Prediction of large conformational changes of a protein binding pocket associated with ligand binding

Kanta Sakai and Takashi Ishida*

Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro, Tokyo, Japan

ABSTRACT: Docking simulation, a key technique in virtual screening, typically treats proteins as rigid bodies. However, proteins are inherently flexible, and ligand binding can induce significant conformational changes, affecting prediction accuracy. This study proposes a new approach to identify protein binding pockets that exhibit substantial conformational changes upon ligand binding, potentially impacting docking simulation accuracy. In this research, we developed a prediction model using graph neural network to identify protein pockets with large conformational changes. To train the model, we constructed a dataset by calculating conformational changes in ligand-binding sites between multiple holo structures corresponding to the apo structure. We evaluated the performance of the prediction model and the results demonstrated that our model could identify proteins with significant conformational changes, although the prediction accuracy remains low, with an AUC of 0.58 on the test data. This study highlights the potential of deep learning approaches in addressing the challenges of protein flexibility in virtual screening and docking simulations.

INTRODUCTION

Virtual screening is a computational technique used to select drug candidate compounds from vast compound libraries, estimated to contain between 10^{30} and 10^{60} potentially usable compounds[1,2]. This technique primarily evaluates the presence or absence of activity between compounds and drug target proteins, as it is crucial for drug candidate compounds to bind to target proteins and exhibit pharmacological effects[3]. There are two main approaches to virtual screening: ligand-based virtual screening (LBVS)[4], which uses information from known active compounds, and structure-based virtual screening (SBVS)[5], which utilizes the three-dimensional structure of proteins. LBVS employs methods such as chemical similarity, pharmacophore modeling, and machine learning to build regression or classification prediction models based on known experimental information, enabling rapid selection from extensive compound libraries. This method is applicable even when the target protein's structure is unknown, provided there are experimentally known active compounds. However, it relies on the assumption that similar molecules have similar properties, which can lead to missing novel compound candidates[6]. On the other hand, SBVS evaluates the binding affinity between proteins and compounds based on physicochemical interactions such as van der Waals forces, Coulomb forces, and hydrogen bonding, enabling the selection of compounds when the target protein's structure is known. SBVS has the advantage of discovering novel drug candidate compounds without relying on known experimental information.

One of the key technologies in SBVS is docking simulation, a computational technique that predicts the interaction between compounds (ligands) and target molecules. The process of docking simulation starts with acquiring the three-dimensional structure of the target protein, followed by identifying the binding site, preparing the compound library, executing the docking, scoring the binding affinity, and finally selecting lead compounds[7,8]. However, proteins are inherently flexible, and the impact of conformational changes in the binding site due to ligand binding on docking simulation is significant.

The conformational changes in the binding site of proteins due to ligand binding affect docking simulation[9]. Experimental data-based ligand-bound structures (holo structures) have been shown to have higher docking simulation accuracy than unbound structures (apo structures)[10,11] . However, many proteins lack experimentally obtained holo structures. Therefore, docking simulations that consider protein conformational changes[12] have been considered, but they are not common due to increased computational requirements. Thus, methods for generating holo structures from apo structures have been researched to prepare suitable three-dimensional structures for docking simulation.

There is abundant research on estimating holo structures as suitable three-dimensional structures for docking simulation, using techniques such as molecular dynamics (MD) simulations and machine learning in recent years[13]. However, these holo structure estimation methods have problems, and there is no practical method for estimating holo structures. Methods using MD simulations require information about the ligand or holo structure, which is not available for proteins without known holo structures[14]. Moreover, holo structure estimation methods using MD simulations with templates, which are popular approaches for proteins without known holo structures, have the problem of limited applicability to proteins for which suitable templates cannot be prepared[15]. Additionally, the case study of ButF demonstrates the limitations of template-based methods and MD simulations with constraints based on known information[16].

Machine learning-based holo structure estimation methods have been developed in recent years, offering advantages such as lower computational requirements after training and no need for specific templates. However, these methods can lose generalization performance unintentionally due to the size and bias of the training dataset. Generative models used in these methods might have low generalization performance due to the small scale of the dataset and the low accuracy of the generated structures[17,18].

