RESEARCH ARTICLE

PASSerRank: Prediction of allosteric sites with learning to rank

Hao Tian[1] | Sian Xiao[1] | Xi Jiang[2] | Peng Tao[1]

[1]Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas, USA

[2]Department of Statistics, Southern Methodist University, Dallas, Texas, USA

Correspondence
Peng Tao, Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, TX, USA.
Email: ptao@smu.edu

Abstract

Allostery plays a crucial role in regulating protein activity, making it a highly sought-after target in drug development. One of the major challenges in allosteric drug research is the identification of allosteric sites. In recent years, many computational models have been developed for accurate allosteric site prediction. Most of these models focus on designing a general rule that can be applied to pockets of proteins from various families. In this study, we present a new approach using the concept of Learning to Rank (LTR). The LTR model ranks pockets based on their relevance to allosteric sites, that is, how well a pocket meets the characteristics of known allosteric sites. After the training and validation on two datasets, the Allosteric Database (ASD) and CASBench, the LTR model was able to rank an allosteric pocket in the top three positions for 83.6% and 80.5% of test proteins, respectively. The model outperforms other common machine learning models with higher F1 scores (0.662 in ASD and 0.608 in CASBench) and Matthews correlation coefficients (0.645 in ASD and 0.589 in CASBench). The trained model is available on the PASSer platform (https://passer.smu.edu) to aid in drug discovery research.

KEYWORDS
learning to rank, machine learning, protein allostery, web server

1 | INTRODUCTION

Allostery is a biological process where an effector molecule binds to an allosteric site that is distant to the active site of a protein. This binding results in conformational and dynamic changes that can regulate the protein's function, making it a key aspect of cellular signaling and is considered as the second secret of life.[1-4] Despite its importance, the allosteric mechanisms of most proteins remain elusive. A universal protein allosteric mechanism has yet to be formulated.[5,6]

Allostery offers several advantages in drug development. Compared to orthosteric site binding, allosteric site binding provides a controlled regulation of protein function that can either activate or inhibit the binding of ligands at orthosteric sites.[7] Additionally, allosteric modulators are reported to have fewer side effects with no additional pharmacological effects once allosteric sites are saturated.[8]

Furthermore, allosteric sites experience low evolutionary pressure, ensuring the safety of on-target drugs.[9,10] These benefits make allosteric drug development a promising field and offer substantial advantages over orthosteric drug development.

Identifying appropriate allosteric sites is a major challenge in allosteric drug development.[11,12] In recent years, numerous computational methods for allosteric site identification and prediction have been developed. With the help of machine learning (ML), Allosite[13] applies support vector machine (SVM) to learn the physical and chemical features of protein pockets. Another ML-based approach, the three-way random forest (RF) model developed by Chen et al.,[14] is capable of predicting allosteric, regular, or orthosteric sites. PASSer[15-17] is a recently developed method that combines extreme gradient boosting (XGBoost)[18] with a graph convolutional neural network[19] to learn physical and topological properties without any prior

information. In addition to ML, traditional methods such as normal mode analysis[20] and molecular dynamics[21] are widely used to investigate the communication between regulatory and functional sites, including SPACER[22] and PARS.[23]

Some efforts were spent to develop allostery databases for training and benchmark purposes. Allosteric Database (ASD) contains 1949 entries of protein-modulator complexes with annotated allosteric residues.[24] ASBench is a smaller benchmark dataset generated based on the ASD.[25] Both ASD and ASBench are used in the current study for the training and benchmarking purposes. The CASBench benchmark set contains 91 proteins which structures are annotated based on the experimental information from ASD, Catalytic Site Atlas, and Protein Data Bank.[26,27] ASD and CASBench can be accessed at http://mdl.shsmu.edu.cn/ASD and https://biokinet.belozersky.msu.ru/casbench, respectively. These datasets play a crucial role in training and evaluating allosteric site prediction models.

Most previous research on prediction models focused on developing universal models for allosteric site prediction. These models intend to make "absolute" predictions (either as labels or probabilities) for all pockets detected in different types of proteins, which is a challenging and time-consuming task. Learning to Rank (LTR), as an emerging area, was first applied in information retrieval[28] and has been used in many bioinformatics studies, ranging from drug-target interaction prediction[29] to compound virtual screening.[30] Unlike "absolute" predictions, LTR models provide "relative" predictions by ranking objects from the most to the least relevant to a target, making it a more achievable and reasonable approach for allosteric site prediction.

