ACS OMEGA

Open Access

Machine Learning-Driven Methods for Nanobody Affinity Prediction

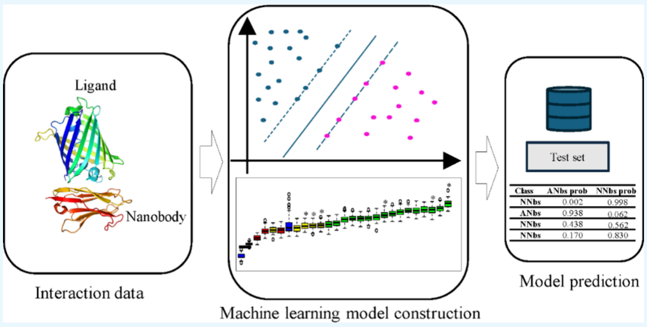
Hua Feng,[#] Xuefeng Sun,[#] Ning Li, Qian Xu, Qin Li, Shenli Zhang, Guangxu Xing, Gaiping Zhang, and Fangyu Wang*

Cite This: *ACS Omega* 2024, 9, 47893−47902

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Because of their high affinity, specificity, and environmental stability, nanobodies (Nbs) have continuously received attention from the field of biological research. However, it is tough work to obtain high-affinity Nbs using experimental methods. In the current study, 12 machine learning algorithms were compared in parallel to explore the potential patterns between Nb−ligand affinity and eight noncovalent interactions. After model comparison and optimization, four optimized models (SVMrB, RotFB, RFB, and C50B) and two stacked models (StackKNN and StackRF) based on nine uncorrelated (correlation coefficient <0.65) optimized models were selected. All the models showed an accuracy of around 0.70 and high specificity. Compared to the other models, RotFB and RFB were not capable of predicting nonaffinitive Nbs with lower precision (<0.44) but showed higher sensitivity at 0.6761 and 0.3521 and good model robustness (F1 score and MCC values). On the contrary, SVMrB, C50B, and StackKNN were able to effectively predict the future nonaffinitive Nbs (specificity >0.92) and reduce the number of true affinitive Nbs (precision >0.5). On the other hand, StackRF showed intermediate model performance. Furthermore, an in-depth feature analysis indicated that hydrogen bonding and aromatic-associated interactions were the key noncovalent interactions in determining Nb−ligand binding affinity. In summary, the current study provides, for the first time, a tool that can effectively predict whether there is an affinity between nanobodies and their intended ligands and explores the key factors that influence their affinity, which could improve the screening and design process of Nbs and accelerate the development of Nb drugs and applications.



Class	ANbs prob	NNbs prob
NNbs	0.002	0.998
ANbs	0.938	0.062
NNbs	0.438	0.562
NNbs	0.170	0.830

INTRODUCTION

Single-domain antibodies (sdAbs) are a special kind of heavy chain antibodies (HCAbs) with only the antigen-binding variable domain of traditional antibodies, which are derived from animals of Camelidae such as camels, llamas, and alpacas.[1] Because they have a low molecular weight of approximately 12−15 kDa and are about 1/10 the size of traditional antibodies (~150 kDa), sdAbs are also known as nanobodies (Nbs).[2] Compared with conventional antibodies, Nbs show obvious advantages, such as high affinity to their targets, high stability under extreme conditions, ease of modification, and low production cost, making Nbs versatile biomolecules widely used in various applications in biotechnology, therapeutics, and diagnostics.[3,4]

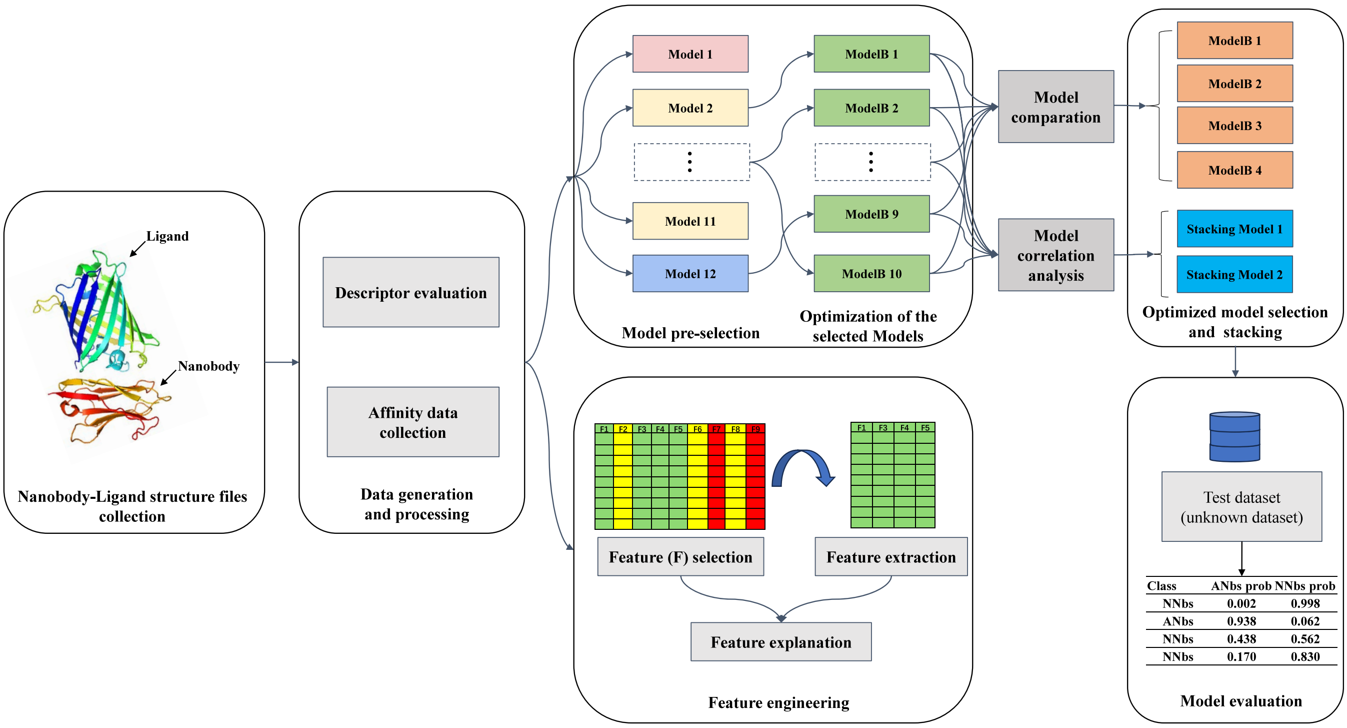
Among the above advantages of Nbs, the affinity of Nbs to their target protein is a critical determinant for their following various applications.[3,5] Currently, experimental methods, such as surface plasmon resonance (SPR), enzyme-linked immunosorbent assay (ELISA), or isothermal titration calorimetry (ITC), are normally used for analyzing the affinity of nanoantibodies, but these methods are time-consuming, are resource-intensive, and require significant amounts of purified proteins.[6] These limitations highlight the need for faster, more accurate, and less expensive methods to identify antibody affinities, which could accelerate the development of nanobody-related applications. The development of computer science and computational approaches has greatly accelerated the development of protein affinity evaluation. As the well-established traditional computational strategies, molecular docking and molecular dynamics simulation have been used to analyze protein−protein affinity using scoring functions, such as AutoDock,[7] HPEPDOCK,[8] ZDOCK,[9] GROMACS,[10] and HTMD.[11] Although these computational strategies have significantly improved the efficiency of high-affinity protein screening in the past decades, the prediction inaccuracy and significant consumption of computational resources are still shortcomings of these methods. Besides, the requirement for specialized expertise also makes these methods time-consuming and labor-intensive.[12−14]

