Accurate Prediction of Protein Allosteric Sites Through Automated Machine Learning

Sian Xiao, Hao Tian, and Peng Tao*

Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas, United States of America

E-mail: ptao@smu.edu

Abstract

Allostery is a fundamental process in regulating proteins' activity. The discovery, design and development of allosteric drugs demand for better identification of allosteric sites. Several computational methods have been developed previously to predict allosteric sites using static pocket features and protein dynamics. Here, we present a computational model using automated machine learning for allosteric site prediction. Our model, PASSer2.0, advanced the previous results and performed well across multiple indicators with 89.2% of allosteric pockets appeared among the top 3 positions. The trained machine learning model has been integrated with the Protein Allosteric Sites Server (`https://passer.smu.edu`) to facilitate allosteric drug discovery.

1 Introduction

In biochemistry, allostery is a fundamental process that regulates a protein's functional activity. In allosteric regulation, an effector molecule binds a protein at a site (allosteric site) other than its active site, often resulting in conformational and dynamical changes.[1,2]

There are many reasons why allosteric drug development holds promise: the medicines delivered are more selective and less toxic; they produce fewer side effects;[3,4] they can either activate or inhibit proteins; they can be used in conjunction with orthosteric drugs. Due to these advantages, the usage of allosteric drugs has gradually increased in recent years.

So far, several methods have been developed to detect and predict allosteric sites in proteins, such as normal mode analysis,[5] molecular dynamics (MD) simulations[6] and machine learning (ML) models.[2,7–9] Several current methods are available as web servers or open-source packages, such as AllositePro,[10] AlloPred,[11] SPACER,[12] PARS[13] and PASSer.[9] These studies have demonstrated the feasibility of allosteric sites prediction using methods which combine pocket features and protein dynamics. In these studies, features are calculated by site descriptors describing chemical and physical properties of protein pockets, while the protein dynamics are extracted from molecular dynamics (MD) simulations.

The past decade has witnessed the rapid development of machine learning. ML methods have been shown to be superior in the classification of protein pockets. Allosite[2] and AlloPred[11] used support vector machine (SVM)[14] with curated features. Chen et al[15] used random forest (RF)[16] to construct a three-way predictive model. Our previous study[9] used an ensemble learning method combining the results of eXtreme gradient boosting (XGBoost)[17] and graph convolutional neural networks (GCNN).[18]

Recently, automated machine learning (AutoML) has emerged as a novel strategy to implement machine learning methods to solve real world problems.[19–21] It has been widely applied in nucleic acid[22] and disease studies.[23,24] As the name suggests, AutoML helps to automate the machine learning pipeline, from data processing, model selection and ensembling to hyperparameter tuning without expert knowledge. This saves human power from the time-consuming and iterative tasks of machine learning model development.[19] Also, AutoML offers the opportunities to produce simpler solutions with superior model performance.[21]

In this study, we applied AutoGluon,[25] an AutoML framework developed by Amazon Web Services, for the prediction of protein allosteric sites. Our model, PASSer2.0, is shown

to be robust and powerful under various indicators.

2 Methods

2.1 Protein data

The data used in this work was collected from the Allosteric Database (ASD).[26] There are a total of 1949 entries of allosteric sites, each with different proteins and modulators.[27] To ensure protein quality and diversity, 90 proteins were selected using the previous rules:[2] protein structures with either resolution below 3 angstroms or missing residues in the allosteric sites were removed; redundant proteins that have more than 30% sequence identity were filtered out. The names and IDs of the 90 proteins are extracted from the Protein DataBank[28,29] and listed in Table S1.†

2.2 Site descriptors

FPocket algorithm[30] is used to detect pockets on the surface of the selected proteins. A pocket is labeled as either 1 (positive) if it contains at least one residue identified as binding to allosteric modulators or 0 (negative) if it does not contain such residues. Therefore, a protein structure may have more than one positive label. A total of 2123 pockets were detected with 133 pockets being labeled as allosteric sites. For each of the detected pockets, 19 numerical features are calculated from FPocket and are listed in Table S3.†

2.3 Automated Machine Learning

State-of-the-art ML performance normally requires extensive domain knowledge and experience. This process includes data preparation and preprocessing, feature engineering, model selection and hyperparameter tuning, which are time-consuming and challenging even for experienced practitioners. Automated machine learning aims to free human effort from

this process. Developed by Amazon Web Services, AutoGluon[25] automates these ML tasks and achieves the best performance with no machine learning experience required. Moreover, AutoGluon includes techniques for multi-layer stacking that can further boost ML performance.

