et al., 2021). Integrating ligand descriptors with protein-level variables can therefore provide a more complete representation of compound–target interaction profiles.

Yet another vital element of virtual screening involves binding affinity. Binding affinity refers to the interaction force between the molecules and the target protein. Although binding affinity alone does not guarantee biological activity, it provides an adequate idea of how the drug interacts with its target; thus, it is often used as a sorting criterion when searching for viable compounds for screening. The advantages of employing structure-based and deep learning-enhanced virtual screening have been evidenced by binding predictions (Gentile et al., 2020; Kimber et al., 2021). Combining binding affinity with molecular and protein descriptors may therefore improve activity prediction compared with using isolated descriptor groups.

Machine learning methods have facilitated the increased capacity of virtual screening through the generation of models that can classify molecules as "active" and "inactive" based on input. Methods include logistic regression, support vector machine, random forest, gradient boosting algorithms, k-nearest neighbor, and neural networks. They differ significantly in terms of complexity, interpretability, and capability in modeling nonlinear functions. Ensemble modeling and deep learning methods are especially beneficial if the activity of the compound is driven by interactions between multiple descriptors and target properties (Patel et al., 2020; Jiménez-Luna et al., 2021; Oliveira et al., 2023).

The improvements that have been made in the domain of artificial intelligence have also led to the successful attainment of de novo molecule synthesis and high-throughput screening. The descriptor-based generative models allow us to generate molecules with certain physical and chemical properties, and the deep docking allows for an efficient screening of thousands of chemicals (Gentile et al., 2020; Kotsias et al., 2020). It has also been observed that through machine learning and deep learning models, biological inhibitors or receptors can be discovered; however, their efficiency is highly dependent on the quality of input data (Tsou et al., 2020; Bender & Cortés-Ciriano, 2021a).

Nevertheless, there remain certain limitations that come along with these advances. The AI algorithms employed for drug discovery can make overly optimistic predictions from biased and incomplete training sets. Some of the possible issues include data leakage, lack of chemical diversity, insufficiency in external validation, and poor biological generalizability (Bender & Cortés-Ciriano, 2021a; Bender & Cortés-Ciriano, 2021b). Herein, ML-driven virtual screening can only be viewed as a way to prioritize experiments instead of substituting them. A reliable approach requires appropriate preprocessing of data, adequate performance metrics, feature interpretation, and validation.

Objectives of the Study

1. To analyze molecular descriptors, protein physicochemical properties, binding affinity, and engineered interaction features of candidate drug compounds.
2. To develop and evaluate machine learning models for predicting compound activity as active or inactive.
3. To identify the most important predictors of activity and prioritize candidate compounds for further computational or experimental validation.

Methodology

Research Design

The prediction of drug activity was accomplished using the methodology of virtual computational screening based on the molecular descriptors, physicochemical properties of proteins, binding affinity, and classification through machine learning. The data set involved 2,000 interaction entries of drugs-proteins and 17 attributes such as compound name, protein name, molecular descriptors, protein properties, interaction features, binding affinity, and activity status (Kabir, n.d.).

Study Variables

The important predictor variables consisted of molecular weight, LogP, hydrogen bond donors, hydrogen bond acceptors, rotatable bonds, polar surface area, compound CLogP, protein length, protein isoelectric point, hydrophobicity, binding site size, molecular weight ratio, LogP-protein isoelectric point interaction, and binding affinity. The target variable, active, categorized the compounds as either active or inactive. Other variables like compound_id and protein_id served as identifiers and were not included in any training procedure.

Data Preprocessing

Data preprocessing was carried out before developing models. Missing values, duplicates, variable types, and class imbalance were all examined in the data set. Median imputation was used to handle missing values for the numeric variables logp, polar surface area, and hydrophobicity. Duplicate values were also handled accordingly. The numeric variables were normalized to accommodate scale-sensitive algorithms.