Considering the absence of practical methods for estimating holo structures with usable accuracy, identifying proteins with significant conformational changes in the binding site due to ligand binding is crucial for minimizing the negative impact on docking simulation. This approach could also contribute to the

development of holo structure estimation methods. For holo structure estimation using MD simulations, limited operation for target proteins that affect the accuracy of virtual screening can solve the problem of high computational requirements. For holo structure estimation using machine learning, excluding proteins with small structural differences between the apo and holo structures of the ligand-binding site from the dataset is expected to improve accuracy. Therefore, limiting the dataset to proteins with large structural differences in the ligand-binding site between the apo and holo structures is considered to lead to improved accuracy of holo structure estimation methods.

In light of the challenges associated with holo structure estimation and the importance of identifying proteins with significant conformational changes in the binding site, this study aims to develop a deep learning-based method to distinguish proteins that exhibit substantial conformational changes in the pocket due to ligand binding. Specifically, we will focus on estimating whether the conformational changes in the pockets between the apo and holo structures are significant, with the goal of enhancing the accuracy and applicability of docking simulations and virtual screening in drug discovery.

MATERIALS AND METHODS

The flow of this study is outlined in Figure 1. The process begins with searching for the corresponding apo structures for the dataset of holo structures and calculating the conformational changes in the pocket between the apo and holo structures. As multiple holo structures may correspond to a single apo structure, we determine the label based on the maximum value of the pocket's structural change. Next, graph data are generated from the apo structures using the types and coordinates of amino acid residues and the centroid coordinates of the pocket. The node features include the type of amino acid residue and the distance from the pocket centroid, while the edge features use weights based on the distance between amino acid residues. Finally, we train and evaluate the prediction model using the graph data and the label based on the maximum structural change in the pocket.

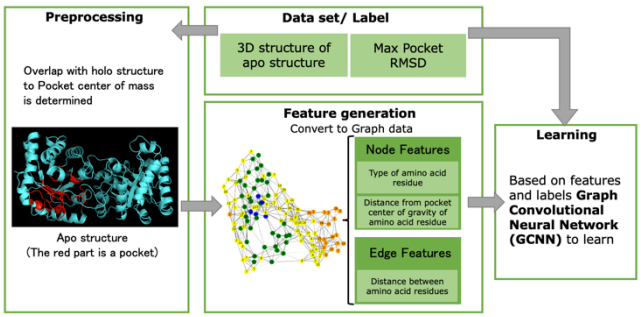

Figure 1. Outline of Research Methods

Datasets construction. For this study, we aim to discriminate proteins with significant conformational changes in the pocket due to ligand binding using deep learning, with the input being the three-dimensional structure of apo proteins. To achieve this, data on the conformational changes in the pocket between apo and holo proteins is required. Existing datasets only record one holo structure for each apo structure, not considering multiple holo proteins corresponding to one apo protein and are also small in scale[19,20]. In this research, we constructed a new dataset that records the pocket conformational changes for multiple holo proteins corresponding to an apo protein. The detailed procedure for constructing this dataset is described in the Supporting Information.

Protein Structure Graph Generation. Protein structure is deeply related to its function, and detailed analysis of its structural features is crucial for predicting pocket conformational changes due to ligand binding. However, directly handling the complex three-dimensional structure of proteins is computationally very challenging. Therefore, it is necessary to simplify the data representation while retaining the necessary information. In this study, we converted the protein three-dimensional structure data into graph data, which is commonly used as input for deep learning, to capture the essential features of the three-dimensional structure formed by the protein's molecular chain and represent this structural information in a form processable by machine learning models.

The edge feature is the weight using the distance between the coordinates of the amino acid residues, adopting $\frac{1}{1+x}$ (distance$= x$) as the weight. For the condition of forming edges, in the Test and Validation data, edges are formed between amino acid residues with a distance of 8Å or less. In the Train data, for distances greater than 6Å, the probability of forming an edge approaches zero as the distance approaches 10Å, according to the following formula: $\frac{10-x}{4}$ (if $6 < x < 10$). This condition for forming edges is aimed at data augmentation and is referred to as "random edges." The representative point of the amino acid residue is the C_β atom coordinate, which is used to consider the influence of the distance between side chains.