With the increasing availability of huge protein affinity data, machine learning (ML), as a data-driven method, has emerged

Received: October 25, 2024
Revised: November 5, 2024
Accepted: November 12, 2024
Published: November 19, 2024

ACS Publications

https://doi.org/10.1021/acsomega.4c09718



* "B" represents the model/hyperparameter that configuration showed the best performance after optimization.

Figure 1. Workflow of the current study.

as a potential tool for protein affinity evaluation with its ability to discover hidden patterns in data.[12] Compared to the traditional computational methods, ML methods can enhance the performance of protein affinity evaluation, lower the consumption of computational resources, and meanwhile show a higher interpretability.[12,15] Ballester et al. predicted the affinity between proteins using a random forest model based on the occurrence number of protein−ligand atom type pairs, which was shown to be competitive when compared to methods that were popular at the time of the article's publication[16]; Romero-Molina et al. constructed a support vector machine model to evaluate the affinity of protein−peptide and protein−protein, which could also generate and rank protein mutants.[17] Ahmed et al. incorporated convolutional neural networks to predict the protein−ligand binding affinity, which showed a better performance compared to the some existing methods.[18] Undoubtedly, all these methods could make efficient predictions for large-scale protein data, which greatly improve the screening efficiency of affinitive proteins. More importantly, ML methods could drastically rule out nonaffinitive proteins and reduce the number of potential affinitive protein candidates for subsequent experimental verification, which could lower the risk of clinical trial failure and accelerate the procedure of drug development.[19] However, due to the unique structural features of Nbs, the currently available ML methods may not be capable of Nb affinity evaluation. Therefore, an affinity prediction tool for nanobodies is urgently needed to cater to the increasing amount of nanobody data and research needs in related fields.

In the current study, 12 ML algorithms were preliminarily used to learn the hidden pattern between Nb−ligand binding affinity type and different noncovalent interaction data of Nb−ligand complexes, which were further optimized for their hyperparameters. With the properly handed data sets, four optimal models and two stacking models built by nine models were selected and compared on the test data set. Finally, different models showed different predictive advantages, and all can effectively evaluate Nbs with affinity. In summary, for the first time, the current study constructed a series of ML models for evaluating the affinity between Nbs and their corresponding ligands, which could facilitate the discovery and screening of high-affinitive nanobodies. Figure 1 shows the procedures of model construction and prediction in the current study. The model and the related data can be downloaded from https://github.com/greenGM/Nbaffinity.

■ EXPERIMENTAL SECTION

Data Set Collection and Processing. A total of 991 pdb files of different Nbs and their corresponding ligands were collected from the RCSB-PDB database (https://www.rcsb.org/). In addition, the different descriptors of hydrophobic interactions (HI), disulfide bridges (DB), ionic interactions (IoInt), aromatic−aromatic interactions (AAI), aromatic−sulfur interactions (ASI), cation−pi interactions (CPI), hydrogen bonding main−main chain interactions (HBMM), hydrogen bonding main−side chain interactions (HBMS), and hydrogen bonding side−side chain interactions (HBSS) between the Nbs and their ligands were calculated by ProtInter (https://github.com/maxibor/protinter), which is a tool designed to calculate noncovalent interactions of single chain proteins in a protein complex pdb file. For each pair of Nbs and ligand, all noncovalent interactions mentioned above were calculated for eight parameters: number (count), mean, standard deviation (std), and the quartiles values (min, 25%, 50%, 75%, and max) of distances between key sites in all interaction regions of these two proteins as described in the user manual. To prevent model overfitting problems due to the scarcity and single pattern of the obtained nanobody−ligand interaction data, about 444 other pairs of protein−protein interaction data were added as supplementary data to increase the data variability, and at least one protein of each pair had a length of less than 250 amino acids, making them closer to the length of the nanobodies (about 120−150aa). Doing so could enhance the data learning of subsequent algorithms. All of the MIC values of these Nb−ligand pairs were collected from their corresponding published papers manually. Finally, 72 features, including all the noncovalent interaction data and MIC values, were obtained.

Then, all the data were further classified into two groups based on minimum inhibitory concentration (MIC) values: affinitive Nbs (ANb, MIC < 2000, n = 359) and nonaffinitive Nbs (NNb, MIC ≥ 2000, n = 1076). After removing the features with correlation coefficients higher than 0.7, 35 features were left, and all the obtained data were split into the training set and test set by a ratio of 8:2, which were further preprocessed by two transform ways for the following modeling procedure: scaled and centered (TrainS set and TestS set)/scaling; centering and principal component analysis (TrainP set and TestP set).

Comparison of Different ML Algorithms. A series of popular supervised ML algorithms for classification, including the generalized linear model (GLM), naive Bayes (NB), linear discriminate analysis (LDA), k-nearest neighbors (KNN), random forest (RF), classification and regression trees (CART), support vector machines with radial kernel (SVMr), C5.0 decision tree (C50), bagged CART (BAG), random ferns (RFer), rotation forest (RotF), and multilayer perceptron (MLP), were employed to preliminarily learn the hidden pattern in the TrainS set and TrainP set using the R package *Caret*[20] without tuning hyperparameters for each algorithm. Five times 10-fold cross-validation (CV) was set in this procedure. Then, all the performances of the models were previewed based on the values of accuracy, which were generated from the training process. The proper configurations of the training sets (TrainS set or TrainP set) and algorithms with satisfactory performance were selected for the subsequent model rebuilding and optimization process.

Model Optimization and Stacking. All the selected models were reconstructed, and the hyperparameters for each algorithm were also tuned based on the rules provided by the *Caret* package. To avoid overfitting issues, 10 times 10-fold CV strategy and a "smote" sampling method were also set in this process. During the optimization, different performance metrics including accuracy, sensitivity, specificity, and the values of area under the curves (AUC) of the receiver operating characteristic (ROC) were calculated and compared for the optimal combination of hyperparameters for each algorithm.