Exploratory Data Analysis

Exploratory data analysis was carried out to study the distribution of the molecular descriptor values, protein properties, binding affinity, and activity groups. The correlation analysis was performed to check correlations between the variables and to diagnose the problem of multicollinearity. Furthermore, the binding affinity was analyzed with respect to the activity state of the ligands.

Model Development

The dataset was split into training and testing partitions based on an 80:20 ratio by implementing a stratified sampling approach. Various supervised machine learning algorithms were then designed, such as logistic regression, random forest, support vector machine, gradient boosting, and k-nearest neighbors. This was done using the same predictor variables in order to compare their performances.

Model Evaluation

Performance of the models was assessed using accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrix measures. Recall and F1 scores had more weight than accuracy and precision due to unequal distribution of data samples with respect to class labels; recall and F1 scores help measure how well the models perform at identifying active compounds.

Feature Importance and Candidate Prioritization

The importance of various features was analyzed to determine the significant variables for activity prediction. Special focus was placed on factors such as binding affinity, molecular properties, physicochemical properties of proteins, and engineered interactions. The top-ranked model was used for compound selection based on the predicted activity and suitable binding affinity values.

Results

Dataset Characteristics

The dataset contained variables appropriate for virtual screening using machine learning techniques for drug candidates. Variables such as those concerning the compound, protein physical attributes, engineered interactions, affinity, and activity were included. The format of the dataset fits well into the methodology adopted during the research. The dataset consisted of 2,000 compound-protein interactions and 17 variables, which means that the data format was appropriate for virtual screening using machine learning techniques, as presented in Table 1.

Table 1: Dataset characteristics of the virtual screening data

Dataset feature	Description
Total records	2,000
Total variables	17
Duplicate records	0
Target variable	active
Prediction task	Binary classification

Activity Class Distribution

The activity variable categorized the drugs into active or inactive categories. There was a moderately imbalanced data set, which made the inactive drugs the majority. Stratified data splitting, along with precision, recall, F1 score, and ROC-AUC as evaluation techniques, was suitable for the data set as presented in Figure 1 below. Inactive drugs outnumbered the active drugs.