In this study, we present the state-of-the-art ML model on allosteric site prediction with LTR. The LTR model is implemented using LambdaMART. LambdaMART combines gradient boosting decision tree (GBDT) with the loss function derived from LambdaRank, a LTR algorithm. Compared with other ML models such as XGBoost, SVM, and RF, LambdaMART achieved the highest F1 score and Matthews correlation coefficient (MCC). Moreover, this model has a better ability to rank actual allosteric sites at top positions. The trained LambdaMART model is freely available at PASSer (https://passer.smu.edu) to facilitate related research.

2 | METHODOLOGY

2.1 | Allosteric protein databases

Two databases were used to train and validate different ML models, including the ASD and CASBench.

In the latest version of ASD, there are 1949 entries of protein-modulator complexes. To ensure data quality, a cleaning process is applied to the protein-modulator complexes based on standards proposed in the Allosite study.[13] Three standards, including high-resolution protein structures with a resolution smaller than 3 Å, the presence of a complete structure in the allosteric site, and a low sequence identity threshold of 30%, was applied to select high-quality and sequence-diverse proteins in the overall training set. If two or

more proteins have high sequence identity, the one with the shortest modulator-pocket distance is retained to ensure the finest labeling. The modulator-pocket distance calculation is described below in Section 2.2. A total of 207 proteins were selected in the overall training set and were randomly split into a training set (80%) and a test set (20%). To facilitate the cleaning process, a data processing pipeline script has been created and made available as open source on GitHub (https://github.com/smu-tao-group/PASSerRank).

The CASBench dataset was used as an external test set. The CASBench benchmark set comprises proteins annotated with allosteric sites, but only those entries that include both allosteric ligands and sites were included. Additionally, proteins that were already present in the ASD dataset were removed to ensure the validity of the benchmark set. Through cleaning, 71 out of 91 proteins were selected for validation.

2.2 | Pocket descriptors and labeling

FPocket is an open-source software for protein pocket detection.[31] In this work, FPocket was applied on each protein to detect protein pockets. On average, 21 pockets were detected in each protein, with a total of 4413 pockets in 207 proteins. For each detected pocket, 19 physical and chemical features are calculated, ranging from score, volume, solvent accessible surface area, to hydrophobicity. A complete list of feature names is shown in Figure 2. Specifically, the FPocket score is calculated using a scoring function which includes five features (number of alpha spheres, mean local hydrophobic density, proportion of apolar alpha sphere, polarity score, and alpha sphere density). These features are normalized to 0–1 scale within each protein and the function is fitted using FPocket training dataset.[31] The FPocket score has little overlap with the other 18 features used in the present study.

To label each pocket as an allosteric or nonallosteric site, we have automated the process by assigning the closest pockets to the modulator as the allosteric site. The center of mass is first calculated for all pockets and the modulator, and then the pairwise distances between the pockets and the modulator are computed. The pocket with the shortest distance is labeled as positive (allosteric site), while all other pockets are labeled as negative (nonallosteric site). However, if the closest distance is greater than 10 Å, this entry is removed from the dataset, as such a large distance may indicate inaccurate pocket detection and negatively impact model performance. It should be noted that this labeling strategy may lead to false negative cases where the pockets labeled as nonallosteric are allosteric in other protein-modulator settings. However, the potential false negative cases are considered unlikely and less problematic because the main purpose of the prediction method is identifying allosteric sites as true positives.

2.3 | Learning to rank

Prior researches on allosteric site prediction focus on developing a universal model that can accurately predict allosteric sites in all proteins. However, in practice, it is more important to identify the most

promising pockets within each individual protein.[32] Therefore, a ML model that is capable of ranking pockets in order of their likelihood to be allosteric sites is more desirable and attainable than a binary classification model that provides absolute predictions for all pockets.