The node features used are the type of amino acid residue and the distance from the pocket centroid coordinate of the amino acid residue. The type of amino acid residue is one-hot encoded into a total of 21 types, including 20 standard amino acids and others, after converting non-standard amino acids to similar standard amino acids. The distance from the pocket centroid coordinate of the amino acid residue is input directly as a numerical value.

The edge feature is the weight using the distance between the coordinates of the amino acid residues, adopting $\frac{1}{1+x}$ (distance$= x$) as the weight.

Labeling. The label of a protein structure graph is defined based on its conformational changes between apo and holo structures. We use the root mean square distance (RMSD) of c-alpha atoms of pocket residues and define the labels as follows:

- Negative: Proteins with a pocket RMSD in the range of $0.0\,\text{Å} \leq$ Pocket RMSD$< 1.5\,\text{Å}$ are considered to have small conformational changes.
- Positive: Proteins with a pocket RMSD in the range of $1.5\,\text{Å} \leq$ Pocket RMSD$< 10\,\text{Å}$ are considered to have large conformational changes.

Proteins with a pocket RMSD of 10Å or more are excluded because such large-scale conformational changes often involve OPEN/CLOSE conformational changes and it is not the scope of this research.

Data Splitting and Augmentation. To demonstrate that the proposed method can be applied to apo proteins whose holo structures are not known to have high sequence similarity, we divided the data so that proteins with high sequence similarity to the Test data are not included in the Train data. Specifically, we defined protein groups with sequence similarity (using BLAST+[21]) of 50% or more as groups with high sequence similarity and divided the 830 groups with high sequence similarity into Train, Test, and Validation data so that the protein units do not overlap across the data and the ratio is close to 8:1:1.

To address the imbalance between Positive and Negative classes in the Train data, we employed “Random OverSampling” from the imbalanced-learn library in scikit-learn[22] to equalize the

number of samples in each class. Additionally, we applied the random edge augmentation technique described in the Feature Extraction section to increase the size of the Train dataset by tenfold, enhancing the diversity of the training data and potentially improving the robustness of the model.

Performance Evaluation. For performance evaluation, we adopted the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve as a generic and balanced performance metric for the prediction model, which is a binary classification problem that determines whether the structural change in the binding pocket is large or small. AUC is a reliable evaluation compared to Accuracy, especially for imbalanced datasets, as it considers the trade-off between false positive rate and true positive rate. Additionally, the ROC curve visually represents the performance trade-off of the model, facilitating comprehensive comparison and analysis between different models and settings.

Deep Learning Model. We chose a simple Graph Convolutional Neural Network (GCNN)[23] that requires relatively few features, can efficiently generate necessary data, and is effective for learning with a small amount of data, as the dataset is small, and it is necessary to use the entire protein as input data. The GCNN model uses graph data with amino acid residues of the apo protein as nodes and the bonds between residues as edges. The optimization method is Adam[24], and the loss function is Binary Cross Entropy Loss.

The model configuration was set using Optuna[25], a Python library for automating hyperparameter optimization. The parameters and their search ranges explored with Optuna are shown in Table 1. The shapes of the convolutional layers are pyramid, rhombus, and uniform, with the channel number increasing in the first layer and decreasing towards the output layer in the pyramid, the channel number reaching the maximum in the middle layer in the rhombus, and all layers having the same channel number in the uniform.

Table 1. Hyperparameters and their search ranges used in the optimization process with Optuna.

Parameter	Search Range
Number of Convolutional Layers	3, 4, 5, 6, 7, 8, 9, 10
Shape of Convolutional Layers	pyramid, rhombus, uniform
Number of Channels	1 to 512
Pooling Layer	mean, max, add
Dropout	0.1 to 0.7
Batch Normalization	using or not using
Learning Rate	1e-5 to 1e-2 (searched on a log scale)

RESULTS

Dataset generation. We obtained 19,444 holo proteins from the PDBbind version 2020 database[26,27] and 3,535 corresponding apo proteins. Defining proteins with sequence homology of 99% or more as identical, we identified 1,480 unique proteins. No proteins with multiple pockets were found in this dataset.

