KYOTO UNIVERSITY

DISSERTATION

Machine Learning-Based Methods for Predicting the Most Stable Conformation and Binding Affinity of Protein-Drug Complexes

機械学習に基づくタンパク質-薬剤分子の

最安定配座および結合親和性の予測手法

Author:

Koji Shiota

Supervisor:

Prof. Akutsu Tatsuya

Abstract

In drug discovery, virtual screening (VS) is an indispensable technique for efficiently extracting drug candidates from huge compound libraries. Among the technologies required for VS, Protein-Ligand (P-L) binding affinity prediction is an essential technology that forms the basis of VS, and many methods have been developed since the emergence of VS. Recently, a number of machine-learning based P-L binding affinity prediction methods have been reported with great success. In this dissertation, we introduce two novel machine-learning based P-L binding affinity prediction methods. The first is AQDnet and the second is multi-shelled ECIF.

We have developed AI QM Docking Net (AQDnet), which utilizes the three-dimensional structure of P-L complexes to predict binding affinity. This system is novel in two respects: first, it significantly expands the training dataset by generating thousands of diverse ligand configurations for each protein–ligand complex and subsequently determining the binding energy of each configuration through quantum chemistry computation. Second, we have devised a method that incorporates the atom-centered symmetry function (ACSF), highly effective in describing molecular energies, for the prediction of protein–ligand interactions. These advancements have enabled us to effectively train a neural network to learn the P-L quantum energy landscape (P-L QEL). Consequently, we have achieved a 92.6% top 1 success rate in the CASF-2016 docking power, placing first among all models assessed in the CASF-2016, thus demonstrating the exceptional docking performance of our model.

Extended connectivity interaction features (ECIF) is a method developed to predict protein–ligand binding affinity, allowing for detailed atomic representation. It performed very well in terms of Comparative Assessment of Scoring Functions 2016 (CASF-2016) scoring power. However, ECIF has the limitation of not being able to adequately account for interatomic distances. To investigate what kind of distance representation is effective for P-L binding affinity prediction, we have developed two algorithms that improved ECIF's feature extraction method to take distance into account. One is multi-shelled ECIF, which takes into account the distance between atoms by

dividing the distance between atoms into multiple layers. The other is weighted ECIF, which weights the importance of interactions according to the distance between atoms. A comparison of these two methods shows that multi-shelled ECIF outperforms weighted ECIF and the original ECIF, achieving a CASF-2016 scoring power Pearson correlation coefficient of 0.877.

Acknowledgements

My supervisor, Professor Akutsu, has provided me with invaluable assistance and direction during my PhD studies, for which I am really grateful. His knowledge, understanding, and perseverance have been really helpful to me as I have grown academically and finished this thesis.

I express my gratitude to the JT team for their generosity, counsel, and cooperation. Their help has given me a wider perspective on my job and made a substantial contribution to my research. In addition, I am deeply grateful to JT for the full support of my tuition.

I would also like to sincerely thank each and every person in Prof. Akutsu's laboratory. Working with such intelligent and dedicated people has been tremendously illuminating and inspiring. I have found the friendly and supportive environment of the lab to be very beneficial to my PhD program.

I have to sincerely thank my friends for helping me along the way with everyday life support. Their continuous understanding, friendship, and support have been invaluable throughout difficult times.

Finally, I would like to thank my family for their unconditional love and continuous support. It has helped me to overcome even the most difficult times. I sincerely thank them for all they have done for me.

Contents

Figure Contents

Table Contents

Chapter 1

Introduction

1.1. Background

Virtual screening (VS) is a method for efficiently extracting potential drugs from a large library of compounds. Initially developed to streamline high-throughput screening (HTS), VS has recently played a very important role in drug discovery.

Virtual Screening (VS) has been developed as an innovative alternative to High Throughput Screening (HTS), offering a more rapid and less labor-intensive approach. Therefore, we will first explain HTS and then provide much more detail about VS.

1.1.1 High Throughput Screening

HTS is a key method used in the early stage of drug discovery to quickly find potential drugs from large collections of compounds. This technique simultaneously tests several chemical compounds against target proteins through automated procedures. Discovering 'hit' compounds with the necessary biological activity is the main objective of HTS.

Subsequently, these hits can be refined to enhance their potency, specificity, and drug-likeness (Hughes, et al., 2011).

HTS employs a variety of assay technologies. For example, biochemical assays evaluate the interaction of a compound with a target protein or enzyme, while cell-based assays evaluate the effect of a compound on cellular processes and phenotypes. The choice of assay type is guided by biological relevance to the disease mechanism (Blay, et al., 2020). Automation and robotics play a key role in HTS, facilitating the handling of small amounts of reagents and biological samples with high precision and reproducibility (Hansel, et al., 2022; Lorenz and Dejan, 2009).

Identifying novel drug candidates takes less time and money thanks to HTS's simplification of the early discovery phase. Furthermore, compared to conventional methods, it enables a more thorough exploration of chemical space, raising the possibility of finding novel and potent medicinal molecules. The efficiency and scalability of HTS, in spite of its high initial setup cost, make it a vital instrument in the contemporary drug discovery process, speeding up therapeutic research and ultimately enabling the delivery of new therapies to patients.

1.1.2 Virtual Screening

In the realm of drug development, virtual screening (VS) is a computational tool used to make it easier to identify interesting compounds from enormous libraries of chemical structures. VS uses computer algorithms to predict the interaction between chemicals and biological targets, such as proteins or enzymes, linked to a disease, in contrast to high throughput screening, which depends on physical trials (Hughes, et al., 2011).

Similar to HTS, the primary objective of VS is to identify "hit" compounds that exhibit the necessary biological activity. But unlike HTS, VS is applied to a bigger sample size of compounds and effectively reduces the enormous pool of possible compounds to a manageable subset that shows strong biological activity potential. This minimizes the need for expensive and time-consuming experimental experiments (Blay, et al., 2020; Hughes, et al., 2011).