Then, because the highly correlated models could not contribute more to the performance of the stacking model, model correlation analysis was processed. After removing the highly correlated models, two stacking models were constructed based on all the rest of the optimal models using KNN and RF, respectively. And a 10-fold CV strategy was also set during the stacking process.

Model Performance on the Test Set. The single optimized model with the best performance and all staking models were selected and further validated on the TestS set. The accuracy, sensitivity (recall), specificity, precision, F1 score, and Matthews's correlation coefficient (MCC) of these models were calculated by the ratios of true positive (TP), false positive (FP), true negative (TN), and false negative (FN):

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

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{sensitivity/recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1 score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}$$

To further evaluate the model performance, ROC curves and AUC for each model were also plotted based on the sensitivity and specificity of each model using the *pROC* package.

Analysis of Feature Importance. In the current study, the importance of these features could reveal how the noncovalent interactions affect the binding affinity between the Nbs and their ligands. The importance of the features was evaluated by the Boruta algorithm[21] with a hyperparameter *ntree* set at 5000. In addition, the *P* values of these features were also evaluated using the *rfPermute* package.[22]

All figures in the current study were plotted by the *ggplot2* package.[23]

RESULTS AND DISCUSSION

Because of its compact structure, high affinity, and stability, Nb has emerged as a potential alternative that can replace traditional antibodies in basic research, disease diagnosis, and therapy.[24,25] Currently, several protein−ligand affinity prediction tools have been developed,[12,13] which may not be capable of accurately predicting the affinity of Nbs due to their shortened information in protein sequences and specific 3D structures. Besides, with the increase in the amount of data and attention given to Nbs, an affinity predicting tool specific to Nbs is urgently needed. In the current study, 12 kinds of ML models were constructed and compared based on the data of noncovalent interactions of the complexes of Nb−ligand. This is the first study on affinity prediction for Nbs using ML methods.

Data Set Processing. A total of 991 pairs of Nb−ligand complex were analyzed by ProtInter, and 72 descriptors related to noncovalent interactions between proteins were obtained, which were used as the features for model building. To avoid the overfitting problem due to the Nb data similarity during modeling and effectively show the useful patterns to ML algorithms,[26] we increased the variability of the Nb data by adding 444 other pairs of protein−protein interaction data,[27] which allow the data patterns to be better learned by the algorithms; in fact, the model constructed using these mix data performed better in predicting nanobody−ligand interactions than the model constructed using pure nanobody data (which is not shown in current study). Furthermore, the features with zero values higher than 75% and/or correlation coefficients higher than 70% were removed, and a 1435 × 36 data set was finally constructed with only the affinity class information as

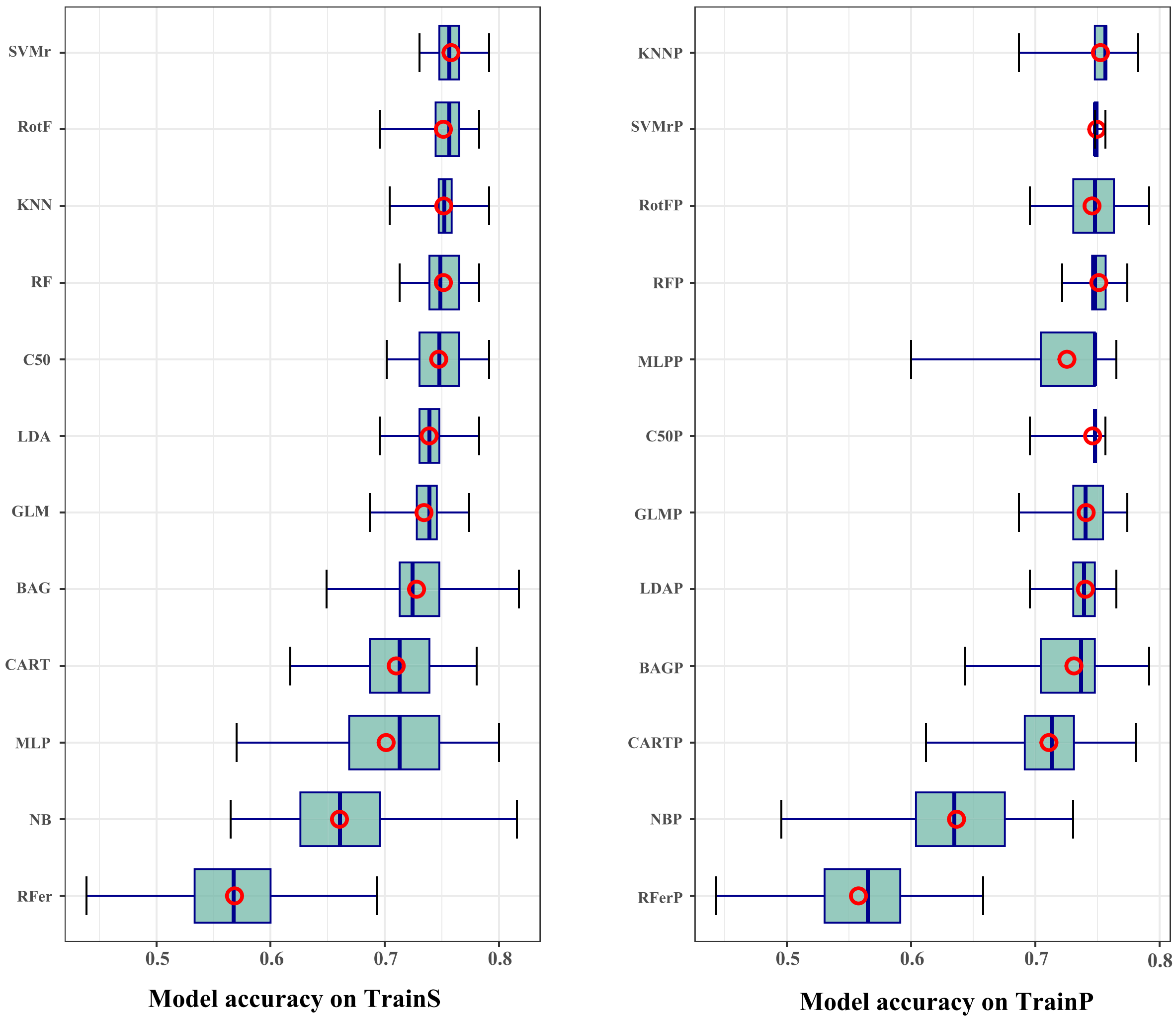


Figure 2. Preview of the performance of different models based on TrainS and TrainP sets.

well as 35 interaction features, which was further split into the training and test sets for the model building process. In detail, a total of 1149 cases were included in the training set, of which 288 were for ANb and 861 for NNb, while the rest of the data (286) was left for the test set, of which 71 were for ANb and 215 were for NNb. All of these two sets were further transformed using different data processing methods (TrainS set and TestS set; TrainP set and TestP set).