Data Splitting. The data was split so that groups with high sequence similarity (groups with sequence similarity of 50% or more) did not overlap between the Train, Test, and Validation data. The splitting was done at the protein level in a ratio close to 8:1:1. Specifically, groups with a number of proteins equal to or less than one-tenth of the number of proteins allocated to the Test and Validation data were randomly assigned to the Test and Validation data, and the remaining data was assigned to the Train data. After random splitting, manual adjustments were made to the Test and Validation data to ensure that the ratio of Positive to Negative labels did not significantly deviate. The results of the data splitting are shown in Table2

Table 2. Results of Datasets Splitting

data type	label	Groups with High Sequence Similarity	Proteins	Apo Protein Binding Pockets
Train	Negative	185	359	901
	Positive	476	852	2205
Test	Negative	83	87	95
	Positive	26	45	124
Validation	Negative	71	74	91
	Positive	36	57	116

Performance Evaluation. Using the parameters determined by Optuna, where the convolutional layer shape was uniform, the number of channels was 34, the pooling layer was mean, dropout was 0.634, batch normalization was applied, and the learning rate was 3.836e-5, we constructed the GCNN model. The model was trained with the dataset, and the performance was evaluated on the Test data.

The ROC curve for the test data is shown in Figure 2. The AUC was 0.58. It shows that the model's learning was better than random prediction. Especially, the true positive rates at the low false positive rate regions are relatively better and it indicates that large conformational change of protein pocket would be predictable.

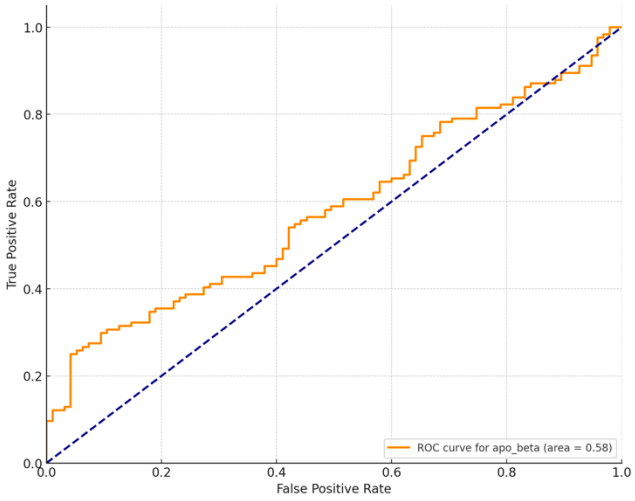

Figure 2. An ROC curve of the proposed method for the test dataset. The AUC = 0.58.

Case Study. We examined cases where the model accurately predicted proteins with large conformational changes in the pocket and cases where predictions were significantly off. For accurately predicted cases, we found that the model could distinguish

proteins with large conformational changes relatively well but had low accuracy for proteins with small conformational changes. Additionally, even among correctly predicted apo proteins, the pocket structural change and the predicted value were not proportional, indicating that the model was not capturing features that influence the pocket's structural change. In the case study (Figure 3), among the examples where predictions were significantly off, there were some proteins that were incorrectly predicted to have large conformational changes. Notably, within this group, some proteins had structural similarities to proteins that were correctly predicted to have large conformational changes. This suggests that the learning process may become challenging due to being influenced by partial structural similarities. For instance, Example A in Figure 3 shows a protein that was correctly predicted to have a large structural change, whereas Example B shows a protein that was incorrectly predicted to have a large structural change but shares some structural similarities with Example A.