AutoGluon is advantageous in: (1) simplicity: straightforward and user-friendly APIs; (2) robustness: no data manipulation or feature engineering required; (3) predictable-timing: ML models are trained within the allocated time; (4) fault-tolerance: the training process can be resumed after interruption. Also, AutoGluon is an open-source library with transparency and extensibility.

In the current study, AutoGluon v0.2.0 is applied with base models listed in Table S2.†

2.4 Performance indicators

For binary classification, the results can be classified as Table 1. This confusion matrix can visualize and evaluate models performance.

Table 1. Binary classification results in confusion matrix

	Real positive	Real negative
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

Various indicators were used to quantify model performance: (1) precision measures how well the model can predict real positive labels; (2) recall measures the ability to classify True Positive and True Negative; (3) F1 score is the weighted average of precision and recall. The indicators are calculated as shown from equation 1-3. The higher the values of these indicators, the better the model's performance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

3 Results and Discussion

From the perspective of data quality, data collection and processing is important in machine learning problems. In terms of data collection, the annually updated Allosteric Database (ASD) provides allosteric proteins and allosteric sites with high resolution, bringing opportunities for allosteric site prediction. The selected proteins are processed by the geometry-based algorithm, FPocket, to identify pockets, which are further used to predict allosteric sites. Compared with other web server and open-source pocket detection packages, FPocket is superior in execution time and the ease to be integrated with other models.

Data imbalance happens in a classification problem where the samples are not equally distributed per class. In the current study, there are a total of 2123 pockets detected with 133 being labeled as positive (allosteric sites). The positive labels only make up 6.26% of the dataset, thus leading to an imbalanced data. Data imbalance normally leads to an unsatisfactory model performance, as the training model could not learn sufficiently from the limited minority examples.

There are mainly two effective ways, oversampling and undersampling, to handle imbalanced dataset. Oversampling expands the size of the minority class by randomly duplicating existing examples or generating new but similar examples. However, this could results in overfitting for some machine learning models. Also, in the context of protein allosteric sites, the generated allosteric sites may not be biologically reasonable. Due to these reasons, random undersampling was used to balance the dataset. Specifically, for each positive label, a

fixed number of negative labels were randomly selected in the same protein.

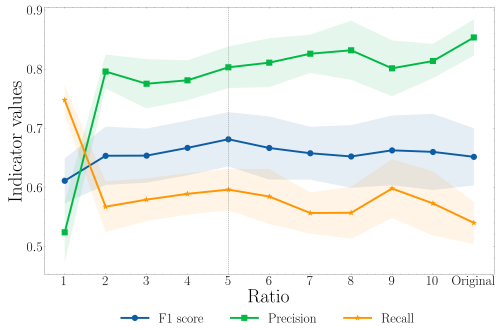

Figure 1. The ratio between negative and positive labels are tuned. The ratio value was adjusted from 1 to 10. Model was trained in 10 independent runs for each ratio. The mean and standard deviation of each metric was calculated. A ratio of 5 is considered reaching a balance between recall and precision with the highest F1 score.

The ratio between the number of negative and positive labels were tuned with results plotted in Figure 1. The ratio value was changed from 1 to 10 along with the original data distribution. Precision, recall and F1 score were calculated for each ratio. A ratio of 5 (number of negative labels : number of positive labels = 5 : 1) reached a balance between the precision and recall with the highest F1 score.

Feature selection is of fundamental importance in machine learning projects. It enables faster training time, reduces model complexity and reduces overfitting. Many machine learning models can benefit from this process with better performance.[31] In the current study, feature selection was conducted through the calculation of feature importance, which is defined as permutation importance.[32] A feature's importance value indicates the performance drop when model making predictions with this feature being randomly shuffled. Higher feature importance represents a more important role for the model's performance. As for negative feature importance, the model suffers from the information introduced by this feature, which undermines the model performance. Thus, the features with negative contribution will be removed to achieve a better performance.

As mentioned above, there are 19 numerical features obtained from FPocket. These features are related with the physical and chemical situations of protein sites. The feature

importance is calculated with mean and standard deviation for each feature with results listed in Table S3.† There are several features (number of alpha spheres, mean local hydrophobic density, polarity score, alpha sphere density and center of mass) with negative feature importance. These features are dropped to improve model performance. The prediction performance before and after dropping these features are listed in Table 2. It is shown that the values of all three indicators increased.