In this study, we implemented the LTR algorithm using GBDT and the LambdaMART method. GBDT is a popular ML approach that iteratively learns from decision trees and ensembles of their predictions. Here, we use LightGBM,[33] one of the two popular implementations of GBDT, over XGBoost.[18] LambdaMART is an LTR method that trains GBDT with the lambdarank loss function. The lambdarank loss function optimizes the value of the normalized discounted cumulative gain (NDCG) for the top K cases, and is calculated using discounted cumulative gain (DCG) and ideal discounted cumulative gain (IDCG) as:

$$DCG@K = \sum_{i=1}^{K} \frac{2^{G_i} - 1}{\log_2(i+1)}, \tag{1}$$

$$IDCG@K = \sum_{i=1}^{K} \frac{2^{|G|_i} - 1}{\log_2(i+1)}, \tag{2}$$

$$NDCG@K = \frac{DCG@K}{IDCG@K}, \tag{3}$$

where G_i is the gain (graded relevance value) at position i and $|G|$ is the ideal ranking.

The LGBMRanker module in the LightGBM package (v3.3.4) was used to implement the LambdaMART algorithm with GBDT as boosting type and lambdarank as the objective function.

2.4 | ML models

In addition to the LTR model, other commonly used ML models in allosteric site prediction were considered for comparison. XGBoost and RF are tree-based models. As previously stated, XGBoost is an implementation of the GBDT model that could also be used to train the LTR model. The RF model employs a bagging approach, training several independent decision trees in parallel. The prediction of RF is obtained through the weighted average of the outputs of all decision trees. The SVM classifier, on the other hand, learns a high-dimensional hyperplane that separates data points based on their labels. The XGBoost algorithm was implemented using the XGBoost package (version 1.7.3), and the RF and SVM classifiers were implemented using the Scikit-learn package (version 1.2.0).[34]

SHapley Additive exPlanations (SHAP) value is a method to increase model interpretability by quantifying feature importance. It has been implemented recently to explain tree-based models.[35] In this study, the SHAP values of 19 features from FPocket were calculated and compared. The method is implemented in the SHAP package (v0.41.0).[36]

2.5 | Performance criteria

Several metrics were calculated to compare and evaluate different ML models. Precision, recall, and specificity are good indicators for binary classification. The F1 score is a weighted measure of precision and recall. Moreover, it is reported that the MCC is a more suitable indicator than the F1 score and accuracy in binary classification evaluation.[37]

$$Precision = TP/(TP + FP). \tag{4}$$

$$Recall = TP/(TP + FN). \tag{5}$$

$$Specificity = TN/(TN + FP). \tag{6}$$

$$F1\ score = 2 * Precision * Recall/(Precision + Recall). \tag{7}$$

$$Accuracy = (TP + TN)/(TP + FP + FN + TN). \tag{8}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}. \tag{9}$$

The percentage of actual allosteric sites that are ranked among the top 1, 2, and 3 positions is calculated. This metric has been commonly used in evaluating various allosteric site prediction models.[23,38] The actual allosteric sites are compared with the predicted top three most probable pockets in each protein, and the percentage is calculated and accumulated for each position.

3 | RESULTS

FPocket was used to identify potential pockets in each protein. To ensure that the allosteric site and modulator are in contact, a distance threshold was imposed on the closest pocket. The effect of different distance thresholds on the percentage of proteins included in the training set is shown in Figure 1A, with a final distance threshold of 10 Å chosen to avoid the inclusion of incorrectly labeled pockets. Consequently, 91.1% of proteins from the ASD were included in the training set. To avoid over-representation of highly similar proteins, the pairwise sequence similarity was calculated at protein level between each newly selected protein and all previously selected proteins. If the similarity was higher than a specified threshold, the protein structure was discarded. The effect of different sequence identity thresholds is shown in Figure 1B, with a final threshold of 30% chosen. After these steps, 207 proteins were included in the overall training set.

We randomly selected 80% of these proteins as the training set and used the remaining 20% as the testing set. A total of four ML models, including LambdaMART, XGBoost, RF, and SVM, were trained through fivefold grid search with cross validation. The grid search takes an exhaustive search strategy over all combinations of prespecified parameter values. All models were trained on a