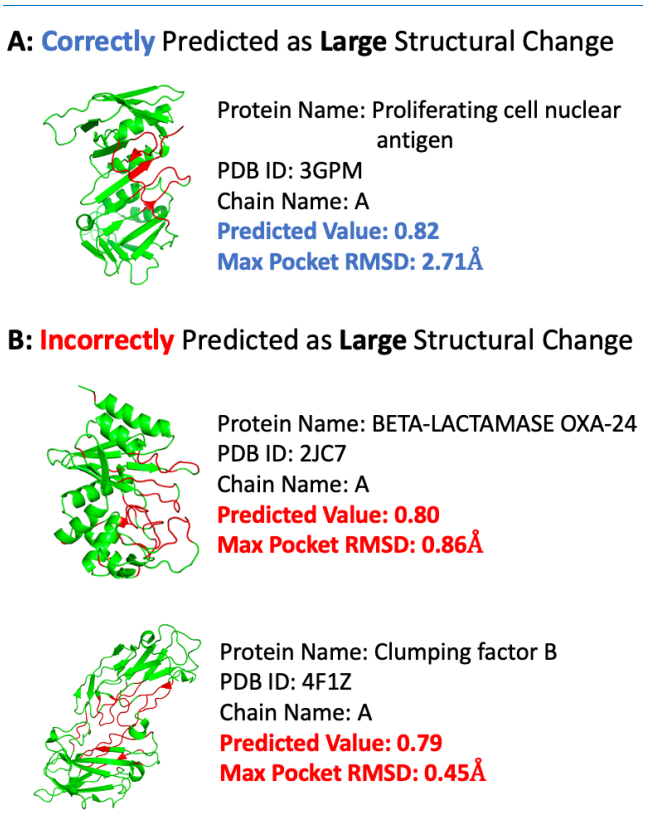

Figure 3. This figure illustrates examples from the case study where the deep learning model was used to predict large conformational changes in protein pockets due to ligand binding. The red highlighted areas indicate the pocket regions. The model's predictions are compared to the actual maximum pocket root mean square deviation (RMSD) to assess its accuracy.

DISCUSSION

Analysis of Dataset Distribution Using t-SNE. The dataset was constructed such that protein pairs with sequence similarity of 50% or more do not overlap between the Train, Test, and Validation data. However, protein pairs with sequence similarity below 50% exist across the different data sets. By visualizing the sequence similarity relationships of the dataset using t-SNE[28] (Figure 4), it was observed that some protein pairs across the data sets have higher sequence similarity compared to pairs within the same dataset.

This study aims to learn the features of apo proteins that undergo conformational changes in the pocket due to ligand binding through deep learning. Since the input includes amino acid residue information, it is likely that features from proteins with high sequence similarity are more easily learned. However, depending on the bias of the dataset and not achieving generalization performance is undesirable, as it would mean being heavily influenced by proteins with high sequence similarity. Therefore, as a comparison, we generated a distance matrix from sequence similarity and performed predictions using the k-nearest neighbors method[29]. The results for different k values on the Validation and Test data are shown in Table 3. From the results, it can be inferred that learning is occurring from information other than sequence similarity, as the proposed method has higher accuracy on the Validation data. On the other hand, the k-nearest neighbors method has higher accuracy on the Test data than the proposed method, indicating that there is a different trend in the Test data despite the high sequence similarity proteins not having the same correct labels in the Validation data. The inclusion of such differently characterized data may make learning more challenging. Moreover, when compared based on Validation Accuracy, the model has better accuracy, so this method is not selected, and the proposed method is superior from the perspective of generalization performance.

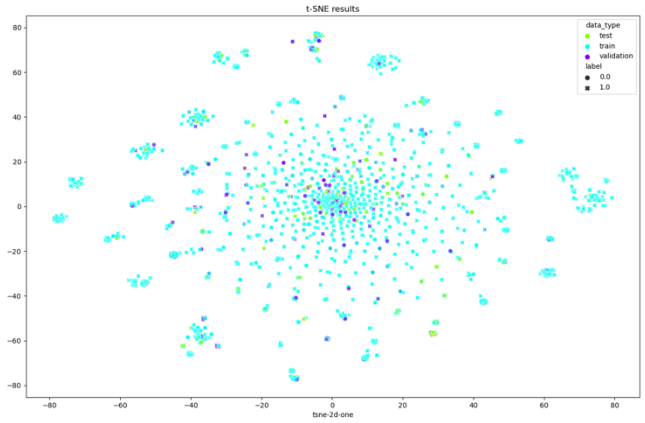

Figure 4. t-SNE by sequence homology of datasets. .

Table 3. Comparison of Validation Accuracy and Test Accuracy for different k values

k value	Validation Accuracy	Test Accuracy
3	0.47	0.60
5	0.48	0.60
10	0.52	0.63
15	0.50	0.62
Our Method (threshold=0.5)	0.66	0.56

Effect of Random Edge Augmentation. In this study, we employed a unique data augmentation technique called "random edges" due to the limited size of our dataset. However, such data augmentation has not been used in other studies, and the

effectiveness of random edges remains unclear. To assess the efficacy of random edges, we conducted a comparative experiment.