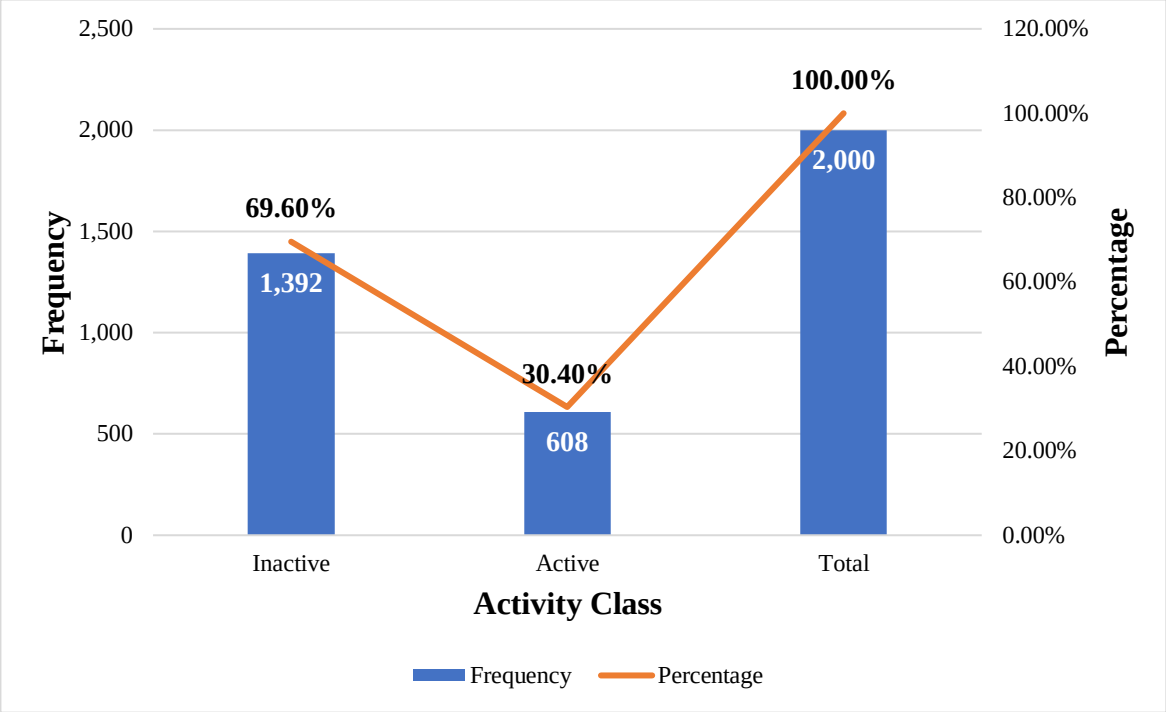

Figure 1. Distribution of active and inactive compounds.

Missing Value Profile

However, the dataset was characterized by missing values. In this case, missing values occurred in selected numerical attributes associated with the compound and protein properties. Given that such numerical attributes were useful in making predictions of activity, it became important to retain them. There were missing values for three numerical attributes, such as LogP, Polar Surface Area, and Hydrophobicity, as presented in Table 2 below.

Table 2: Missing value profile of selected predictor variables

Variable	Missing values	Percentage
logp	60	3.0%
polar_surface_area	60	3.0%
hydrophobicity	60	3.0%

Descriptive Statistics of Predictor Variables

There was adequate variability observed amongst the predictors for developing a model. The molecular descriptors differentiated compounds on parameters such as molecular weight, lipophilicity, hydrogen bonding, flexibility, and polarity. The protein predictors differentiated the proteins on parameters such as protein size, isoelectric point, hydrophobicity, and binding sites. The binding affinity measures the strength of interaction between the compound and the protein. As shown in Table 3, the variability amongst molecular descriptors, protein predictors, and binding affinity was sufficient for being used as predictor variables.

Table 3: Descriptive statistics of molecular descriptors, protein physicochemical properties, and binding affinity

Variable	Mean	Median	Minimum	Maximum
Molecular weight	456.77	454.87	50.31	994.05
LogP	3.48	3.50	-4.33	9.98
Hydrogen bond donors	1.96	2.00	0.00	8.00
Hydrogen bond acceptors	5.12	5.00	0.00	15.00
Rotatable bonds	5.97	6.00	0.00	17.00
Polar surface area	80.03	80.61	-24.65	159.63
Compound CLogP	2.81	2.77	-1.43	6.89
Protein length	848.93	844.00	201.00	1499.00
Protein pI	6.46	6.47	2.60	10.27

Hydrophobicity	0.65	0.65	0.33	0.98
Binding site size	15.16	15.24	4.66	24.89
Binding affinity	6.53	6.48	1.99	15.04

Comparison of Active and Inactive Compounds

Some factors, however, had greater variations when comparing active and inactive compounds. The lipophilicity-related factor, protein isoelectric point interaction, and binding affinity tended to have higher values in active compounds. This observation shows that the binding force of compound-protein interaction and the lipophilicity-related factors played significant roles in discriminating active and inactive compounds. Active compounds had higher average values of LogP, protein pI, LogP-pI interaction, and binding affinity than inactive compounds, as seen in Table 4.