VS approaches can be broadly classified into two categories: structure-based drug discovery (SBDD) and ligand-based drug discovery (LBDD) (Yu and MacKerell, 2017). SBDD involves the use of three-dimensional structures of biological targets to identify compounds that are likely to bind to their active sites. This method requires detailed knowledge of the target's structure, typically obtained through techniques like X-ray crystallography or NMR spectroscopy. LBDD, on the other hand, does not rely on the structural information of the target. Instead, it utilizes the properties and characteristics of known active compounds to search for new compounds with similar features, based on the principle that similar molecules tend to exhibit similar biological activities. This doctoral dissertation is particularly focused on methods related to SBDD.

There are various benefits to incorporating VS into the drug discovery process. By facilitating the quick and affordable screening of sizable chemical libraries, it greatly expedites the first phases of drug discovery and offers a better throughput than conventional experimental methods. Furthermore, VS can reveal new scaffolds or chemotypes that might not have been found using empirical techniques, increasing the chemical variety of substances that can be studied further.

1.1.3 Stages of Virtual Screening

VS can be divided into the following four major steps: selection of target proteins, identification of ligand binding sites, identification of ligand binding poses, and comparison of binding affinities among ligands (Dhakal, et al., 2022). Each of these is discussed in detail below. Figure 2.1 shows a schematic representation of each stage of VS.

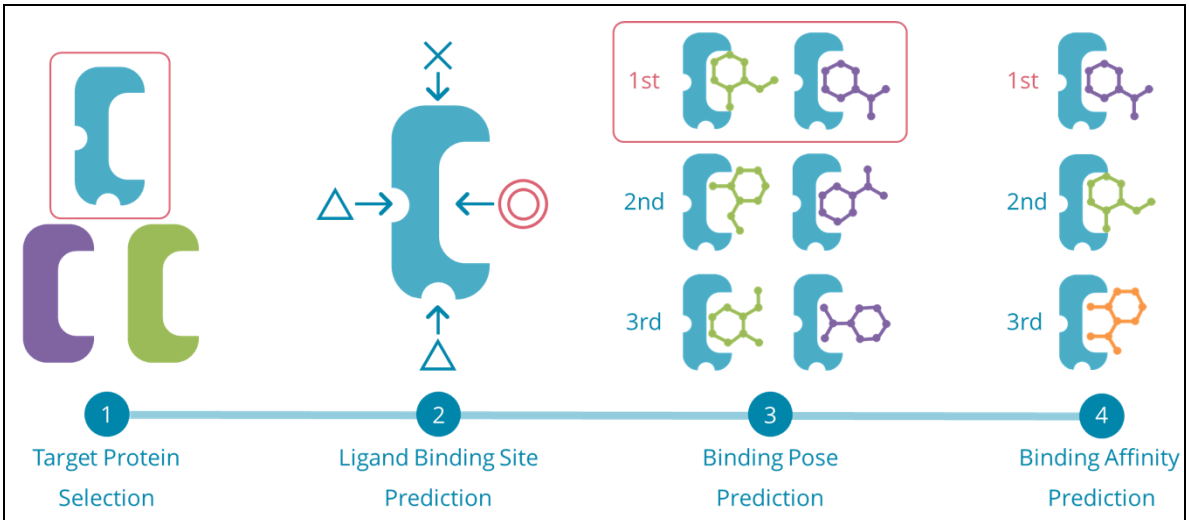


Figure 1. 1 Stages of virtual screening

A simple schematic representation of the stages of VS. The first step in VS begins with the selection of target proteins. The second step is to predict the ligand binding site to identify the pocket where the ligand will bind. Third, the ligand binding pose prediction step identifies the conformation in which the ligand binds to the protein. Finally, the binding affinity prediction compares the binding affinities between ligands.

1. Selection of the Target Protein:

Selecting the target protein is the initial step in virtual screening, and it's an important one that sets up the whole process. The goal here is to identify a protein that is critical to the relevant pathway and may be a target for treatment. A detailed examination of databases and literature is necessary to discover more about the composition, role, and connection between the protein and the illness. Selection is based on a protein's druggability, or the ability to change the protein's activity using a small chemical. Proteomic, metabolic, and genomic analyses are performed using sophisticated bioinformatics methods to validate the target's significance and drug-discovery potential.

2. Identification of the Ligand Binding Site:

Following the selection of the target protein, the next stage is to locate the protein's ligand binding site which is the area where ligand molecules can bind (Pérot, et al., 2010;

Zheng, et al., 2013). This stage aims to locate a protein's active site and offer recommendations for locating substances that bind to the target protein efficiently. When available, experimental data is used to support computational techniques like homology modeling and molecular docking in order to do this. Determining the binding site is essential to comprehending the interactions between ligands and the protein as well as forecasting the binding affinity of possible therapeutic candidates.

3. Determination of the Ligand Binding Pose (Docking):

The determination of the ligand binding pose refers to predicting the orientation and conformation of a ligand within the binding site of the target protein. This stage aims to model how a ligand fits into the binding site, which is essential for understanding the molecular basis of ligand-receptor interactions. Computational docking simulations are extensively used in this phase, where various poses of a ligand within the binding site are generated and evaluated based on their energy profiles. The goal is to identify the most stable poses that suggest the potential for strong interactions, thereby helping to design more effective drug molecules. The AQDnet in Chapter 3 excels especially in the performance of this phase.

4. Comparison of Binding Affinities among Ligands (Scoring):

The final stage of virtual screening involves the comparison of binding affinities among different ligands. The objective is to rank the screened compounds based on their predicted binding affinity to the target protein, thereby identifying the most promising drug candidates. This is typically accomplished through scoring functions that estimate the binding energy between the ligand and the target protein. Although physics-based scoring functions have been dominant for this stage in the past, machine learning-based scoring functions have been very successful in recent years, and a variety of machine learning-based scoring functions have been reported. The compounds are then ranked according to their scores. This comparative analysis facilitates the selection of top candidates for further experimental validation and optimization. The multi-shelled ECIF in Chapter 4 especially has superior performance in this phase.