Table 2. Performance indicator values for feature selection. Several features with negative feature importance were dropped to improve the model performance

Indicators	Original features	Tuned features
Precision	0.803 ± 0.063	0.824 ± 0.064
Recall	0.596 ± 0.070	0.647 ± 0.071
F1 score	0.681 ± 0.071	0.725 ± 0.077

Stacking is an ensemble algorithm in machine learning. It applies a meta-learning algorithm to learn to combine different machine learning models in the best arrangement to boost model performance. The main idea of stacking is using predictions of machine learning models from the previous level as input variables for models on the next level.[33] AutoGluon framework uses a multi-layer stacking with k-fold bagging to reduce model's variance. The number of layers and the value of k are heuristically determined within the framework.

The model's performance under non-stacking and stacking are listed in Table 3. F1 score was drastically increased through stacking. While the overall performance is improved, the values of the two indicators (precision and recall) behaved differently. Precision is higher than recall in the non-stacking setting while recall is higher than precision in the stacking setting.

Table 3. Fine-tuning results of model stacking. Performance indicators were calculated for each setting in 10 independent runs

Indicators	Non-stacking	Stacking
Precision	0.824 ± 0.064	0.692 ± 0.077
Recall	0.647 ± 0.071	0.832 ± 0.078
F1 score	0.725 ± 0.077	0.748 ± 0.087

Recall is the fraction of true positive labels being retrieved, and precision is the fraction

of correct predictions on the positive labels. For the purpose of identifying true allosteric sites, a high value in recall is more desired than a high value in precision, because it is more desired to retrieve all real allosteric sites and do further computational study than to have a model that seldom classifies pockets as allosteric sites.

Machine learning models often suffer from underfitting with limited training time. Intuitively, increasing the training time leads to the improvement of performance. In order to fully train each basic classifier, model performance is evaluated with increasing training time as shown in Figure 2.

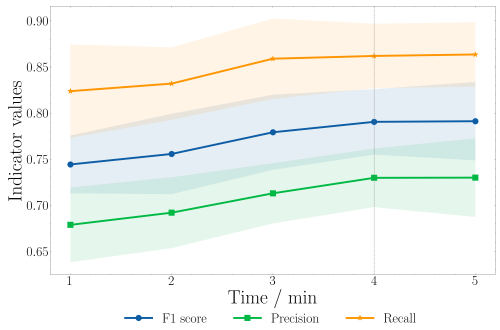

Figure 2. Model performance with different training time. A training time of four was chosen with high F1 score. There was no significant increase in all three indicators for longer training time.

When the training time was increased from two minutes to four minutes, the performance indicators increased as expected. However, there was no significant improvement as the training time was increased to five minutes. In addition, the model tends to be unstable with higher variance in F1 score. Therefore, the training time was chosen as four minutes.

After refinements described above, we achieved the final model. The model was tested using the testing set with results listed in Table 4. Compared with the results from Allosite and PASSer, our model exhibited higher F1 score, precision and recall rates. The relatively small standard deviation indicates the stability of our model.

It is expected that a powerful machine learning model is capable of ranking allosteric sites on the top positions. In the current study, we evaluated the ranking power of our models

Table 4. Model evaluation and comparison. We applied our model to the test set for evaluation. PASSer2.0 displays better performance than PASSer and Allosite in all three performance indicators

Indicators	PASSer2.0	PASSer	Allosite[2]
F1 score	0.790 ± 0.071	0.782	0.761
Precision	0.730 ± 0.063	0.726	0.688
Recall	0.862 ± 0.073	0.847	0.852

by calculating the fraction of allosteric sites on the top 1, 2, and 3 positions and compared it with other models (Table 5). The result indicates that 89.2% of allosteric sites appeared among the top three positions from all detected pockets in the same protein.

Table 5. Probabilities of ranking allosteric sites among the top 3 positions. 89.2% of allosteric sites appeared among the top 3 positions, higher than the previous results

	Top 1	Top 2	Top 3
PARS[13]	44%	62%	73%
AlloPred[11]	57.5%	70.0%	NA[a]
PASSer[9]	60.7%	81.6%	84.9%
PASSer2.0	62.5%	83.2%	89.2%

[a] Not available in the reported results.

In addition, we tested our model using two proteins that are not included in the selected 90 proteins. These two proteins represent two types of allosteric proteins: dynamics-driven and conformational-driven. The second PDZ domain (PDZ2) is a typical dynamics-driven protein in human PTP1E protein. PDZ2 undergoes allosteric process upon binding with peptides.[34] The light-oxygen-voltage domain of Phaeodactylum tricornutum Aureochrome 1a (AuLOV) is a conformational-driven allosteric protein.[35] AuLOV is a monomer in the dark state and undergoes dimerization upon blue light perturbation.[36] Our prediction model ranks the allosteric sites of these two proteins as the top 1 with probabilities of 46.1% and 95.3%, respectively. Both of these results increased through the comparison with the previous results of 45.1% and 89.5%. This indicates that our model could provide reliable predictions for conformational-driven allosteric proteins. However, this model needs further improvement for dynamics-driven allosteric proteins, though is still better than PASSer.