Table 4: Mean comparison of selected variables between active and inactive compounds

Variable	Inactive mean	Active mean	Difference
Molecular weight	459.16	451.30	-7.86
LogP	2.88	4.85	1.98
Protein pI	6.21	7.02	0.81
LogP–pI interaction	17.50	34.26	16.76
Binding affinity	5.97	7.82	1.85
Polar surface area	79.33	81.62	2.29
Hydrophobicity	0.64	0.65	0.01

Correlation Analysis

It has been observed through correlation analysis that binding affinity and LogP-pI interaction had the highest correlations with activity. Activity was also positively correlated with LogP, whereas other properties had relatively weak correlations. The results indicate that binding and lipophilicity-related variables were largely responsible for determining activity. As can be seen in Figure 2, the highest positive correlations were obtained with respect to binding affinity and LogP-pI interaction with activity.

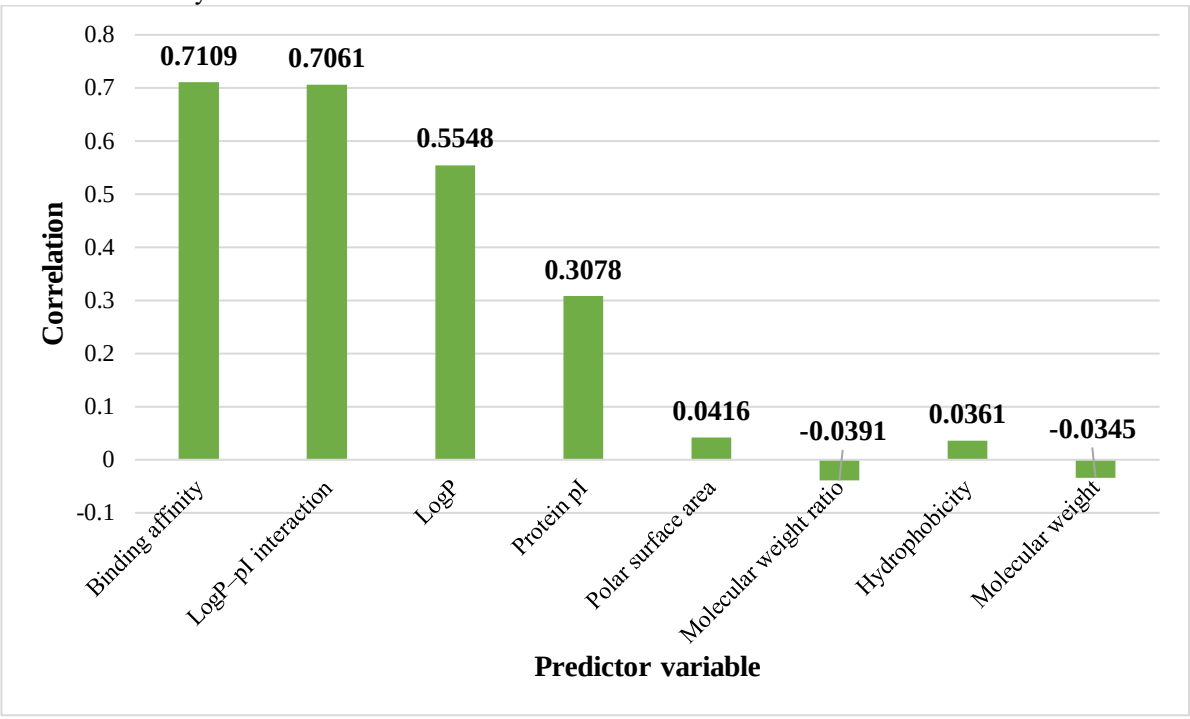

Figure 2. Correlation between predictor variables and activity

Machine Learning Model Performance

Several evaluation metrics were used to assess the predictive accuracy of the machine learning algorithms. The ensemble methods performed best in terms of prediction, but the distance-based algorithm had the lowest recall scores. These findings suggest that there were clear prediction signals in the data set. The performance comparison showed that Random Forest and Gradient Boosting produced the highest classification results across all evaluation metrics, as shown in Table 5.