In conclusion, each step of the virtual screening process is designed to incrementally refine the search for potential drug candidates by utilizing computational techniques to

model and predict interactions between small molecules and the target protein. This process allows for the efficient and cost-effective identification of promising compounds for further development in the drug discovery pipeline.

It is theoretically possible for a single method to solve the latter two stages simultaneously. However, in practice, many methods that excel in the docking task do not perform well in the scoring task, and conversely, many methods that excel in the scoring task have difficulty in performing docking. Referring to the CASF-2016 results, the top ranked method for the Docking power test are Physics-based methods. Many of their scoring power test results have a Pearson's $R < 0.6$, which is significantly lower than the results of recent state-of-the-art machine learning-based methods (Pearson's $R > 0.8$). On the other hand, many of the recent state-of-the-art machine learning-based methods do not report the results of the docking power test. This is presumably because many of the methods did not obtain favorable results in the docking power test. Many of the machine learning-based methods train their models using only the crystal structure and do not learn the energy difference from the most stable conformation, which is very important in docking. Therefore, it is theoretically difficult for them to evaluate Docking. For these reasons, although it is theoretically possible to perform Docking and Scoring with a single method, we expect that treating docking and scoring as different tasks will yield more favorable results.

In this doctoral dissertation, we focused specifically on the latter two stages of VS: binding pose prediction and binding affinity prediction.

Our reasons for targeting these stages are as follows:

1. The initial stages (target protein selection and binding site prediction) rely heavily on empirical experimental data.

2. The bottleneck in the drug discovery process lies in the latter stages.

In drug discovery, selecting target proteins is usually based on genetic analysis, animal model experiments, or literature reviews. Even if a highly accurate machine learning model predicts a protein's significance for a disease, its results are unlikely to take precedence over experimental findings.

Binding site prediction is typically done by experts using X-ray crystallography or cryo-electron microscopy to analyze protein-ligand complexes. Machine learning predictions, no matter how accurate, are rarely preferred over these methods.

However, binding pose prediction and binding affinity prediction are very costly and time-consuming. Despite finding many potential targets and binding sites, it's impossible to screen them all, making these stages bottlenecks in drug discovery. Therefore, replacing these phases with machine learning models could significantly benefit drug discovery. This study aims to develop superior methods for binding pose prediction and binding affinity prediction.

1.1.4 Protein-Ligand Binding Affinity Prediction

Protein-Ligand (P-L) binding affinity prediction is one of the most important VS techniques. It plays a crucial role in the phases of binding pose prediction (Docking), which predicts the relative position of the ligand to the protein, and the comparison of binding affinities among ligands (Scoring). P-L binding affinity prediction has traditionally been dominated by physics-based methods, which are computational techniques that utilize the principles of physics to analyze the interaction between protein and ligand.

However, a number of machine learning-based methods have been developed in recent years. This is due to the significant increase in the number of available 3D P-L binding structures and the increase in computational resources. In addition, the field of P-L binding affinity prediction has actively incorporated innovations in the area of computer science. Examples include Deep Neural Network (DNN), Convolutional Neural Network (CNN) and transformer and graph neural networks. These technologies have greatly improved the accuracy of P-L binding affinity prediction. However, there are still challenges in prediction accuracy and computation time for application to VS.

In the following, each phase of binding pose prediction (docking) and ligand-to-ligand binding affinity comparison (scoring) in VS is described respectively from the perspective of machine learning applications.

In binding pose prediction (Docking), physics-based methods have been the predominant method from the past to the present, with physics-based methods performing better than machine learning-based methods. To begin with, machine learning-based methods are rarely evaluated in benchmark datasets that assess docking performance (e.g. Comparative Assessment of Scoring Functions 2016 (CASF -2016) Docking power test). However, physics-based methods need longer computation times for higher accuracy methods. It is difficult to calculate all of the conformations generated from the huge number of compounds in the compound library by the physics-based method in a realistic amount of time in an actual VS. The AQDnet in Chapter 3 is a DNN that learns the results of quantum mechanics (QM) calculations and is a method with very good docking performance. Because AQDnet is a DNN, it can output prediction results extremely fast compared to physics-based methods, and at the time of reporting, it showed the best performance compared to existing methods in the CASF-2016 docking power test.

In comparison of binding affinities among ligands (Scoring), machine learning-based methods generally perform better than physics-based methods. In the CASF-2016 scoring power test, the recently reported machine learning-based method dominates the top ranks. Extended connectivity interaction features (ECIF) show very high performance (Pearson's R 0.866) in the CASF-2016 scoring power test. However, ECIF has the challenge that it represents the atoms on the ligand and protein side in great detail, but only considers the inter-atomic distance if the two atoms are located at 6 Å or not. In Chapter 4, we introduce the Multi-Shelled ECIF, which is a modification of ECIF to take into account inter-atomic distances and improve its performance. Multi-shelled ECIF achieved a Pearson's R of 0.877 in the CASF-2016 scoring power test, the best performance compared to existing methods at the time of reporting.

1.2. Contribution

In this dissertation, we introduced two novel studies in the field of P-L binding affinity prediction.

In our first study, we present our AI QM Docking Net (AQDnet) findings, a novel system predicting P-L binding affinity using their 3D structures. Our approach uniquely generates thousands of the ligand configurations and computes their binding energies by quantum chemistry computation. Incorporating atom-centered symmetry functions (ACSF), we trained a neural network to understand the P-L quantum energy landscape (P-L QEL), achieving a leading 92.6% success rate in CASF-2016 docking power.