(A) (B)
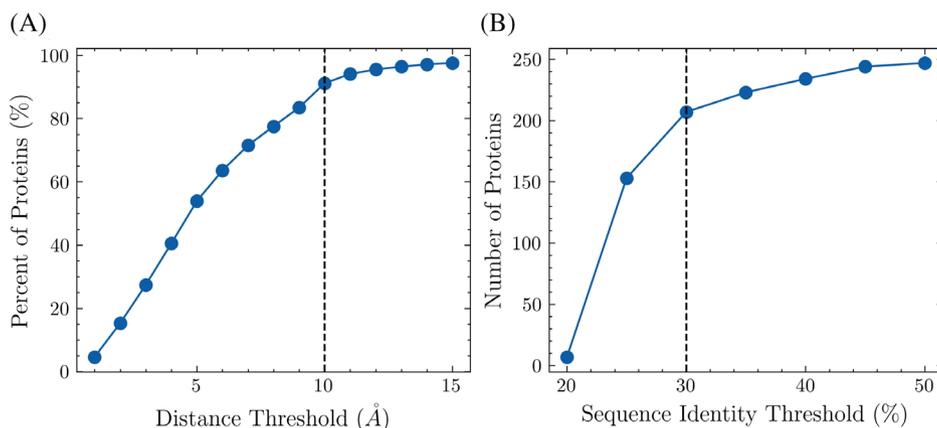
FIGURE 1 The number of proteins included in the training set, along with different distance and sequence identity thresholds. (A) The minimum distances between the center of masses from pockets to the modulator in each protein were calculated. A protein-modulator complex is discarded if the minimum distance is higher than the threshold. With the threshold being set as 10 Å, 91.1% of proteins were included. (B) To ensure uniqueness, proteins with high sequence identity were removed. The threshold was set as 30%. A total of 207 proteins are included in the training set.

high-performance-computing platform with a 60GB memory V100 graphical processing unit. Different combinations of parameter values were tested to optimize the performance of the prediction models. A full list of parameter names and values is available at https://github.com/smu-tao-group/PASSerRank/blob/main/src/utils/params.json. The set of parameters which had the best performance on the training set was determined. For comparison, the performance of FPocket is reported, in which the pocket with the highest score was treated as the positive (allosteric) prediction and others as negative (nonallosteric) predictions based on FPocket results. Similarly for the LambdaMART predictions, the pocket with the highest prediction score in each protein was labeled as positive. This explains that precision and recall metrics have the same number in LambdaMART and FPocket models, respectively, as there is only one positive prediction. If this positive prediction is wrong, we have a false positive, and there will also be a false negative, leading to the same number of FP and FN and thus the same value of precision and recall defined in Equation 4 and 5.

All models were evaluated using the testing set of ASD. The results are listed in Table 1. The percentage of true allosteric sites that appeared in the predicted top 1, 2, and 3 positions was calculated and abbreviated as Top 1, 2, and 3, respectively. The performance of four ML models was compared with FPocket. Both LambdaMART and XGBoost exhibited better performance than FPocket under all or most metrics. RF and SVM were comparable to FPocket with higher F1 scores, MCC, and Top 3 percentage. LambdaMART achieved the best performance in eight out of nine metrics among all models.

These models were further evaluated using the CASBench dataset. The CASBench training dataset was prepared with the same procedures as the ASD training data. In addition, the proteins included in the ASD training data were excluded in the CASBench set to ensure the evaluation validity. The same metrics were

calculated, and the results are listed in Table 2. Compared with the numbers reported in Table 1, the performance of all models was decreased but within an acceptable range. Overall, LambdaMART is superior to FPocket and leads in 7 out of 9 metrics. This demonstrates the ability of LambdaMART to rank protein pockets in terms of the relevance to allostery, which leads to a high F1 score, MCC, and Top 3 percentage.

The feature importance of the LambdaMART model was analyzed using SHAP values. As shown in Figure 2, the SHAP value distributions and mean SHAP values were displayed in descending order. Figure 2 shows the distribution and mean SHAP values of the features in descending order. The results indicate that the FPocket score was the most important feature and significantly outperformed all other features. This highlights the effectiveness of the FPocket score in differentiating between allosteric and nonallosteric sites. Other features that were found to be important include the volume score, flexibility, charge score, and total solvent-accessible surface area (SASA). As seen from the SHAP value distribution, allosteric sites (represented in red) tend to have high FPocket scores, high volume scores, high charge scores, but low flexibility and low total SASA.