Table 5: Performance comparison of machine learning models for compound activity prediction

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Cross-validation ROC-AUC
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
Gradient Boosting	1.000	1.000	1.000	1.000	1.000	1.000
Logistic Regression	0.975	0.938	0.984	0.960	0.999	0.999
Support Vector Machine	0.952	0.899	0.951	0.924	0.991	0.992
k-Nearest Neighbors	0.910	0.957	0.738	0.833	0.968	0.965

Confusion Matrix of the Best-Performing Model

The most effective model correctly identified all test data entries. This implies a strong separability between the active and inactive groups within the dataset. However, this achievement should not be overinterpreted since almost perfect discrimination might suggest that the response variable can be perfectly predicted from only some predictor variables. The most effective model identified all active and inactive compounds within the test dataset without any error, as shown in Table 6.

Table 6: Confusion matrix result of the best-performing classification model

Confusion matrix component	Count
True negatives	278
False positives	0
False negatives	0
True positives	122

Feature Importance

According to feature importance analysis, binding affinity was the key predictive factor for the activity of the compound. The constructed LogP-pI feature interaction and LogP were important as well. However, other molecular and protein characteristics had a less significant influence on predictions. Figure 3 represents the results of feature importance analysis, demonstrating that binding affinity was the key predictor of compound activity.