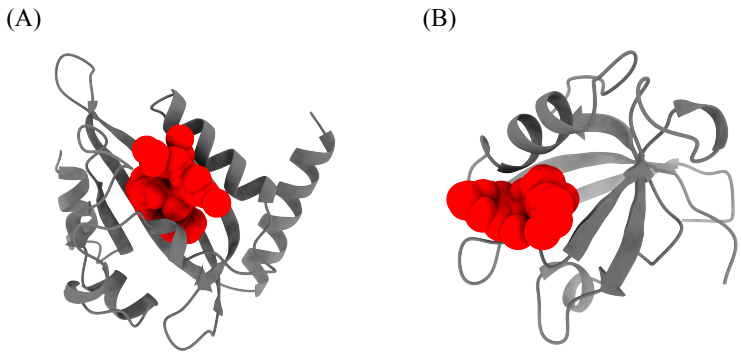

Figure 3. Predictions of two proteins that are not included in the training set: (A) AuLOV (PDB ID: 5DKK); (B) PDZ2 (PDB ID: 3LNY). Red regions are the most probable pockets in the predictions with probabilities of (A) 95.3% and (B) 46.1% and are also the true allosteric sites.

4 Conclusion

Several machine learning based methods have been developed for allosteric site prediction over the past few years. In this study, we applied an emerging ML technique, automated machine learning, to further improve protein allosteric sites prediction models. The AutoML framework is capable of automating the machine learning model pipeline. The developed allosteric sites prediction model, PASSer2.0, is superior to many previous studies. It achieved high values for performance indicators and a good ranking power with a higher percentage of ranking allosteric sites at top positions.

Code and Data availability

The authors declare that all data supporting the findings of this study is available within the paper. Data and processing scripts are available on GitHub at `https://github.com/smu-tao-group/passer_autoML`.

Competing interests

There are no conflicts to declare.

Acknowledgment

Computational time was generously provided by Southern Methodist University's Center for Research Computing. Research reported in this paper was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R15GM122013.

Author Contributions

Sian Xiao: Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing; **Hao Tian:** Conceptualization, Investigation, Software, Validation, Visualization, Writing – review & editing; **Peng Tao:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

References

(1) Srinivasan, B.; Forouhar, F.; Shukla, A.; Sampangi, C.; Kulkarni, S.; Abashidze, M.; Seetharaman, J.; Lew, S.; Mao, L.; Acton, T. B., et al. Allosteric regulation and substrate activation in cytosolic nucleotidase II from L egionella pneumophila. *The FEBS journal* **2014**, *281*, 1613–1628.

(2) Huang, W.; Lu, S.; Huang, Z.; Liu, X.; Mou, L.; Luo, Y.; Zhao, Y.; Liu, Y.; Chen, Z.; Hou, T., et al. Allosite: a method for predicting allosteric sites. *Bioinformatics* **2013**, *29*, 2357–2359.

(3) Wagner, J. R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging computational methods for the rational discovery of allosteric drugs. *Chemical reviews* **2016**, *116*, 6370–6390.

(4) Nussinov, R.; Tsai, C.-J.; Csermely, P. Allo-network drugs: harnessing allostery in cellular networks. *Trends in pharmacological sciences* **2011**, *32*, 686–693.

(5) Panjkovich, A.; Daura, X. Exploiting protein flexibility to predict the location of allosteric sites. *BMC bioinformatics* **2012**, *13*, 1–12.

(6) Laine, E.; Goncalves, C.; Karst, J. C.; Lesnard, A.; Rault, S.; Tang, W.-J.; Malliavin, T. E.; Ladant, D.; Blondel, A. Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor. *Proceedings of the National Academy of Sciences* **2010**, *107*, 11277–11282.

(7) Amor, B. R.; Schaub, M. T.; Yaliraki, S. N.; Barahona, M. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nature communications* **2016**, *7*, 1–13.

(8) Bian, Y.; Jing, Y.; Wang, L.; Ma, S.; Jun, J. J.; Xie, X.-Q. Prediction of orthosteric and allosteric regulations on cannabinoid receptors using supervised machine learning classifiers. *Molecular pharmaceutics* **2019**, *16*, 2605–2615.

(9) Tian, H.; Jiang, X.; Tao, P. PASSer: prediction of allosteric sites server. *Machine Learning: Science and Technology* **2021**, *2*, 035015.