The trained LambdaMART model has been made accessible to the public through the PASSer platform (https://passer.smu.edu). Users can access the model either through the webpage or through the command line interface using the PASSer API (https://passer.smu.edu/apis/). To demonstrate the efficacy of the model, two examples that were not part of the training and validation sets are presented in Figure 3. These examples show the predicted allosteric sites of the light-oxygen-voltage domains of *Phaeodactylum tricornutum* Aureochrome 1a[39] and *Avena Sativa* phototropin 1[40] obtained using the LambdaMART model. The top 3 pockets are highlighted in red, orange, and yellow, respectivley, and with the corresponding predicted relevance scores. The top 1 predicted pockets illustrated in red are actual allosteric sites in both cases.

TABLE 1 Performance comparison of machine learning models on the Allosteric Database dataset.

Metric	LambdaMART	XGBoost	RF	SVM	FPocket
Precision	**0.662** ↑	0.586 ↑	0.528 ↓	0.444 ↓	0.556
Accuracy	**0.968** ↑	0.961 ↑	0.956 ↓	0.944 ↓	0.958
Recall	0.662 ↑	0.609 ↑	0.677 ↑	**0.758** ↑	0.556
Specificity	**0.983** ↑	0.979 ↑	0.970 ↓	0.953 ↓	0.978
F1 score	**0.662** ↑	0.596 ↑	0.593 ↑	0.559 ↑	0.556
Matthews correlation coefficient	**0.645** ↑	0.577 ↑	0.575 ↑	0.554 ↑	0.536
Top 1	**59.5%** ↑	56.6% ↑	58.0% ↑	57.5% ↑	55.6%
Top 2	**73.9%** ↑	69.6% ↓	71.0% ↓	69.6% ↓	71.5%
Top 3	**83.6%** ↑	80.7% ↑	79.7% ↑	78.3% ↑	76.8%

Abbreviations: RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

TABLE 2 Performance comparison of machine learning models on the CASBench dataset.

| Metric	LambdaMART	XGBoost	RF	SVM	FPocket
Precision	**0.608** ↑	0.504 ↓	0.431 ↓	0.395 ↓	0.550
Accuracy	**0.963** ↑	0.953 ↓	0.941 ↓	0.932 ↓	0.956
Recall	0.608 ↑	0.657 ↑	0.767 ↑	**0.803** ↑	0.550
Specificity	**0.980** ↑	0.968 ↓	0.950 ↓	0.939 ↓	0.977
F1 score	**0.608** ↑	0.569 ↑	0.551 ↑	0.529 ↓	0.550
MCC	**0.589** ↑	0.551 ↑	0.548 ↑	0.534 ↑	0.527
Top 1	56.3% ↑	52.5% ↓	44.1% ↓	**57.3%** ↑	55.5%
Top 2	**73.7%** ↑	70.0% ↓	68.4% ↓	73.2% ↑	71.4%
Top 3	**80.5%** ↑	77.0% ↑	76.6% ↓	76.0% ↓	76.7%

Abbreviations: MCC, Matthews correlation coefficient; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

4 | DISCUSSION

The collection and cleaning of training data is a crucial step for the development of a high-performing ML model. The study by Huang et al.[13] applied three rules to select protein structures from the ASD dataset and curated a training set of 90 proteins. However, there is no available script to automate this process, which can result in an unfair model comparison. To address this issue, an open-source script to prepare ML-ready dataset is provided. This pipeline offers a simple and customizable benchmark preparation for evaluating various ML models. It should be noted that in the current design, proteins with multiple modulators in the same chain are discarded, as this can result in inconsistent ratios of allosteric and nonallosteric sites in each protein. Further refinement of the data cleaning process can lead to higher-quality training data. It could be feasible to develop pretrained models based on the nonallosteric data to be used to refine the allosteric site prediction models. However, the amount of nonallosteric protein data significantly exceeds the amount of allosteric protein data, leading to the data imbalance issue in the model development for allosteric sites prediction. In the current implementation, FPocket could detect over 10 pockets in one protein and typically only one is allosteric. Therefore, a pretrained model based on the nonallosteric

data was not considered in this study. During dynamical processes of protein with significant conformational changes, hidden allosteric sites might appear with new conformations. The current method could be applied on selected conformations from dynamical process of proteins to reveal hidden allosteric sites associated with conformational driven allostery. There are other approaches to address this issue, including a recently reported prediction model, TopoAlloSite, for the prediction of cryptic allosteric sites.[41]