Performance Comparison of Different Algorithms. Twelve models were initially built by using the two different handed train sets, and the accuracy of these models was compared to gain an overview of the algorithm performance on learning data patterns of these training sets. As shown in Figure 2, besides NB and RFer, the accuracy of all of the other models was higher than 0.7012, and SVMr showed the highest accuracy at 0.7581. In the comprehensive comparison with other models, SVMr, KNN, RF, and C50 models showed a relatively good performance after analyzing the quartiles of the model's accuracy (Table S1, which includes all information on accuracy values of all the constructed models using different TrainS and TrainP sets). Besides, compared with other algorithms, ensemble algorithms showed better advantages.[28] Three of the top five models in terms of accuracy (>0.7473) were ensemble models (RF, RotF, and C50), although the top two were SVMr (accuracy = 0.7581) and KNN (accuracy = 0.7517). Although the use of the TrainP set resulted in a change in accuracy for some of the models, there was not much difference in performance compared to the models constructed using the TrainS set, so based on the TrainS set, all the models, except for NB and RFer, were further optimized for their hyperparameters.

Model Optimization and Stacking Model Construction. As shown in Table 1, three hyperparameter sets for each algorithm were evaluated, and the model performance metrics, including accuracy, sensitivity, specificity, and AUC, were calculated. Because the Caret package does not provide adjustable hyperparameters for GLM, LDA, and BAG, the optimizations have only been applied in the other models. Hyperparameter sets showing optimal model performance are labeled in bold style. Most models showed a higher specificity than sensitivity, which may result from the imbalance in the number of samples in ANb and NNb. Only three models (SVMr with sigma = 0.1763886, C = 252.0389; RF with mtry = 6; and C50 with trials = 93, model = tree, winnow = FALSE) were observed with accuracy higher than 0.7, among which only RF and C50 showed AUC values higher than 0.7, which mean that these two had a better model robustness than the others. The lower accuracy of the remaining optimized models compared to their performance in pretraining indicated an overfitting problem in the pretraining process, which could also

Table 1. Hyperparameter Optimization for the 10 Selected Candidate Models[a]

models	hyperparameter	accuracy	sensitivity	specificity	AUC
SVMr	sigma = 0.1663886 C = 0.889113	0.7200	0.2409	0.8803	0.6539
	sigma = 0.1743886 C = 48.889113	0.7466	0.1826	0.9352	0.6581
	sigma = 0.1763886 C = 252.0389	**0.7494**	**0.1868**	**0.9376**	**0.6583**
RotF	K = 4 L = 2	0.6191	0.5648	0.6373	0.6597
	K = 15 L = 17	0.6310	0.6134	0.6370	0.6836
	K = 7 L = 28	**0.6362**	**0.6327**	**0.6373**	**0.6904**
RF	mtry = 10	0.7147	0.3600	0.8334	0.7051
	mtry = 5	0.7079	0.3522	0.8270	0.7007
	mtry = 6	**0.7198**	**0.3584**	**0.8407**	**0.7062**
KNN	k = 8	0.5853	0.6644	0.5589	0.6544
	k = 12	0.5768	0.6973	0.5365	0.6581
	k = 4	**0.6004**	**0.6045**	**0.5991**	**0.6411**
C50	trials = 46 model = tree winnow = FALSE	0.7449	0.1909	0.9302	0.6995
	trials = 73 model = tree winnow = FALSE	0.7480	0.1878	0.9353	0.7018
	trials = 93 model = tree winnow = FALSE	**0.7490**	**0.1895**	**0.9361**	**0.7042**
CART	cp = 0.008968724	0.5910	0.6274	0.5789	0.6298
	cp = 0.009837963	0.5829	0.6427	0.5629	0.6336
	cp = 0.006944444	**0.6073**	**0.5794**	**0.6166**	**0.6239**
MLP	layer1 = 3 layer2 = 11 layer3 = 35	0.5865	0.5810	0.5884	0.6298
	layer1 = 24 layer2 = 33 layer3 = 45	0.6582	0.4054	0.7429	0.6304
	layer1 = 30 layer2 = 33 layer3 = 35	**0.6735**	**0.4060**	**0.7628**	**0.6311**
LDA		**0.5477**	**0.7140**	**0.4921**	**0.6384**
GLM		**0.5494**	**0.6942**	**0.5011**	**0.6412**
BAG		**0.6863**	**0.3743**	**0.7908**	**0.6822**

[a] Bold text represents the optimal hyperparameter configuration for each model.

be confirmed by their low AUC values (<0.7). Interestingly, RotFB with K = 7 and L = 28 showed a relatively balanced performance on metrics accuracy, sensitivity, specificity, and AUC, which may be due to the fact that the algorithm performs principal component analysis on feature subsets during the modeling process enabling the model to better capture the variations and structure of the data.[29] Furthermore, these results also imply that the model may have acceptable performance on unknown data. After optimization, the optimized models were named as follows: GLMB, LDAB, SVMrB, RFB, C50B, RotFB, etc.

The model correlation analysis indicated that the coefficients for the optimized models were generally lower than 0.6109 (Figure 3). The high coefficient (0.9001) between GLMB and LDAB means that removing one of them does not affect the performance of the subsequently stacked models. Then, with LDAB, the remaining ones were stacked by using the KNN and RF algorithm 10-fold CV strategy, which were named StackKNN and StackRf. StackRF showed accuracy at 1, while the metric for StackKNN was 0.9530.