The ROC curve for the Test data without random edges is shown in Figure 5, with the results overlaid with those obtained using random edges. The AUC without random edges was 0.48, indicating a significant drop in accuracy and a result worse than random judgment. This decrease in accuracy could be due to the GCNN model's parameters being optimized for a dataset increased tenfold by random edges.

We also examined the results when optimizing the parameters using Optuna without random edges. The results showed a Validation AUC of 0.70 and a Test AUC of 0.54. Although the Validation AUC was higher than that obtained with random edges, the Test AUC remained low. This suggests that the dataset is too small for the learning method without random edges, leading to low generalization, and that random edges have a certain effect.

However, the use of random edges also has its drawbacks. As shown in the supporting information (Figure S2 and Figure S3), the prediction accuracy is higher when random edges are used, but the difference between the training and validation data caused by random edges leads to increased adverse effects of overfitting as learning progresses.

In conclusion, while random edges contribute to improved prediction accuracy, they also introduce challenges in terms of overfitting, highlighting the need for careful consideration when using data augmentation techniques in machine learning models for protein structure prediction.

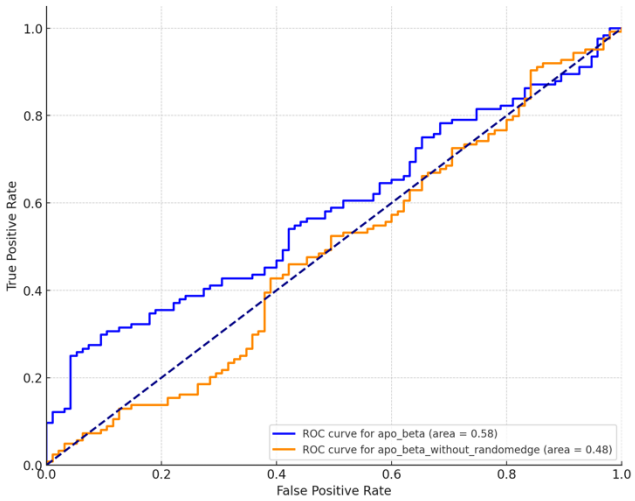

Figure 5. Comparison of ROC curves with and without random edge augmentation.

CONCLUSIONS

In this study, we proposed a method that involves creating a new dataset that considers the pocket conformational changes between apo proteins and their corresponding multiple holo structures. We hypothesized that by training with this dataset, we could discriminate proteins with large conformational changes in the pocket due to ligand binding. We conducted experiments to build the dataset and improve the model. The results showed that the AUC for the Test data of the constructed dataset was 0.58, indicating superior performance to random guessing. This suggests that the proposed method has some learning effect, especially in accurately identifying proteins with large conformational changes in the pocket due to ligand binding.

However, the results of this study are not sufficient for practical application as a preprocessing step for the holo structure prediction problem. One possible reason for the lack of improvement in the model's prediction accuracy is the insufficient size of the dataset.

Furthermore, as an additional experiment, we generated a distance matrix from sequence similarity and conducted predictions using the k-nearest neighbors method. By comparing these results, we demonstrated that the model could learn without depending on sequence similarity. However, the apparent differences in the properties of the Validation and Test data suggest that the differences in properties due to groups with high sequence similarity may be larger than expected, indicating that the learning results may depend on the data splitting.

The dataset created in this study is larger compared to existing datasets, but considering the experimental results, it can be said that it is not sufficiently large. There is also a problem of bias in the number of apo protein structures included in the constructed dataset. Specifically, there are 2 proteins with more than 100 apo structures, 4 proteins with more than 50, and 9 proteins with more than 20, which may have caused biased learning if these proteins are included in the Train data. Furthermore, proteins with many experimentally obtained structures may contain unintended biases, posing a problem from the perspective of diversity. Searching for proteins with sequence similarity of 80% or more within the dataset resulted in only proteins that are already present in the dataset, indicating that the construction of the dataset using PDBbind, while including more proteins than existing studies, may not be utilizing specific proteins effectively. One solution is to increase the amount of data. As a method to increase the diversity and volume of the dataset, utilizing all pairs of holo and apo proteins present in the PDB can be considered. However, it is necessary to clearly define ligands, holo proteins, and apo proteins and filter protein structures from the collected PDB that are valuable as a dataset. The number of proteins included in the PDB is constantly increasing, and as of January 2024, 215,092 structures are registered[30]. The cost of constructing such a large-scale dataset is a significant problem.