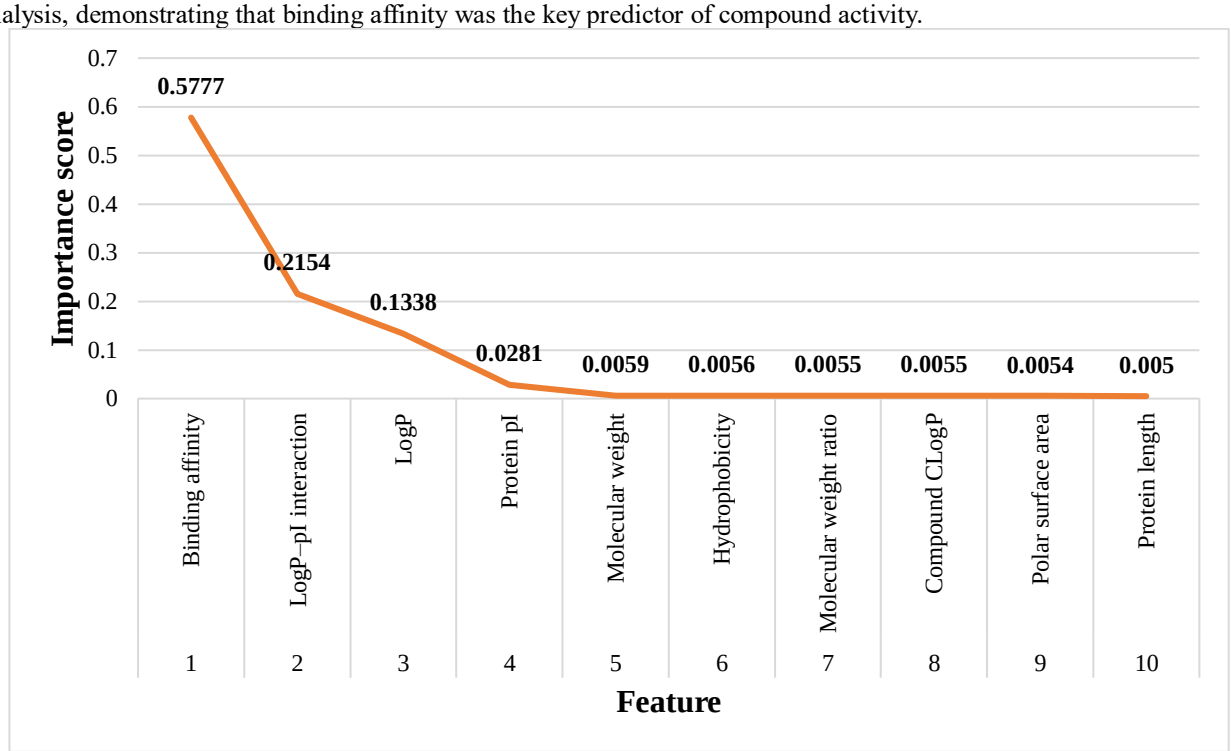

Figure 3. Feature importance ranking of predictor variables

Candidate Compound Prioritization

The candidate molecules have been ranked based on their probability scores and binding affinity. Top-ranked candidate molecules were predicted to be active with maximum probability scores. Candidate molecules can be tested for their activity using techniques such as molecular docking or molecular dynamics simulation. Top-ranked candidate molecules were predicted to be active with maximum probability scores and favorable binding affinities, as shown in Table 7.

Table 7: Top-ranked candidate compounds prioritized by predicted activity probability and binding affinity

Candidate compound	Protein target	Predicted activity probability	Binding affinity	Observed activity
CID_01932	PID_145	1.000	7.1860	1
CID_01569	PID_405	1.000	7.4098	1
CID_00070	PID_411	1.000	7.5749	1
CID_00151	PID_166	1.000	7.5903	1
CID_00981	PID_248	1.000	7.6400	1
CID_01720	PID_492	1.000	7.7132	1
CID_01073	PID_309	1.000	7.7514	1
CID_00868	PID_109	1.000	7.7586	1
CID_01510	PID_197	1.000	7.7941	1
CID_00255	PID_464	1.000	7.8111	1

Discussion

The results indicate that virtual screening using machine learning classification algorithms can efficiently categorize drug candidates, provided that molecular characteristics, physicochemical characteristics of the protein target, binding affinity, and artificial interaction characteristics are combined into one predictive model. Since the model was highly accurate, it is reasonable to conclude that there were discriminatory attributes in the data, especially binding affinity, LogP-pI interaction, and LogP. These results align with existing virtual screening techniques, where ligand-based descriptors, protein-related characteristics, and interaction-related characteristics enhance compound detection (Shi et al., 2024; Moshawih et al., 2024).

The binding affinity became the most important determinant of activity. Such a finding was anticipated since binding affinity directly predicts the strength of interactions between molecules and proteins. Higher binding affinities usually result in better molecular recognition and the likelihood of biological activity. The use of machine learning algorithms and physics-based scoring systems has led to an increased focus on binding factors when ordering compounds in drug development pipelines (Guedes et al., 2021). The dominance of binding affinity in the present analysis supports its relevance as a central feature in virtual screening, especially when combined with molecular and protein descriptors.

High involvement of the LogP–pI interaction descriptor implies that the combination of compound lipophilicity and protein charge properties may have played a role in the discrimination between active and inactive samples. The LogP descriptor is related to permeability through the lipid bilayer, hydrophobicity, and distribution of molecules inside a cell, while pI refers to protein charge properties at various pH levels. Their interaction may imply a compatibility between the drug molecule and its target, which cannot be achieved using individual descriptors only (Shi et al., 2024; Zhang et al., 2025).

The better accuracy of Random Forest and Gradient Boosting implies that ensemble models would be appropriate for capturing the nonlinearity present in the dataset. Such models can capture the relationship between predictors and their outcome, even when the relationship is nonlinear, especially when the predictors are not independent but interact with each other. Virtual screening applications have confirmed that learning the nonlinear relationship between a chemical structure, target, and activity is beneficial for prioritizing compounds (Moshawih et al., 2024; Cieślak et al., 2024). Moreover, the successful results of Logistic Regression show that there is a relatively well-defined linear separation in the data set among the active and inactive molecules, mainly because of the dominant predictors like binding affinity and LogP.