Model Performance on the Test Set. To further analyze the generalizability of the models on the unknown data, four single models—SVMrB, RotFB, RFB, and C50B—and two stack models—StackKNN and StackRF—were evaluated on the test set. As shown in Table 2, all the models showed comparable performance on the overview accuracy, but as expected, the higher specificity compared to sensitivity indicated a better prediction on class NNb than ANb, which can be explained by the imbalance of ANb and NNb samples in the training set caused by uncertainty during data collection.[30,31] Compared with the other models, RotFB and RFB showed relatively lower accuracy (0.6853 and 0.7273), specificity (0.6884 and 0.8512), and precision (0.4174 and 0.4386) but higher sensitivity (0.6761 and 0.3521), F1 score (0.6415 and 0.6075), and MCC (0.3211 and 0.2198). These models could correctly predict the majority of ANb classes (sensitivity/recall). However, these accurately predicted ANbs constituted a relatively small proportion of the proteins predicted as ANbs (precision), which may lead to an increased workload during experimental validation. Besides, RotFB showed the highest F1 score, MCC, and AUC values, indicating a better model quality than the others. The opposite trend is found in SVMrB, C50B, and StackKNN with higher accuracy (0.7587, 0.7587, and 0.7517), specificity (0.9488, 0.9395, and 0.9256), and precision (0.5417, 0.5357, and 0.5000) but lower sensitivity (0.1831, 0.2113, and 0.2254), F1 score (0.5645, 0.5786, and 0.5796), and MCC (0.2056, 0.2192, and 0.2068). These models showed the ability to effectively reduce the number of experimentally validated candidates, especially excluding the number of potential NNb with high specificity, and the ratios between the numbers of ANbs (TP) and NNbs (FP) of the predication were higher than 1:1 (precision >0.5), but this comes at the cost of lowering the rate of correct prediction of truly valid ANbs (sensitivity/recall <0.2254). The relatively low F1 score, MCC, and AUC values were observed in these three models. Furthermore, all the ensemble models (RotFB, RFB, and C50B) showed better F1 score and MCC than the unensemble model (SVMrB), which were also proven by previous studies.[31] Although StackRF showed compromised performance, an overfitting problem was observed in the training process with a high accuracy at 1. Furthermore, the ROC curves and AUC values (Figure 4) also confirmed that the RotF was the most robust model with the highest AUC value at 0.7386. Besides, the AUC values of RFB and StackRF were 0.7135 and 0.7051, while all of the remaining modes showed AUC values lower than 0.7.

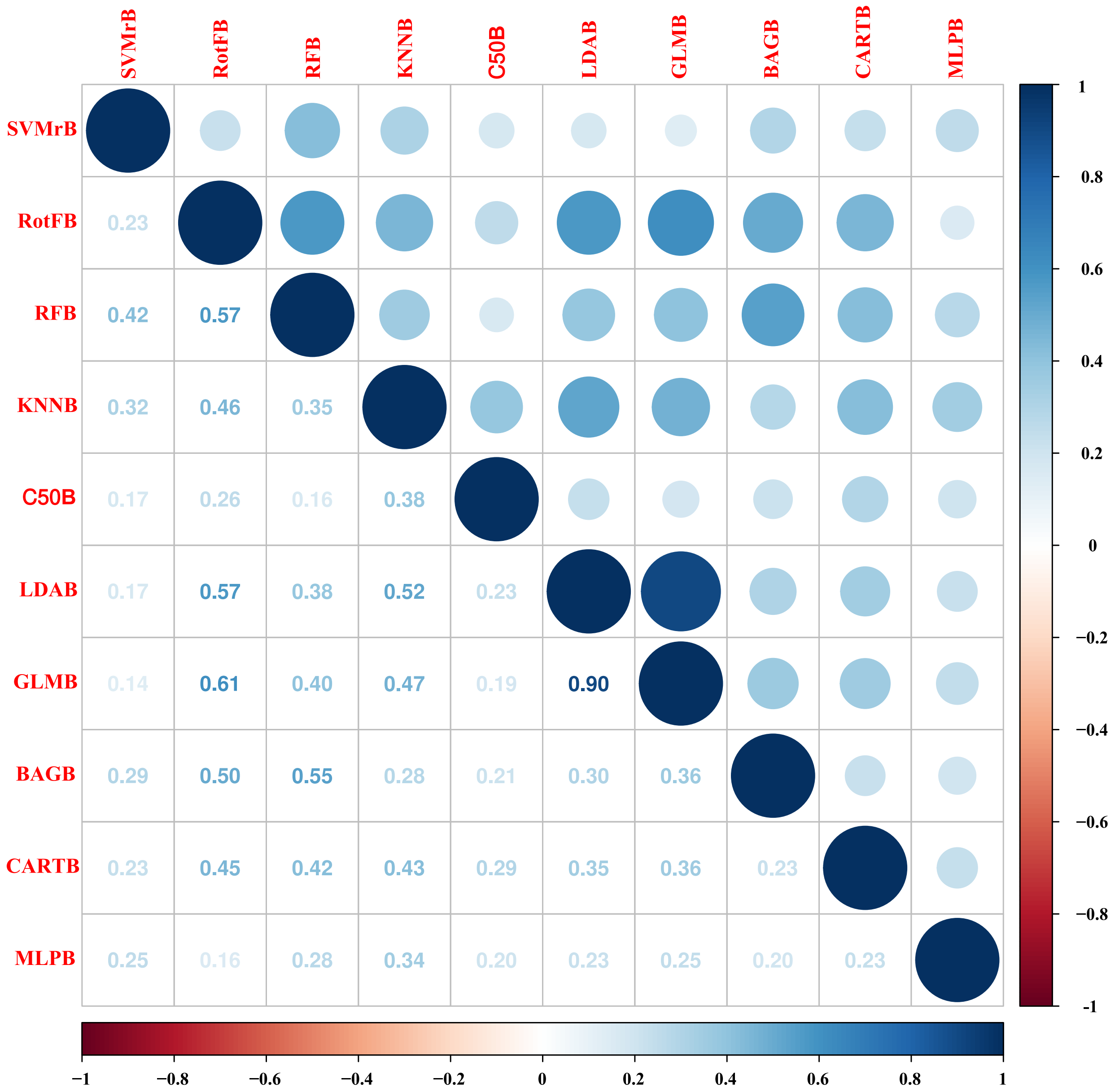


Figure 3. Correlation analysis of all of the optimized models.

Table 2. Performance Comparation of the Four Selected Optimized Models and Two Stacked Models on the TestS set

models	accuracy	sensitivity/recall	specificity	precision	F1 score	MCC
SVMrB	0.7587	0.1831	0.9488	0.5417	0.5645	0.2056
RotFB	0.6853	0.6761	0.6884	0.4174	0.6415	0.3211
RFB	0.7273	0.3521	0.8512	0.4386	0.6075	0.2198
C50B	0.7587	0.2113	0.9395	0.5357	0.5786	0.2192
StackKNN	0.7517	0.2254	0.9256	0.5000	0.5796	0.2068
StackRF	0.7448	0.2535	0.9070	0.4737	0.5863	0.2043

Importance Analysis of the Features. Noncovalent interactions are one kind of the key factors of biological processes, which are essential for maintaining protein advanced structure (tertiary and quaternary structure), as well as for mediating interactions in the protein−ligand.[32] As shown in Figure 5a, there were 27 features confirmed as important features in classifying whether Nbs were affinitive; the *P* values of 14 of them were lower than 0.05, which distributed into all types of noncovalent interactions but HI. Among the other eight features, there are five tentative features, one of which fell into the shadow value range along with three other features that were considered unimportant (see Table S1 for more detail on the feature importance).