ASSOCIATED CONTENT

Supporting Information.

AUTHOR INFORMATION

Corresponding Author

Takashi Ishida – Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro, Tokyo, Japan
Email: ishida@c.titech.ac.jp

Authors

Kanta Sakai – Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro, Tokyo, Japan
Email: sakai@cb.cs.titech.ac.jp

Author Contributions

K.S provided the methodology, experiments, implementation and wrote the original draft. T.I supervised the research and wrote the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT
This study was carried out using the TSUBAME3.0 supercomputer at Tokyo Institute of Technology.

REFERENCES

1. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: A molecular modeling perspective. *Med Res Rev*. 1996;16(1):3-50.

2. Walters WP. Virtual Chemical Libraries. *J Med Chem*. 2019;62(3):1116-1124.

3. Paul SM, others. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. 2010;9:203-214.

4. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20:318-331.

5. Lionta E, Spyrou G, Vassilatis DK, Cournia Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr Top Med Chem*. 2014;14:1923-1938.

6. Wang Z, others. Combined strategies in structure-based virtual screening. *Physical Chemistry Chemical Physics*. 2020;22:3149-3159.

7. Nichol AQ, Dhariwal P, Ramesh A, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol 162. ; 2022:16784-16804.

8. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model*. 2021;61(8):3891-3898.

9. Najmanovich R. Protein flexibility upon ligand binding: Docking predictions and statistical analysis. *ArXiv*. 2013.

10. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem*. 2003;46(14):2895-2907.

11. Aggarwal R, Gupta A, Priyakumar UD. APObind: A Dataset of Ligand Unbound Protein Conformations for Machine Learning Applications in De Novo Drug Design. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*. 2021.

12. Ravindranath PA, others. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Comput Biol*. 2015;11:e1004586.

13. Zhang J, Li H, Zhao X, Wu Q, Huang SY. Ligand-Binding-Site Refinement to Generate Reliable Holo Protein Structure Conformations from Apo Structures. *J Chem Inf Model*. 2022;62(22):5806-5820.

14. Seeliger D, de Groot BL. Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput Biol*. 2010;6(1):e1000634.

15. Zhang J, Li H, Zhao X, Wu Q, Huang SY. Holo Protein Conformation Generation from Apo Structures by Ligand Binding Site Refinement. *J Chem Inf Model*. 2022;62(22):5806-5820.

16. Wang D, Weng J, Wang W. An unconventional ligand-binding mechanism of substrate-binding proteins: MD simulation and Markov state model analysis of BtuF. *J Comput Chem*. 2019;40(14):1440-1448.

17. Zhang X, Geffner T, McPartlon M, et al. Bending and Binding: Predicting Protein Flexibility upon Ligand Interaction using Diffusion Models. *NeurIPS 2023 - Generative AI and Biology (GenBio@NeurIPS2023)*. 2023.

18. Nakata S, Mori Y, Tanaka S. End-to-end protein–ligand complex structure generation with diffusion-based generative models. *BMC Bioinformatics*. 2023;24(1):233.

19. Gunasekaran K, Nussinov R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J Mol Biol*. 2007;365(1):257-273.

20. Peng C, Zhang X, Xu Z, et al. D3PM: a comprehensive database for protein motions ranging from residue to domain. *BMC Bioinformatics*. 2022;23(1):70.

21. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 2004;32(Web Server issue):W20-W25.

22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.

23. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: *International Conference on Learning Representations*. ; 2017.

24. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *CoRR*. 2014;abs/1412.6980.

25. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ; 2019:2623-2631.

26. Stärk H, Ganea O, Pattanaik L, Barzilay DrR, Jaakkola T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. In: *Proceedings of the 39th International Conference on Machine Learning*. ; 2022:20503-20521.

27. Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics*. 2014;31(3):405-412.

28. van der Maaten L, Hinton G. Viualizing data using t-SNE. *Journal of Machine Learning Research*. 2008;9:2579-2605.

29. Cunningham P, Delany S. k-Nearest neighbour classifiers. *Mult Classif Syst*. 2007;54.

30. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.