Efforts have been invested in developing a universal model for allosteric site prediction by learning pockets from different proteins without considering the protein itself, such that all detected pockets from proteins in the training set are gathered and shuffled in a pool for training purposes. This approach, however, poses a challenge in model design and requires a model to learn a general rule that applies to all proteins of various families. Additionally, this training process is not reflective of real-world applications, where all pockets in a target protein need to be compared to determine the most probable ones. In light of these challenges, we offer a new perspective to rank pockets in each protein. The model focuses on the protein level and learns a ranking pattern among pockets. The proposed LambdaMART model outperforms other popular ML models such as XGBoost and SVM, with high F1 score and MCC, and is capable of ranking actual
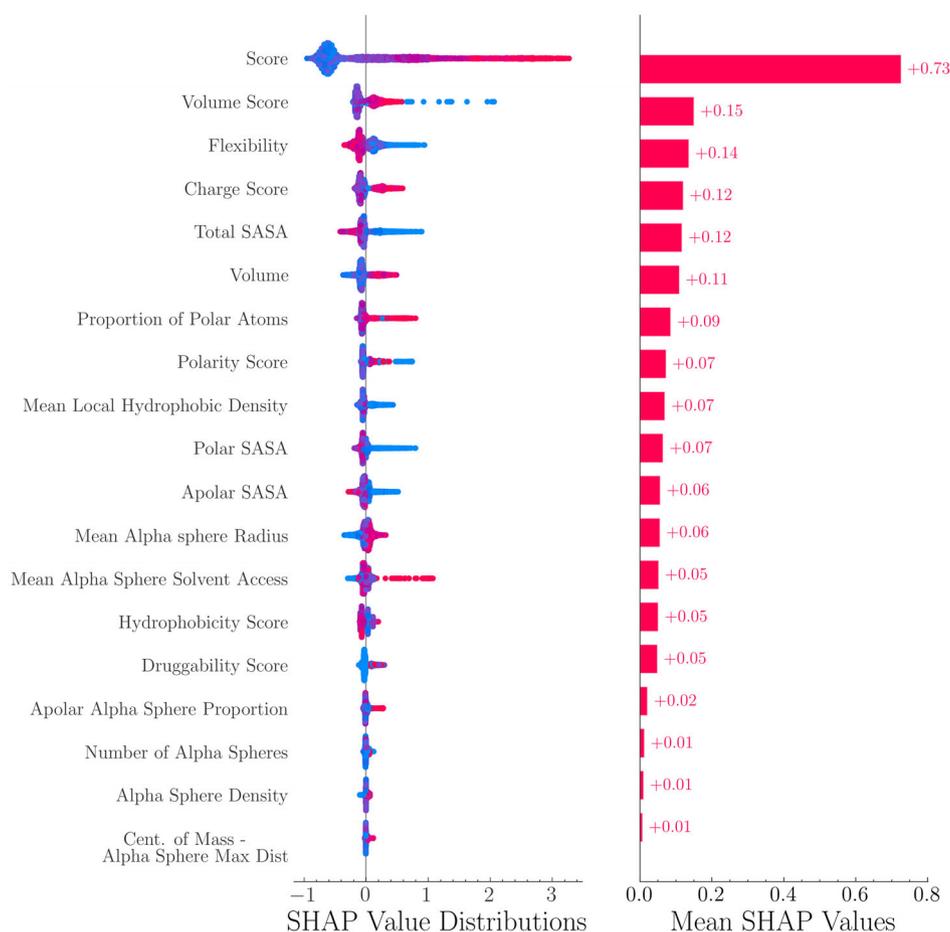
FIGURE 2 SHapley Additive exPlanations (SHAP) value distributions and mean values of 19 features. These features are calculated from FPocket. Red and black colors indicate positive and negative samples, respectively. FPocket score was identified as the most important feature.
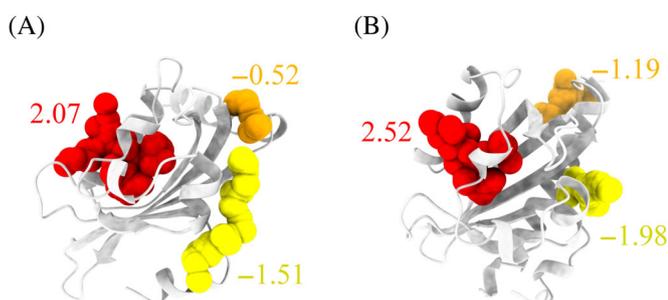
FIGURE 3 Predictions of the light-oxygen-voltage domains from (A) *Phaeodactylum tricornutum* Aureochrome 1a and (B) *Avena Sativa* phototropin 1. In each protein system, top 3 pockets with the highest rank scores are highlighted in red (first), orange (second), and yellow (third) colors, respectively. Corresponding rank scores are displayed next to each pocket in the same color. The top 1 pocket from both examples illustrated in red are the actual allosteric sites.