(10) Song, K.; Liu, X.; Huang, W.; Lu, S.; Shen, Q.; Zhang, L.; Zhang, J. Improved method for the identification and validation of allosteric sites. *Journal of chemical information and modeling* **2017**, *57*, 2358–2363.

(11) Greener, J. G.; Sternberg, M. J. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics* **2015**, *16*, 1–7.

(12) Goncearenco, A.; Mitternacht, S.; Yong, T.; Eisenhaber, B.; Eisenhaber, F.; Berezovsky, I. N. SPACER: server for predicting allosteric communication and effects of regulation. *Nucleic acids research* **2013**, *41*, W266–W272.

(13) Panjkovich, A.; Daura, X. PARS: a web server for the prediction of protein allosteric and regulatory sites. *Bioinformatics* **2014**, *30*, 1314–1315.

(14) Suykens, J. A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters* **1999**, *9*, 293–300.

(15) Chen, A. S.-Y.; Westwood, N. J.; Brear, P.; Rogers, G. W.; Mavridis, L.; Mitchell, J. B. A random forest model for predicting allosteric and functional sites on proteins. *Molecular informatics* **2016**, *35*, 125–135.

(16) Liaw, A.; Wiener, M., et al. Classification and regression by randomForest. *R news* **2002**, *2*, 18–22.

(17) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

(18) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**,

(19) Yao, Q.; Wang, M.; Chen, Y.; Dai, W.; Li, Y.-F.; Tu, W.-W.; Yang, Q.; Yu, Y. Taking human out of learning applications: A survey on automated machine learning. *arXiv preprint arXiv:1810.13306* **2018**,

(20) Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine* **2020**, *104*, 101822.

(21) Elshawi, R.; Maher, M.; Sakr, S. Automated machine learning: State-of-the-art and open challenges. *arXiv preprint arXiv:1906.02287* **2019**,

(22) Chen, Z.; Zhao, P.; Li, C.; Li, F.; Xiang, D.; Chen, Y.-Z.; Akutsu, T.; Daly, R. J.; Webb, G. I.; Zhao, Q., et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic acids research* **2021**, *49*, e60–e60.

(23) Karaglani, M.; Gourlia, K.; Tsamardinos, I.; Chatzaki, E. Accurate blood-based diagnostic biosignatures for Alzheimer's disease via automated machine learning. *Journal of clinical medicine* **2020**, *9*, 3016.

(24) Panagopoulou, M.; Karaglani, M.; Manolopoulos, V. G.; Iliopoulos, I.; Tsamardinos, I.; Chatzaki, E. Deciphering the Methylation Landscape in Breast Cancer: Diagnostic and Prognostic Biosignatures through Automated Machine Learning. *Cancers* **2021**, *13*, 1677.

(25) Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; Smola, A. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505* **2020**,

(26) Huang, Z.; Zhu, L.; Cao, Y.; Wu, G.; Liu, X.; Chen, Y.; Wang, Q.; Shi, T.; Zhao, Y.; Wang, Y., et al. ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic acids research* **2011**, *39*, D663–D669.

(27) Liu, X.; Lu, S.; Song, K.; Shen, Q.; Ni, D.; Li, Q.; He, X.; Zhang, H.; Wang, Q.; Chen, Y., et al. Unraveling allosteric landscapes of allosterome with ASD. *Nucleic acids research* **2020**, *48*, D394–D401.

(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235–242.

(29) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S., et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* **2019**, *47*, D464–D474.

(30) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **2009**, *10*, 1–11.

(31) Pereira, J. C.; Caffarena, E. R.; Dos Santos, C. N. Boosting docking-based virtual screening with deep learning. *Journal of chemical information and modeling* **2016**, *56*, 2495–2506.

(32) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347.

(33) Pavlyshenko, B. Using stacking approaches for machine learning models. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). 2018; pp 255–258.

(34) Zhou, H.; Dong, Z.; Tao, P. Recognition of protein allosteric states and residues: Machine learning approaches. *Journal of computational chemistry* **2018**, *39*, 1481–1490.

(35) Heintz, U.; Schlichting, I. Blue light-induced LOV domain dimerization enhances the affinity of Aureochrome 1a for its target DNA sequence. *Elife* **2016**, *5*, e11860.

(36) Tian, H.; Trozzi, F.; Zoltowski, B. D.; Tao, P. Deciphering the allosteric process of the Phaeodactylum tricornutum Aureochrome 1a LOV domain. *The Journal of Physical Chemistry B* **2020**, *124*, 8960–8972.