In our second study, we present improvements to Extended Connectivity Interaction Features (ECIF) for predicting P-L binding affinity, particularly addressing its inability to account for interatomic distances. We developed two algorithms: multi-shelled ECIF, dividing atomic distances into layers, and weighted ECIF, assigning importance based on these distances. Our results show that multi-shelled ECIF surpasses both weighted ECIF and the original ECIF, achieving a Pearson correlation coefficient of 0.877 in CASF-2016 scoring power.

1.3. Organization

In Chapter 2, we briefly introduce several basic background knowledge about P-L binding affinity prediction and other techniques used in this thesis.

In Chapter 3, we present AQDnet, an advanced framework for P-L docking simulations, offering detailed insights into its methodology and capabilities. AQDnet predicts binding affinity using P-L 3D structures, expanding training datasets via ligand configurations and quantum chemistry computation. The model incorporates ACSF, learning the P-L quantum energy landscape, and excels in the CASF-2016 docking power, achieving a top 1 success rate of 92.6%.

In Chapter 4, we explore the advancement of Extended Connectivity Interaction Features (ECIF) in binding affinity prediction. Our research introduces multi-shelled ECIF and weighted ECIF, with the former notably enhancing the accuracy by segmenting interatomic distances into multiple layers. Multi-shelled ECIF demonstrates superior performance, evidenced by a Pearson correlation coefficient of 0.877 on the CASF-2016 benchmark, outperforming traditional ECIF. The results validate the potential of multi-shelled ECIF in refining drug discovery processes.

In Chapter 5, we provide a final summary of both studies and add some discussion of future research.

1.4. Use of AI Tools

DeepL and chatGPT4 are used for the purpose of English translation and proofreading of English texts.

Chapter 2

Preliminaries

This chapter provides an overview of the training dataset, benchmark dataset, machine learning techniques, and quantum mechanics calculations used in this thesis.

2.1. Protein-Ligand Complex Dataset

2.1.1. PDBbind

An extensive collection of experimental binding affinity data is represented by the PDBbind database, which is an invaluable tool for molecular recognition research. It was launched in 2004 and is updated every year to fill the void left by structural databases such as the Protein Data Bank (PDB) and the requirement for comprehensive binding data for a range of computational and statistical analysis in the field of bioinformatics (Liu, et al., 2015; Wang, et al., 2004).

Curating and making available experimental binding affinities of biomolecular complexes included in the PDB is the main objective of the PDBbind database. It covers a broad spectrum of complexes, including as interactions between proteins and ligands,

proteins and nucleic acids, and meets a variety of research goals, including drug discovery and the theoretical comprehension of molecular interactions.

PDBbind uniquely classifies complexes into general, refined, and core sets, catering to different levels of research specificity and quality requirements. While the general set offers broad coverage, the refined set provides curated collections of complexes with high-quality structural and binding data. The PDBbind core set is a small subset of the high-quality 285 complexes in PDBbind and is used as the Comparative Assessment of Scoring Functions 2016 (CASF -2016) benchmark, described below. This stratification enables users to select datasets tailored to their research, enhancing the utility of PDBbind across various computational and statistical analyses.

The database offers detailed binding data, including dissociation constants (Kd), inhibition constants (Ki), and IC50 values, across a wide spectrum of biomolecular complexes. Additionally, it includes "clean" structural files for protein-ligand (P-L) complexes, facilitating their use in molecular modeling software. The refined and core sets further refine the data quality for advanced studies in docking and scoring function validation.

In summary, PDBbind stands as a pivotal repository that not only enriches the PDB with valuable binding affinity data but also categorizes these into practical subsets. Its evolution over the years reflects a continuous effort to address the dynamic needs of the bioinformatics and computational biology community, underscoring its indispensable role in advancing research in molecular recognition and drug discovery.

2.2. Benchmark Dataset

2.2.1. CASF-2016

The Comparative Assessment of Scoring Functions (CASF) project offers a critical benchmarking tool for assessing how well different scoring functions function in the context of structure-based drug design (Cheng, et al., 2009; Li, et al., 2014; Li, et al., 2014; Su, et al., 2019). The main objective of CASF-2016 is to offer an objective, rigorous benchmark for evaluating scoring functions' effectiveness in terms of how well they predict

P-L binding affinities. A key component of contemporary drug discovery efforts is the prediction of P-L interactions, which is made possible by scoring functions. CASF-2016, the latest iteration, aims to provide an objective measure for the performance of various scoring functions by decoupling the scoring process from docking, hence allowing a more focused evaluation.

The one of key characteristics of CASF-2016 is to test scoring functions across four key aspects: scoring power (the ability to predict ligand binding affinities accurately), ranking power (the ability to correctly rank ligands by their binding affinities), docking power (the capacity to identify the correct ligand binding pose), and screening power (the proficiency in distinguishing active from inactive compounds) (Su, et al., 2019). This multifaceted approach ensures a comprehensive assessment of scoring functions, reflecting their applicability in real-world scenarios. By decoupling the scoring process from the docking procedure, CASF-2016 aims to isolate and directly assess the predictive power of scoring functions, thus facilitating a clearer understanding of their capabilities and limitations. Figure 2.1 and 2.2 show the evaluation methods of the docking power test and the screening power test, respectively.

CASF-2016 offers meticulously curated 285 P-L complexes, characterized by high-resolution crystallographic structures and reliable experimentally determined binding affinity data. This selection ensures the reliability and relevance of the data for accurate performance evaluation of scoring functions. The dataset is derived from the Protein Data Bank (PDB).

The dataset is not confined to crystal structures alone; it is augmented with decoy poses for comprehensive testing of docking and screening powers, which is vital for simulating realistic scenarios where false positives are a common challenge.

Furthermore, the dataset facilitates comparison between different scoring functions by providing evaluation results for established methods like AutoDock Vina across all four test metrics. This inclusion allows for direct, straightforward comparisons between new scoring functions and established benchmarks, simplifying the assessment of advancements or regressions in scoring function performance.