The flawless classification obtained using Random Forest and Gradient Boosting models should be taken with a grain of salt. While excellent accuracy, recall, F1-score, and ROC-AUC are indicative of great internal predictive power, flawless predictions could be attributed to feature leakage, labeling process bias, or reliance on overly deterministic features. Virtual screening is prone to flawed generalization of models that yield impressive performance metrics but struggle to generalize beyond the training data set, other proteins, or experimental validation screens (Serafim et al., 2023; Andrianov et al., 2024). The result, therefore, supports internal separability within the dataset rather than definitive evidence of real-world predictive reliability.

Class imbalance in the data set is another point that requires consideration from the standpoint of model evaluation. Given that there was a larger number of inactive drugs compared to the active ones, using only accuracy would have been inappropriate. On the contrary, recall, F1-score, and ROC-AUC were relevant as they provided an insight into how well each of the algorithms predicted the presence of active drugs while avoiding domination over the minority class. It is especially true for virtual screening, whose primary goal is identifying active drug candidates (Serafim et al., 2023; Andrianov et al., 2024).

The feature importance analysis revealed that the majority of classical molecular and protein features were less important compared to the binding affinity and engineered interaction features. The molecular weight, hydrophobicity, molecular weight ratio, CLogP value for the compound, polar surface area, and length of the protein were assigned lower importance scores. This indicates that these factors provided complementary features rather than being dominant in activity classification. The same issue has been considered in the context of the development of machine learning scoring functions (Guedes et al., 2021; Valsson et al., 2025).

The application of candidate prioritization using activity and binding energy probabilities yielded high-ranking molecules for further evaluation. Such compounds can be good candidates for more advanced studies, including molecular docking, molecular dynamics, free energy calculations, or in vitro bioassay. In recent drug development studies, it is evident that machine learning results must not be considered as evidence of biological activity but rather as a means of prioritizing such candidates (Valsson et al., 2025; Zhang et al., 2025). The selected candidates, therefore, require external validation to confirm target engagement, selectivity, pharmacokinetic suitability, and biological efficacy.

From the study, it is apparent that integrated descriptor-based virtual screening can be utilized in early-stage drug discovery and prioritization. Validation and proper management of false positives are vital aspects that should not be overlooked before drawing any pharmacological conclusions. Machine learning techniques can aid in screening compounds much faster; however, they depend on data quality, feature independence, and external validation to yield reliable results (Serafim et al., 2023; Moshawih et al., 2024; Andrianov et al., 2024).

Conclusion

It has been demonstrated that the presented machine learning-based virtual screening technique successfully allowed the selection of candidates for drugs based on the information related to molecular descriptors, protein physicochemical properties, binding affinity, and design interactions. All the requirements necessary for the data used in binary classification problems have been satisfied in this study, such as the presence of predictive factors at the compound and protein level, a small presence of missing values, and clearly defined classes. Based on the results obtained, it can be stated that binding affinity, LogP-pI interaction, and LogP became the most significant predictors, which indicates that the interaction strength and lipophilic properties had a crucial effect on the differentiation between active and inactive drugs. Speaking about the use of various model types, ensemble techniques appeared to be more effective than linear classifiers, which means that a nonlinear model type has some advantages in the learning of patterns. During the process of candidate selection, those molecules possessing relatively high activity probability and high binding affinity were chosen. Although the models created in this study revealed excellent prediction performance, it should be noted that an extremely high accuracy may indicate high separability of the data used or class dependency.