allosteric sites at the top positions. This demonstrates that it is more effective to learn the relative differences among pockets rather than a universal law applicable to all proteins.

In the context of allosteric site prediction, explainable ML is important as it helps researchers understand how a model arrives at its predictions. This information can be useful in drug design, as it can provide insights into the influencing factors that whether a

pocket is likely to be an allosteric site. Tree-based models, such as RF and gradient boosting decision tree, have good explainability as they can use metrics like Gini impurity to determine feature importance. SHAP values, a method from cooperative game theory, can also be used to quantify the contribution of each feature to the predictions made by a ML model. In this study, the SHAP values were used to indicate that the FPocket score was the most crucial feature, which aligns with the good performance of FPocket as a benchmark model.[16] The SHAP values also revealed that the model tends to predict pockets with high charge, volume, and low flexibility as allosteric sites, which can benefit the development of allosteric drugs.

5 | CONCLUSIONS

The prediction of allosteric sites is crucial to the development of allosteric drugs. While many efforts have been dedicated to constructing a universal model for such prediction, this study presents a novel approach by employing a ranking model through the learning to rank concept. The proposed model outperforms other ML models based on various performance metrics, including a high rate of ranking true allosteric sites at top positions. Furthermore, a customizable pipeline is provided for the preparation of high-quality proteins for training

purposes. The trained model is deployed on the PASSer platform (https://passer.smu.edu) and is readily available for public usage.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in PASSerRank Github repository at https://github.com/smu-tao-group/PASSerRank.

ORCID

Hao Tian https://orcid.org/0000-0002-0186-9811
Peng Tao https://orcid.org/0000-0002-2488-0239