The CASF-2016 dataset is thus a significant resource for the computational chemistry community, offering a comprehensive suite of tools for the rigorous evaluation of scoring functions in drug discovery applications. CASF-2016 represents a significant advance in the benchmarking of scoring functions, offering a more robust and comprehensive tool for the assessment of these critical computational methods. By providing an open-access platform for evaluation, CASF-2016 encourages the development and refinement of scoring functions, ultimately contributing to the acceleration of drug discovery and the enhancement of therapeutic interventions.

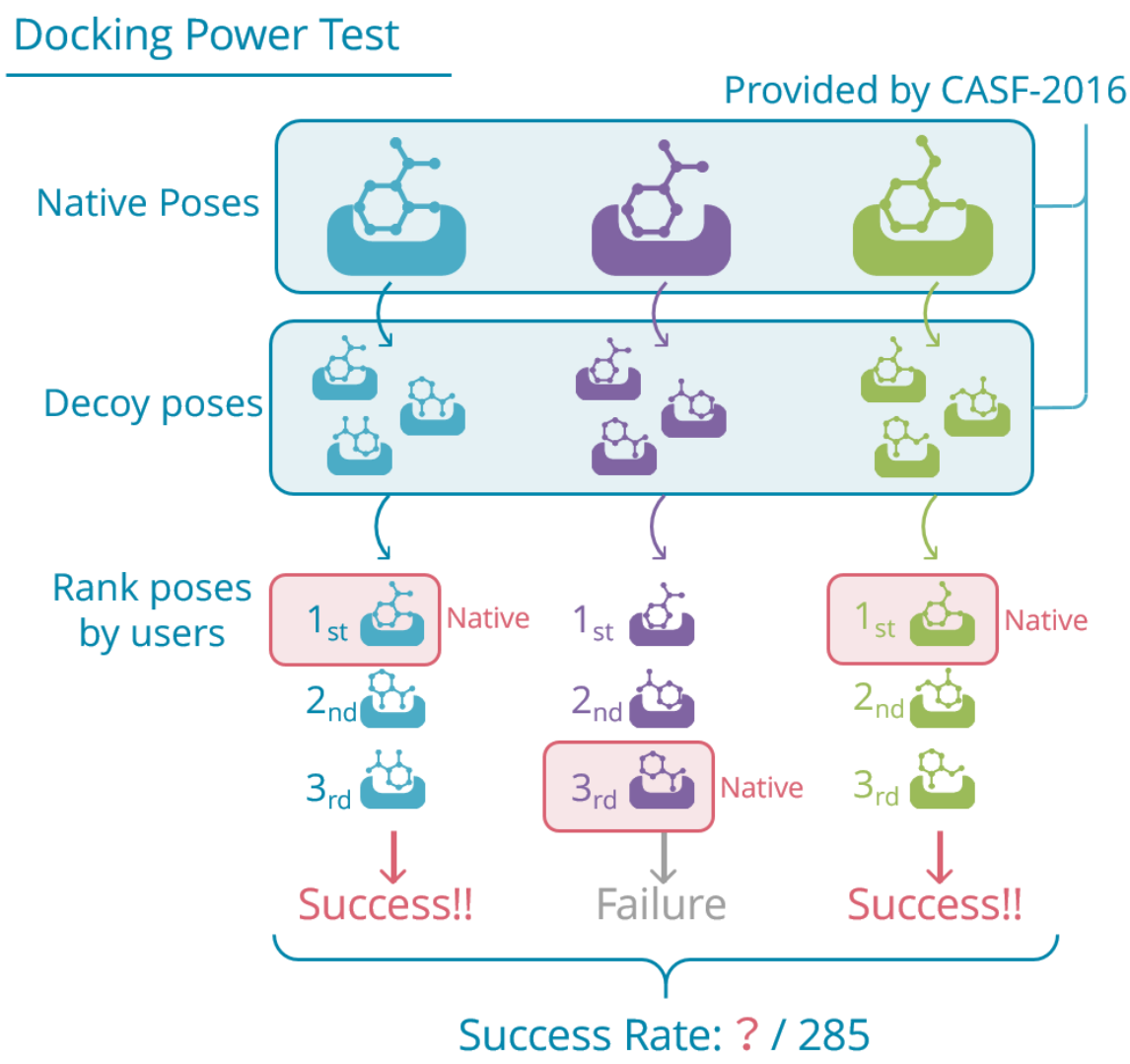


Figure 2. 1 Evaluation method by CASF-2016 docking power test

Schematic diagram of the docking power test evaluation method. First, up to 100 decoy poses are generated from the crystal structures of 285 P-L complexes. Next, the decoy poses and crystal structures of each complex are scored and ranked by a scoring function. If the difference in RMSD between the top ranked (1st~3rd) pose and the crystal structure is less than a predefined cutoff value (e.g., 2.0Å), the docking of that complex is considered successful. This evaluation is performed for all 285 complexes, and the overall success rate, i.e., how many of the 285 complexes were successful, is used as a quantitative measure of docking power.