Among these important features (Figure 5b), hydrogen bond descriptors accounted for 51.9% (14/27, n_HBMM = 6, n_HBMS = 5, n_HBSS = 3), while aromatic associated descriptors, including three ASI, three CPI, and two AAI, account for 29.6% (8/27). And all the rest of the features,

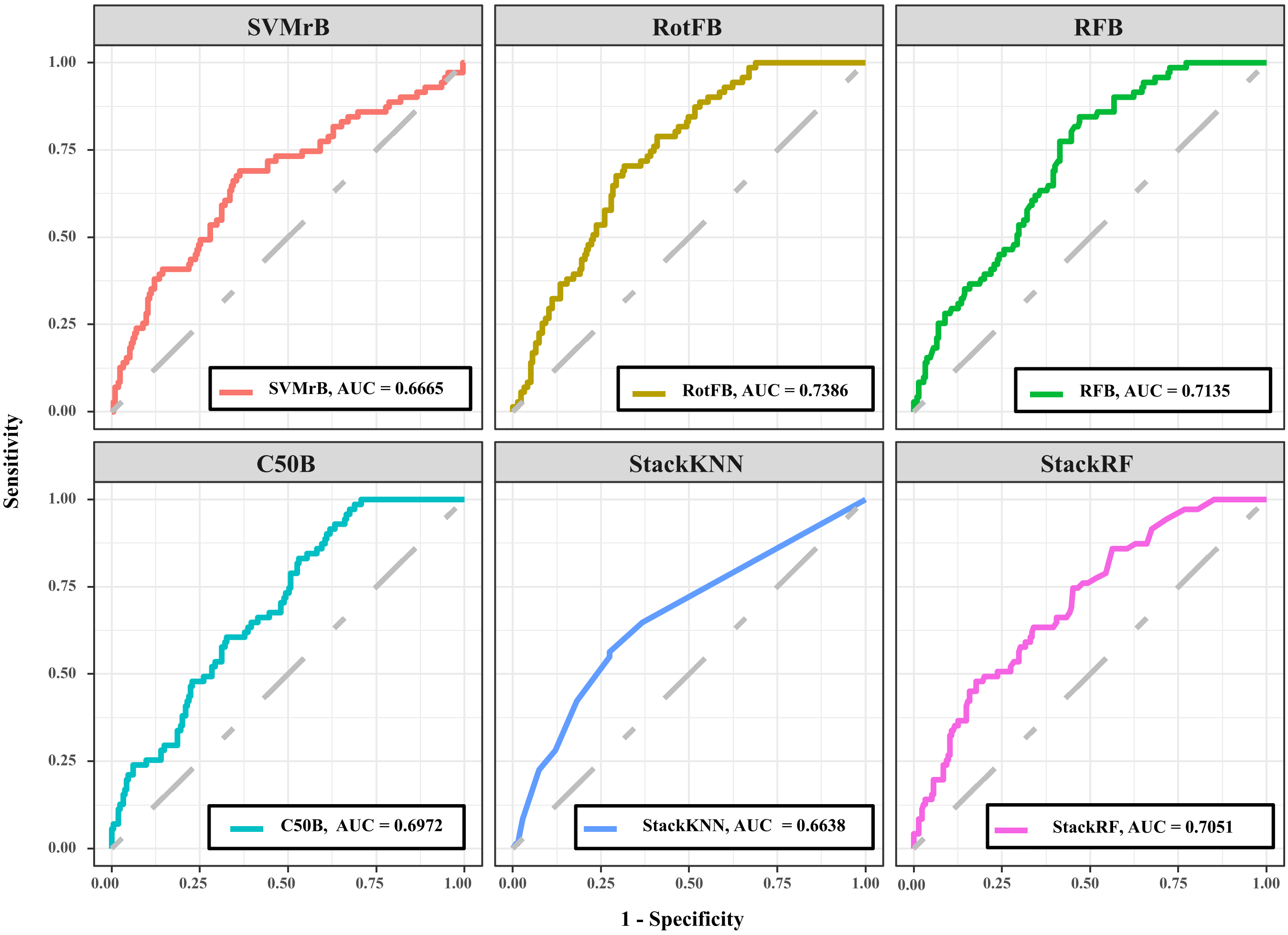


Figure 4. ROC plot with the AUC values for the selected optimized models and two stacking models.

relating to IoInt (n = 3), HI (n = 1), and DB (n = 1), in total accounted for 18.5%. Among these 27 important features, 14 of them related to the main/side chain effect of hydrogen bonding (HB) descriptors, implying that hydrogen bonds are key interactions determining Nb (protein)−ligand binding.[33,34] Furthermore, previous studies also indicated that hydrogen bonds were necessary and play a major role in highly specific Nb−ligand binding as well as a major role in maintaining nanobody stability.[35,36] With eight important features, the model performance contributed by aromatic-associated descriptors was next to the HB, which may be explained by the fact that HB interaction was more extensive than aromatic-associated interactions.[37] In addition, the feature significance analysis (Figure 5c, Table S2) also confirmed the importance of HB and aromatic-associated features, in which seven (including three HBMM, three HBMS, and one HBMSS) and five (two features each in ASI and CPI and one feature in AAI) important features were included with $P <$ 0.05. The less important features in the rest of the descriptors (IoInt, HI, and DB) played a relatively weak role in Nb (protein)−ligand interaction. Interestingly, as the most common interactions in protein−ligand complexes[37] and important factors in other affinity prediction tools, the only one feature in HI, with no significance, contributes less to the model performance in the current study, which may be due to the uniqueness of the structure of nanobodies and low hydrophobic amino acid composition in them,[38] and these reasons could also explain why new affinity prediction models were needed for nanobodies. Besides, Figure 5d showed that there were eight, seven, and six features related to count, std, and min, respectively, and the remaining features were almost evenly distributed to mean (n = 2), 0.25 (n = 1), 0.5 (n = 1), 0.75 (n = 1), and max (n = 1), which may imply that these features could better expose their underlying data patterns to algorithms.