REFERENCES

[1] J. Liu, R. Nussinov, *PLoS Comput. Biol.* **2016**, *12*, e1004966.
[2] S. J. Wodak, E. Paci, N. V. Dokholyan, I. N. Berezovsky, A. Horovitz, J. Li, V. J. Hilser, I. Bahar, J. Karanicolas, G. Stock, P. Hamm, R. H. Stote, J. Eberhardt, Y. Chebaro, A. Dejaegere, M. Cecchini, J. P. Changeux, P. G. Bolhuis, J. Vreede, P. Faccioli, S. Orioli, R. Ravasio, L. Yan, C. Brito, M. Wyart, P. Gkeka, I. Rivalta, G. Palermo, J. A. McCammon, J. Panecka-Hofman, R. C. Wade, A. di Pizio, M. Y. Niv, R. Nussinov, C. J. Tsai, H. Jang, D. Padhorny, D. Kozakov, T. McLeish, *Structure* **2019**, *27*, 566.
[3] K. Krishnan, H. Tian, P. Tao, G. M. Verkhivker, *J Chem Phys* **2022**, *157*, 245101.
[4] A. W. Fenton, *Trends Biochem. Sci.* **2008**, *33*, 420.
[5] R. Nussinov, C.-J. Tsai, *Cell* **2013**, *153*, 293.
[6] S. Xiao, G. M. Verkhivker, P. Tao, *Trends Biochem. Sci.* **2022a**, *48*, 375.
[7] A. Peracchi, A. Mozzarelli, *Biochim. Biophys. Acta* **2011**, *1814*, 922.
[8] N. Wu, L. Strömich, S. N. Yaliraki, *Patterns* **2022**, *3*, 100408.
[9] A. Christopoulos, L. May, V. A. Avlani, P. M. Sexton, *Biochem. Soc. Trans.* **2004**, *32*, 873.
[10] F. De Smet, A. Christopoulos, P. Carmeliet, *Nat. Biotechnol.* **2014**, *32*, 1113.
[11] S. Lu, W. Huang, J. Zhang, *Drug Discovery Today* **2014**, *19*, 1595.
[12] S. Lu, Q. Shen, J. Zhang, *Acc. Chem. Res.* **2019**, *52*, 492.
[13] W. Huang, S. Lu, Z. Huang, X. Liu, L. Mou, Y. Luo, Y. Zhao, Y. Liu, Z. Chen, T. Hou, J. Zhang, *Bioinformatics* **2013**, *29*, 2357.
[14] A. S.-Y. Chen, N. J. Westwood, P. Brear, G. W. Rogers, L. Mavridis, J. B. Mitchell, *Mol. Inf.* **2016**, *35*, 125.
[15] H. Tian, X. Jiang, P. Tao, *Mach. Learn. Sci. Technol.* **2021**, *2*, 035015.
[16] S. Xiao, H. Tian, P. Tao, *Front. Mol. Biosci.* **2022b**, *9*, 879251.
[17] H. Tian, S. Xiao, X. Jiang, P. Tao, *Nucl. Acids Res.* **2023**, *51*, W427. https://doi.org/10.1093/nar/gkad303
[18] T. Chen, C. Guestrin. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining. **2016**, 785.
[19] T. N. Kipf, M. Welling. *arXiv preprint arXiv:1609.02907*. **2016**.
[20] A. Panjkovich, X. Daura, *BMC Bioinform.* **2012**, *13*, 1.
[21] E. Laine, C. Goncalves, J. C. Karst, A. Lesnard, S. Rault, W.-J. Tang, T. E. Malliavin, D. Ladant, A. Blondel, *Proc. Natl. Acad. Sci.* **2010**, *107*, 11277.
[22] A. Goncearenco, S. Mitternacht, T. Yong, B. Eisenhaber, F. Eisenhaber, I. N. Berezovsky, *Nucleic Acids Res.* **2013**, *41*, W266.
[23] A. Panjkovich, X. Daura, *Bioinformatics* **2014**, *30*, 1314.
[24] Z. Huang, L. Zhu, Y. Cao, G. Wu, X. Liu, Y. Chen, Q. Wang, T. Shi, Y. Zhao, Y. Wang, W. Li, Y. Li, H. Chen, G. Chen, J. Zhang, *Nucleic Acids Res.* **2011**, *39*, D663.
[25] W. Huang, G. Wang, Q. Shen, X. Liu, S. Lu, L. Geng, Z. Huang, J. Zhang, *Bioinformatics* **2015**, *31*, 2598.
[26] S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, S. Velankar, *Protein Crystal.* **2017**, *1607*, 627.
[27] A. Zlobin, D. Suplatov, K. Kopylov, V. Švedas, *Acta Naturae* **2019**, *11*, 74.
[28] A. Trotman, *Inform. Retriev.* **2005**, *8*, 359.
[29] X. Ru, X. Ye, T. Sakurai, Q. Zou, *Bioinformatics* **2022**, *38*, 1964.
[30] K. Furui, M. Ohue. *arXiv preprint arXiv:2205.02169*. **2022**.
[31] V. Le Guilloux, P. Schmidtke, P. Tuffery, *BMC Bioinform.* **2009**, *10*, 1.
[32] D. Ni, Z. Chai, Y. Wang, M. Li, Z. Yu, Y. Liu, S. Lu, J. Zhang, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, e1585.
[33] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, *Adv. Neural Inform. Processing Syst.* **2017**, *30*.
[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[35] C. Yin, Z. Song, H. Tian, T. Palzkill, P. Tao, *Phys. Chem. Chem. Phys.* **2023**, *25*, 1349.
[36] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, *Nature Mach. Intell.* **2020**, *2*, 56.
[37] D. Chicco, G. Jurman, *BMC Genom.* **2020**, *21*, 1.
[38] J. G. Greener, M. J. Sternberg, *BMC Bioinform.* **2015**, *16*, 1.
[39] H. Tian, F. Trozzi, B. D. Zoltowski, P. Tao, *J. Phys. Chem. B* **2020**, *124*, 8960.
[40] M. T. Ibrahim, F. Trozzi, P. Tao, *Comput. Struct. Biotechnol. J.* **2022**, *20*, 50.
[41] J. Xie, G. Pan, Y. Li, L. Lai, *J. Chem. Phys.* **2023**, *158*, 105102.