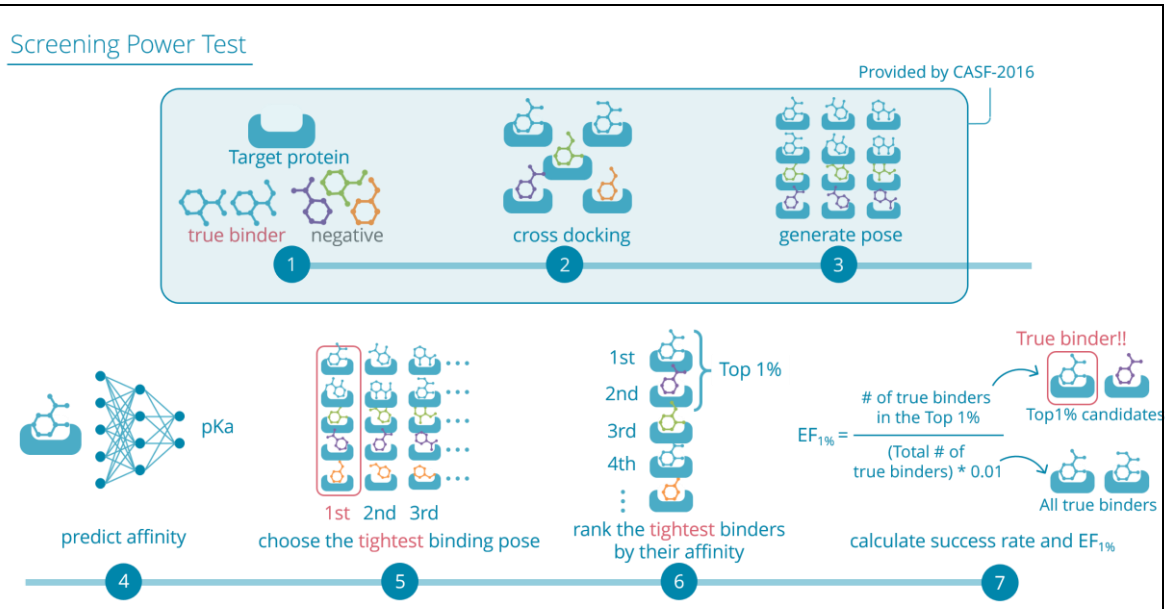


Figure 2. 2 Evaluation method by CASF-2016 screening power test

Schematic diagram of the screening power test evaluation method. First, the test set provided in CASF-2016 includes 57 target proteins, with 5 true binders for each protein. These 5 true binder ligands are defined as positive and the remaining 280 ligands are defined as negative. Second, all ligands are cross-docked against all 57 target proteins. Third, for each cross-docked P-L pair, a total of 100 decoy poses are created. This is all that is provided in CASF-2016. Fourth, all these binding poses are evaluated by a scoring function. Fifth, for each P-L pair, the pose with the best binding score is obtained as the tightest binding pose. The predicted score of this tightest binding pose is used as the score of the ligand for that target protein. Sixth, sort the 285 ligands for each target protein in descending order of score. Finally, the enhancement factor (EF) is calculated for each target protein, and the average EF of all 57 target proteins in the test set is defined as the screening power.

2.2.2. LIT-PCBA

The LIT-PCBA dataset emerges as a response to the critical need for unbiased and realistic datasets designed specifically for benchmarking VS and machine learning methodologies in drug discovery (Tran-Nguyen, et al., 2020). The effectiveness of various VS methodologies is assessed through benchmark studies, which necessitate comprehensive and unbiased datasets to accurately evaluate performance. Traditional datasets such as DUD (Huang, et al., 2006), DUD-E (Mysinger, et al., 2012), and MUV (Rohrer and Baumann, 2009) have been instrumental but are often criticized for potential biases, underscoring the need for more representative collections.

The primary objective of the LIT-PCBA dataset is to provide a rigorous and unbiased benchmarking platform that facilitates a comparative evaluation of VS methods. This dataset is meticulously curated to measure the performance of VS and machine learning approaches in identifying pharmacologically active compounds from a vast chemical space, thereby aiding in the identification of potential therapeutic candidates. The dataset is modeled to simulate real-world high-throughput screening libraries, enabling researchers to develop and test ML models that can accurately distinguish between biologically active and inactive compounds. This aspect is critical for ensuring that computational methods are relevant and applicable to practical drug discovery scenarios.

The LIT-PCBA dataset is derived from 149 dose-response bioassays available in the PubChem bioassays database. LIT-PCBA comprises molecules that have been experimentally confirmed to exhibit either significant activity or inactivity against specific biological targets. The dataset maintains an authentic balance between active and inactive compounds, mirroring the distribution typically encountered in genuine drug discovery campaigns. The compounds included in the LIT-PCBA dataset exhibit a broad range of chemical properties while avoiding biases towards particular molecular frameworks. This diversity ensures that ML models developed using this dataset are versatile and capable of generalizing across the vast chemical space relevant to drug discovery. This processing ensures the dataset's unbiased nature, making it a robust tool for evaluating the efficacy of VS techniques.

LIT-PCBA comprises 15 target sets featuring 7,844 experimentally confirmed active and 407,381 experimentally confirmed inactive compounds, offering a realistic representation of hit rates and potency distributions found in experimental screening. The chemical structures of active and inactive compounds are provided as SMILES (Weininger, 1988) strings, requiring users to generate three-dimensional coordinates. Additionally, for reference to the protein pocket, the dataset includes the crystal structures of target proteins co-crystallized with at least one ligand in MOL2 file format (Tripos, 2007).

In conclusion, the LIT-PCBA dataset marks a significant advancement in the field of computational drug discovery. LIT-PCBA represents a new generation of benchmarking datasets for VS and machine learning, specifically addressing the limitations of previous collections. By maintaining a rigorous standard for data inclusion and processing, LIT-PCBA enables the realistic evaluation of VS methods, facilitating the development and evaluation of more effective drug discovery tools and algorithms. Available for public use, it encourages the advancement of computational screening methods by providing a challenging yet fair testing ground.

2.3. Technologies Used in This Dissertation

2.3.1. Deep Neural Network

Deep neural network (DNN) is a type of machine learning model inspired by the neural circuits of the human brain. DNNs have a multi-layered network structure and have the ability to learn complex features and patterns in data. DNN is used in many fields such as image recognition, natural language processing, and speech recognition due to their ability to process large amounts of training data and their generalization performance (LeCun, et al., 2015; Weibo, et al., 2017).