■ CONCLUSIONS

As the amount of available data on the affinity of Nbs increases, it becomes possible to design ML-based models for predicting the interaction between Nbs and their ligands. In this study, a series of popular algorithms were used to build models for distinguishing ANbs and NNbs. The configurations of 10 models were selected for the following hyperparameter optimization. SVMrB, RotFB, RFB, and C50B were further selected for performance evaluation on the TestS set with two stacking models, StackKNN and StackRF, which were constructed based on nine optimal models. All six models showed different predictive advantages on different performance metrics. Compared to the rest, RotFB, RFB, and StackRF showed satisfactory performance for efficiently selecting a smaller number of affinitive ANbs. However, data shortages, imbalances, and similarities make the models constructed in this study still have great room for improvement, although the models developed in this study meet the needs for screening

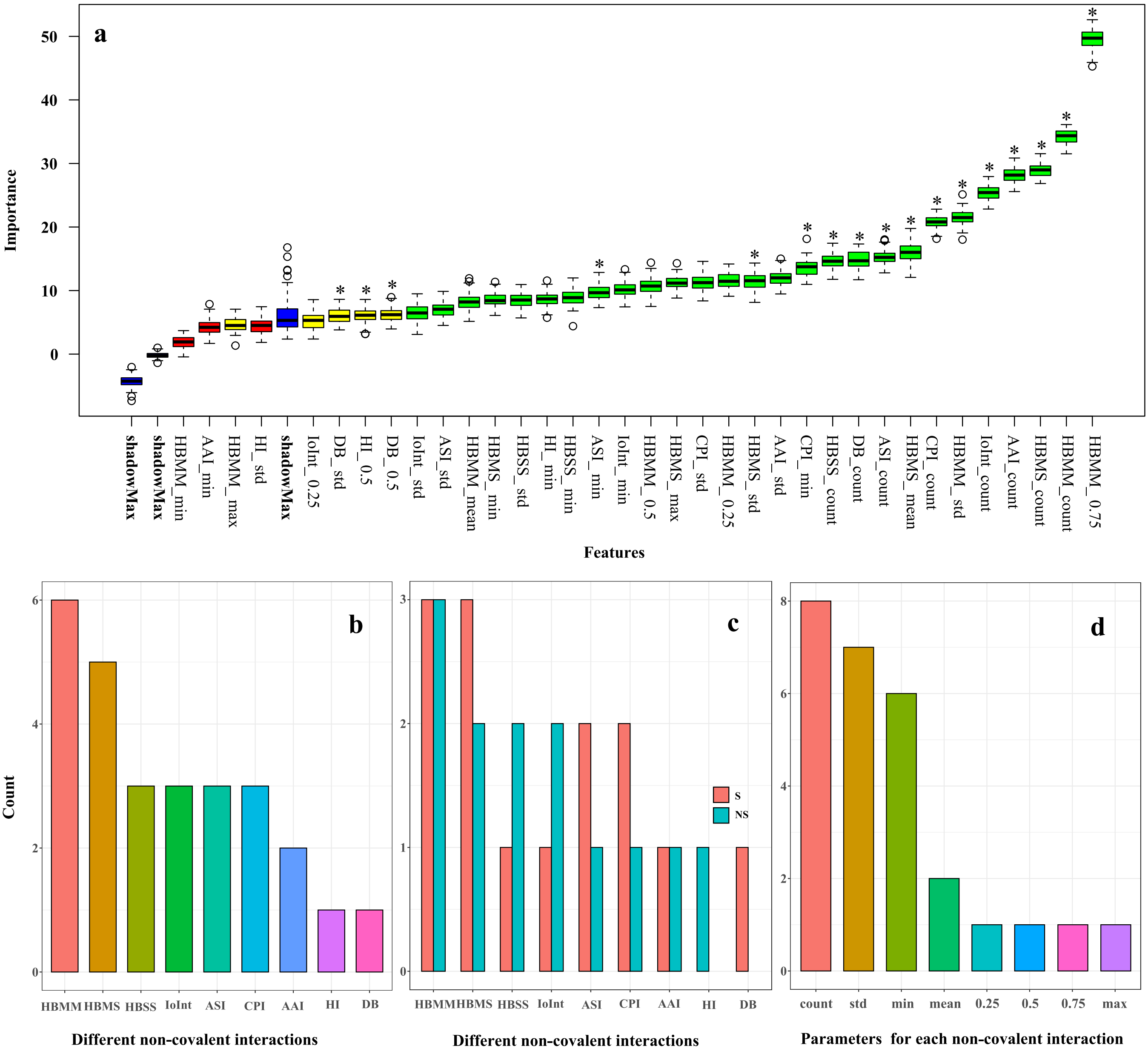


Figure 5. Feature engineering in the current study. (a) The importance of all features with P values (* represented $P < 0.05$.). (b) The number of important features in each noncovalent interaction descriptor. (c) The significance of important features in each noncovalent interaction descriptor. (d) The number of important features in each parameter calculated for the noncovalent interaction descriptors.

affinitive ANbs. In conclusion, the current study provided a tool for Nb/protein affinity evaluation for the first time, which could potentially accelerate research on novel Nb-related reagents and drugs.

■ ASSOCIATED CONTENT

Data Availability Statement

This model and the related data can be downloaded from https://github.com/greenGM/Nbaffinity.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c09718.

Accuracy of all models constructed in the current study by 12 different algorithms using TrainS and TrainP sets (Table S1) and details of the feature importance analysis (Table S2) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Fangyu Wang − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China; Longhu Laboratory, Zhengzhou 450002, China*; Email: sprinkle.w@126.com

Authors

Hua Feng − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China; Longhu Laboratory, Zhengzhou 450002, China*; orcid.org/0000-0001-5737-3172

Xuefeng Sun − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*

Ning Li − *College of Food Science and Technology, Henan Agricultural University, Zhengzhou 450002, China*

Qian Xu − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*

Qin Li − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*

Shenli Zhang − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*

Guangxu Xing − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*

Gaiping Zhang − *Institute for Animal Health, Key Laboratory of Animal Immunology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China; School of Advanced Agricultural Sciences, Peking University, Beijing 100871, China; Jiangsu Co-Innovation Center for the Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou University, Yangzhou 225009, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c09718

Author Contributions

[#]H.F. and X.S. contributed equally to the work.

Author Contributions

H.F. and F.W. conceived the concept and wrote the paper. X.S. and N.L. designed the methodology and performed experiments. Q.X., Q.L., and S.Z. collected all data and analysis results. G.X. and G.Z. advised on the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China (Grant 2023YFF1105001-02) and the Autonomous innovation project of Henan Academy of Agricultural Sciences (2024ZC097).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank associate professor Wang Sen and master's student Hao Yilin from Chongqing University of Technology for their advice on data processing and website construction.

■ REFERENCES

(1) Muyldermans, S. Nanobodies: natural single-domain antibodies. *Annu. Rev. Biochem.* **2013**, *82*, 775−797.

(2) Wang, J.; Kang, G.; Yuan, H.; Cao, X.; Huang, H.; de Marco, A. Research Progress and Applications of Multivalent, Multispecific and Modified Nanobodies for Disease Treatment. *Front Immunol* **2022**, *12*, No. 838082.

(3) Muyldermans, S. Applications of Nanobodies. *Annu. Rev. Anim Biosci* **2021**, *9*, 401−421.