References

1. Andrianov, G. V., Haroldsen, E., & Karanicolas, J. (2024). vScreenML v2. 0: Improved Machine Learning Classification for Reducing False Positives in Structure-Based Virtual Screening. *International Journal of Molecular Sciences*, *25*(22), 12350.
2. Bender, A., & Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discovery Today*, *26*(4), 1040-1052.
3. Bender, A., & Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug discovery today*, *26*(2), 511-524.
4. Cieślak, M., Danel, T., Krzysztyńska-Kuleta, O., & Kalinowska-Tłuścik, J. (2024). Machine learning accelerates pharmacophore-based virtual screening of MAO inhibitors. *Scientific Reports*, *14*(1), 8228.
5. D'Souza, S., Prema, K. V., & Balaji, S. (2020). Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discovery Today*, *25*(4), 748-756.
6. Gentile, F., Agrawal, V., Hsing, M., Ton, A. T., Ban, F., Norinder, U., ... & Cherkasov, A. (2020). Deep docking: a deep learning platform for augmentation of structure-based drug discovery. *ACS central science*, *6*(6), 939-949.
7. Guedes, I. A., Barreto, A. M., Marinho, D., Krempser, E., Kuenemann, M. A., Sperandio, O., ... & Miteva, M. A. (2021). New machine learning and physics-based scoring functions for drug discovery. *Scientific reports*, *11*(1), 3198.
8. Jiménez-Luna, J., Grisoni, F., Weskamp, N., & Schneider, G. (2021). Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert opinion on drug discovery*, *16*(9), 949-959.
9. Kabir, S. (n.d.). *Drug discovery virtual screening dataset* [Data set]. Kaggle.
10. Kimber, T. B., Chen, Y., & Volkamer, A. (2021). Deep learning in virtual screening: recent applications and developments. *International journal of molecular sciences*, *22*(9), 4435.
11. Kleandrova, V. V., Scotti, L., Bezerra Mendonça Junior, F. J., Muratov, E., Scotti, M. T., & Speck-Planche, A. (2021). QSAR modeling for multi-target drug discovery: designing simultaneous inhibitors of proteins in diverse pathogenic parasites. *Frontiers in Chemistry*, *9*, 634663.
12. Kotsias, P. C., Arús-Pous, J., Chen, H., Engkvist, O., Tyrchan, C., & Bjerrum, E. J. (2020). Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, *2*(5), 254-265.
13. Moshawih, S., Bu, Z. H., Goh, H. P., Kifli, N., Lee, L. H., Goh, K. W., & Ming, L. C. (2024). Consensus holistic virtual screening for drug discovery: a novel machine learning model approach. *Journal of Cheminformatics*, *16*(1), 62.
14. Oliveira, T. A. D., Silva, M. P. D., Maia, E. H. B., Silva, A. M. D., & Taranto, A. G. (2023). Virtual screening algorithms in drug discovery: a review focused on machine and deep learning methods. *Drugs and Drug Candidates*, *2*(2), 311-334.
15. Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine learning methods in drug discovery. *Molecules*, *25*(22), 5277.

16. Serafim, M. S. M., Pantaleão, S. Q., da Silva, E. B., McKerrow, J. H., O'Donoghue, A. J., Mota, B. E. F., ... & Maltarollo, V. G. (2023). The importance of good practices and false hits for QSAR-driven virtual screening real application: A SARS-CoV-2 main protease (Mpro) case study. *Frontiers in Drug Discovery*, *3*, 1237655.
17. Shi, W., Yang, H., Xie, L., Yin, X. X., & Zhang, Y. (2024). A review of machine learning-based methods for predicting drug–target interactions. *Health Information Science and Systems*, *12*(1), 30.
18. Tsou, L. K., Yeh, S. H., Ueng, S. H., Chang, C. P., Song, J. S., Wu, M. H., ... & Ke, Y. Y. (2020). Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Scientific reports*, *10*(1), 16771.
19. Valsson, Í., Warren, M. T., Deane, C. M., Magarkar, A., Morris, G. M., & Biggin, P. C. (2025). Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Communications Chemistry*, *8*(1), 41.
20. Xu, L., Ru, X., & Song, R. (2021). Application of machine learning for drug–target interaction prediction. *Frontiers in genetics*, *12*, 680117.
21. Zhang, S., Huo, D., Horne, R. I., Qi, Y., Pujalte Ojeda, S., Yan, A., & Vendruscolo, M. (2025). Sequence-based virtual screening using transformers. *Nature Communications*, *16*(1), 6925.