One of the key advantages of DNNs is their ability to automatically extract interactions between features without the need for manual curation or prior knowledge. This capability allows features to be abstracted at various levels and learned from unstructured data that is difficult for humans to understand. In addition, DNNs have remarkable generalization

performance when used properly. This enables accurate predictions even for unseen data not included in the training set. These features facilitate the application of DNNs to a wide range of tasks, demonstrating their effectiveness.

DNNs are characterized by a multilayer architecture consisting of layers containing tens to thousands of nodes (also called neurons) each. Each node outputs a value obtained by applying an activation function to a linear combination of outputs from nodes in the previous layer. The coefficients in obtaining this linear combination are the training parameters of the DNN. The technical backbone of DNN is the updating of parameters by backpropagation. The backpropagation method is one of the main algorithms used to train neural networks to efficiently adjust the weights of the network and optimize the learning process. This method updates the weights of each node by calculating the error (loss) between the resulting output by the neural network and the true value and propagating the error backward through each layer of the network. There has also been extensive research on the architectural and theoretical aspects of DNNs. Various solutions have been proposed to the problems of trap to local minimum and gradient loss, which were considered problems when DNNs were first developed (He, et al., 2016; Zhang, et al., 2021).

The AQDnet study employs this structure, combining 19 small DNNs (subnetworks) into a comprehensive system. Each subnetwork incorporates the concept of residual learning so that it can learn in as many layers as possible without encountering the vanishing gradient problem.

For AQDnet, because we used features that describe atomic coordinates very precisely to represent energy, we needed a learning model that could directly utilize these features. Therefore, instead of decision tree-based models, which split values into two at each branch, we used neural networks.

2.3.2. Gradient Boosted Decision Tree

Gradient boosted decision tree (GBDT) is a powerful ensemble learning technique in the domain of machine learning, renowned for their effectiveness in both classification and regression tasks (Friedman, 2001; Natekin and Knoll, 2013). By integrating multiple weak

predictive models, typically decision trees, into a robust composite model, GBDTs improve prediction accuracy and performance. This section describes the characteristics of the GBDT and its underlying architecture and methodology.

In the following, we will explain the characteristics of the GBDT from two points of view: prediction accuracy and interpretability. First, in terms of prediction accuracy, the GBDT has a very strong performance, especially for structured data (e.g., tabular data). In machine learning competitions such as Kaggle, GBDT is included in many winning solutions when dealing with tabular data, making GBDT the model of choice over DNN. Regarding interpretability, the GBDT has the advantage that its components are decision trees, so it is easy to understand which features affect predictions. Compared to DNNs, which are often considered a black box where it is difficult to understand how the model is making a particular prediction, the interpretability of the GBDT is extremely valuable in situations where one wants to explain the model's prediction results, such as P-L binding affinity prediction.

Next, we discuss the architecture of the GBDT. The GBDT uses decision trees as its base learner, and multiple trees are constructed sequentially. Each subsequent decision tree learns incrementally from errors in the previous decision tree, employing gradient descent to minimize loss functions specific to tasks such as regression, classification, and ranking. Advanced variants of GBDTs such as XGBoost (Chen and Guestrin, 2016), LightGBM (Ke, et al., 2017), and CatBoost (Prokhorenkova, et al., 2018) introduce optimizations that enhance speed, accuracy, and functionality, including parallel processing, categorical variable handling without preprocessing, and improved regularization.

In the study of Multi-shelled ECIF, GBDT is utilized as the modeling framework, demonstrating exceptional predictive accuracy. Additionally, an analysis of feature importance is conducted, offering valuable insights into identifying critical interactions for predicting P-L binding affinity.

Multi-shelled ECIF uses GBDT to compare with previous studies. The purpose of the Multi-shelled ECIF research is to compare features modified to consider inter-atomic distances with the original ECIF features. If we used neural networks as the learning model and the performance improved, it could be argued that the improvement was due to

switching to neural networks rather than considering distances. To avoid this, Multi-shelled ECIF uses GBDT.

2.3.3. Quantum Mechanics Calculations

The integration of quantum mechanical (QM) calculations into the realm of structure-based drug design, particularly in P-L systems, has been an extremely important recent advance (Cavasotto, et al., 2018). Traditional methods often suffer from a trade-off between speed and accuracy, limiting their ability to adequately model complex molecular interactions. Recent advances, such as the development of the "SQM/COSMO" scoring function, are promising approaches to overcome these limitations and improve predictive power and reliability in diverse P-L complexes (Ajani, et al., 2017; Pecina, et al., 2017; Pecina, et al., 2016).

The main goal of using QM calculations, especially semi-empirical quantum mechanics (SQM) methods such as PM6-D3H4X and DFTB3-D3H4, is to more precisely and accurately describe P-L interactions. The information obtained by these sophisticated methods is essential for understanding the interaction mechanism, allowing more accurate identification of native ligand poses and affinity prediction. Ultimately, it will influence drug design and the drug discovery process, leading to improved reliability of the VS and scoring process.

The SQM/COSMO scoring functions and associated QM approaches stand out in their ability to effectively capture important quantum mechanical effects such as charge transfer, polarization, and dispersion. These effects are generally ignored in classical force field models. By combining semi-empirical methods with empirical corrections and the COSMO model for implicit solvation, these methods provide a comprehensive view of the non-covalent interactions within PL complexes. This approach allows for application to a wide range of chemical spaces and does not require system-specific parameterization.