(4) Hosseindokht, M.; Bakherad, H.; Zare, H. Nanobodies: a tool to open new horizons in diagnosis and treatment of prostate cancer. *Cancer Cell Int.* **2021**, *21* (1), 580.

(5) Salvador, J. P.; Vilaplana, L.; Marco, M. P. Nanobody: outstanding features for diagnostic and therapeutic applications. *Anal Bioanal Chem.* **2019**, *411* (9), 1703−1713.

(6) Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipcigan, F.; Chenthamarakshan, V.; Strobelt, H.; Dos Santos, C.; Chen, P. Y.; et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed Eng.* **2021**, *5* (6), 613−623.

(7) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785−2791.

(8) Zhou, P.; Jin, B.; Li, H.; Huang, S. Y. HPEPDOCK: a web server for blind peptide-protein docking based on a hierarchical algorithm. *Nucleic Acids Res.* **2018**, *46* (W1), W443−W450.

(9) Chen, R.; Li, L.; Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **2003**, *52* (1), 80−87.

(10) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29* (7), 845−854.

(11) Doerr, S.; Harvey, M. J.; Noe, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput* **2016**, *12* (4), 1845−1852.

(12) Wang, H. Prediction of protein-ligand binding affinity via deep learning models. *Brief Bioinform* **2024**, *25* (2), bbae081.

(13) Wang, Y.; Jiao, Q.; Wang, J.; Cai, X.; Zhao, W.; Cui, X. Prediction of protein-ligand binding affinity with deep learning. *Comput. Struct Biotechnol J.* **2023**, *21*, 5796−5806.

(14) Feng, H.; Wang, F.; Li, N.; Xu, Q.; Zheng, G.; Sun, X.; Hu, M.; Xing, G.; Zhang, G. A Random Forest Model for Peptide Classification Based on Virtual Docking Data. *Int. J. Mol. Sci.* **2023**, *24* (14), 11409.

(15) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers* **2021**, *25* (3), 1315−1360.

(16) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26* (9), 1169−1175.

(17) Romero-Molina, S.; Ruiz-Blanco, Y. B.; Mieres-Perez, J.; Harms, M.; Munch, J.; Ehrmann, M.; Sanchez-Garcia, E. PPI-Affinity: A Web Tool for the Prediction and Optimization of Protein-Peptide and Protein-Protein Binding Affinity. *J. Proteome Res.* **2022**, *21* (8), 1829−1841.

(18) Ahmed, A.; Mam, B.; Sowdhamini, R. DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. *Bioinform Biol. Insights* **2021**, *15*, No. 11779322211030364.

(19) Patel, L.; Shukla, T.; Huang, X.; Ussery, D. W.; Wang, S. Machine Learning Methods in Drug Discovery. *Molecules* **2020**, *25* (22), 5277.

(20) Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **2008**, *28* (05), 1−26.

(21) Kursa, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Software* **2010**, *36* (11), 1−13.

(22) Archer, E. EricArcher/rfPermute: version 2.5 (v2.5). *Zenodo* **2021**.

(23) Hadley, W. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York, 2016, https://ggplot2.tidyverse.org.

(24) Schumacher, D.; Helma, J.; Schneider, A. F. L.; Leonhardt, H.; Hackenberger, C. P. R. Nanobodies: Chemical Functionalization Strategies and Intracellular Applications. *Angew. Chem., Int. Ed. Engl.* **2018**, *57* (9), 2314−2333.

(25) Yang, E. Y.; Shah, K. Nanobodies: Next Generation of Cancer Diagnostics and Therapeutics. *Front Oncol* **2020**, *10*, 1182.

(26) Salam, M. A.; Taher, A.; Samy, M.; Mohamed, K. The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. *Int. J. Adv. Comput. Sci. App.* **2021**, *12* (4), 47.

(27) Gangwal, A.; Ansari, A.; Ahmad, I.; Azad, A. K.; Wan Sulaiman, W. M. A. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Comput. Biol. Med.* **2024**, *179*, No. 108734.

(28) Che, D.; Liu, Q.; Rasheed, K.; Tao, X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. *Adv. Exp. Med. Biol.* **2011**, *696*, 191−199.

(29) Rodriguez, J. J.; Kuncheva, L. I.; Alonso, C. J. Rotation Forest: A New Classifier Ensemble Method. *IEEE T Pattern Anal.* **2006**, *28* (10), 1619−1630.

(30) Poongavanam, V.; Kongsted, J. Virtual screening models for prediction of HIV-1 RT associated RNase H inhibition. *PLoS One* **2013**, *8* (9), No. e73478.

(31) Feng, H.; Wang, F.; Li, N.; Xu, Q.; Zheng, G.; Sun, X.; Hu, M.; Li, X.; Xing, G.; Zhang, G. Use of tree-based machine learning methods to screen affinitive peptides based on docking data. *Mol. Inform* **2023**, *42* (12), No. e202300143.

(32) Mati, I. K.; Cockroft, S. L. Molecular balances for quantifying non-covalent interactions. *Chem. Soc. Rev.* **2010**, *39* (11), 4195−4205.

(33) Schiebel, J.; Gaspari, R.; Wulsdorf, T.; Ngo, K.; Sohn, C.; Schrader, T. E.; Cavalli, A.; Ostermann, A.; Heine, A.; Klebe, G. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nat. Commun.* **2018**, *9* (1), 3559.

(34) Pantsar, T.; Poso, A. Binding Affinity via Docking: Fact and Fiction. *Molecules* **2018**, *23* (8), 1899.

(35) Raschka, S.; Wolf, A. J.; Bemister-Buffington, J.; Kuhn, L. A. Protein-ligand interfaces are polarized: discovery of a strong trend for intermolecular hydrogen bonds to favor donors on the protein side with implications for predicting and designing ligand complexes. *J. Comput. Aided Mol. Des* **2018**, *32* (4), 511−528.

(36) Li, J. D.; Wu, G. P.; Li, L. H.; Wang, L. T.; Liang, Y. F.; Fang, R. Y.; Zhang, Q. L.; Xie, L. L.; Shen, X.; Shen, Y. D.; et al. Structural Insights into the Stability and Recognition Mechanism of the Antiquinalphos Nanobody for the Detection of Quinalphos in Foods. *Anal. Chem.* **2023**, *95* (30), 11306−11315.

(37) Ferreira de Freitas, R.; Schapira, M. A systematic analysis of atomic protein-ligand interactions in the PDB. *Medchemcomm* **2017**, *8* (10), 1970−1981.

(38) Jovcevska, I.; Muyldermans, S. The Therapeutic Potential of Nanobodies. *BioDrugs* **2020**, *34* (1), 11−26.