The change in the Gibbs free energy change in P-L binding is expressed by the following equation.

$$\Delta G_{binding} = \Delta E_{int} + \Delta\Delta G_{solv} + \Delta G_{conf} - T\Delta S \quad (1)$$

where ΔE_{int}: the gas-phase interaction energy, $\Delta\Delta G_{solv}$: the solvation energy change upon complex formation, ΔG_{conf}: the change of conformational free energy, $-T\Delta S$: the entropy change upon binding. The SQM/COSMO scoring function focuses on two main elements: calculation of gas-phase interaction energy (ΔE_{int}) and evaluation of solvation energy ($\Delta\Delta G_{solv}$). In the study of the AQDnet, we calculated ΔE_{int} by PM6 level with D3H4X correction and $\Delta\Delta G_{solv}$ by COSMO. The approach is tailored to meet the requirements of a wide range of VS applications and provides a balance between computational efficiency and the level of detail required for accurate predictions.

Advances in QM calculations, including the SQM/COSMO scoring function, will allow for a detailed and accurate depiction of molecular interactions beyond conventional limitations. The results will greatly contribute to the development of more effective and precisely targeted therapeutics and provide a major breakthrough in structure-based drug design and VS. This innovative approach highlights the important role of QM calculations in modern pharmaceutical research and points the way toward more reliable and comprehensive VS methods. In the AQDnet study, SQM/COSMO was used to evaluate the stability of the P-L complex, and the results were used to extend the data and train the DNN, allowing the DNN to learn the P-L quantum energy landscape (P-L QEL).

Chapter 3

AQDnet: Deep Neural Network for Protein-Ligand Docking Simulation

3.1. Background

Virtual screening (VS) is a computational approach that facilitates the identification of bioactive compounds that bind to a specific target protein from an extensive library of compounds. This method can significantly expedite the drug discovery process, reduce the expenses, the time and effort required to evaluate compounds in assays (Gimeno, et al., 2019). Recently, various attempts have been made to leverage the achievements of computer vision and natural language processing technologies, such as convolutional neural networks (Krizhevsky, et al., 2012) and transformers (Vaswani, et al., 2017), for VS. One of the key objectives of VS is to predict the binding affinity of protein-ligand (P-L) interactions. Several virtual screening techniques have been proposed, based on physicochemical calculations and machine learning (Feinberg, et al., 2018; Jones, et al., 2021; Lee, et al., 2019; Ozturk, et al., 2018; Ragoza, et al., 2017; Torng and Altman, 2019). Recently, methods based on deep learning have shown remarkable success in this field. Nonetheless, the accuracy of binding affinity prediction by deep learning models remains

inadequate and demands significant improvement. The cause of this inadequacy lies in the absence of an established learning methodology and an insufficient number of training samples.

Behler discussed the three essential requirements for artificial intelligence (AI) to predict the potential energy, including invariant predictions for system rotation and translation, invariant predictions regardless of the atom processing order, and a unique representation of the three-dimensional molecular structure (Behler, 2015). Various approaches have been developed to represent the molecular structure as graphs (Jiang, et al., 2021; Torng and Altman, 2019), but they may not effectively use the exact relative positions of numerous atoms. Another approach is to use 3D convolutional neural networks (3DCNNs) to process molecular structures as three-dimensional images to predict binding affinity. (Ragoza, et al., 2017) However, 3DCNNs are not invariant to molecule rotation and translation and may neglect energy changes due to slight differences in interatomic distances. Various attempts have been made to predict binding affinity by processing the ligand and protein separately (Lee, et al., 2019; Ozturk, et al., 2018). However, these approaches fail to consider the intricate three-dimensional structure of the ligand-protein complex. OnionNet (Wang, et al., 2021; Zheng, et al., 2019) is an invariant method to system rotation and translation and the order of atoms processed, which predicts binding affinity by counting the number of the elements' contacts that exists at a particular distance as a feature. However, this method uses shells in 0.5 Å increments to determine distances and cannot recognize differences in atomic coordinates smaller than 0.5 Å.

The Atom Centered Symmetry Function (ACSF) (Behler and Parrinello, 2007; Gao, et al., 2020; Smith, et al., 2017) satisfies all three of the above requirements and has proven effective in predicting molecular energies. However, while ACSF has been successfully applied to single-molecule systems, it cannot be directly employed for predicting P-L binding affinity. Furthermore, some of the reported methods for the application of ACSF to predict P-L bond affinity do not fully utilize the information of protein-side atoms. Nonetheless, ACSF possesses two desirable features that are useful in predicting P-L bond affinity: accounting for three-body interactions and generating high-resolution features that can even detect small differences in atomic coordinates.

ACSF is a method developed to replace the Schrodinger equation with neural networks. The relevance of this is discussed below. Quantum mechanics and molecular mechanics methods have traditionally been used to calculate molecular energies. It is supposed that more accurate results are obtained by Quantum mechanics, in particular by solving the many-body Schrödinger equation (MBSE). However, solving MBSE is not possible with current computers, so several alternatives have been proposed. These include DFT (Kohn and Sham, 1965) and Coupled Cluster (Bartlett and Musiał, 2007). These methods were developed to obtain the approximation of the MBSE solution in realistic computation time. ANI, a neural network trained with ACSF as a feature, is also a method developed to obtain the approximation of the solution of MBSE in the same way as DFT and Coupled Cluster, specifically by training a neural network with the calculation results of DFT for a single molecule system. Based on these considerations, ANI, a neural network trained with ACSF as a feature, can be considered to be the approximation of MBSE approximation. Although the ACSF is not a direct representation of MBSE, it contains enough information in its features, such as interatomic distances and many-body interactions, to be used in DFT calculations. It is also the fact that the ACSF can reproduce the results of DFT calculations with sufficient accuracy, and is therefore considered to have a certain validity as an alternative to MBSE.

The majority of current methodologies for predicting the activity of ligand-protein complexes rely solely on information regarding two-body interactions (Wang, et al., 2021; Zheng, et al., 2019). However, P-L intermolecular interactions are essentially many-body interactions. Although many-body interactions are known to play a significant role in predicting the physical properties of chemical compounds, they have rarely been taken into account when predicting binding affinity. Therefore, it is necessary to establish a method to utilize the information of not only two-body interactions but also three- or more-body interactions for the prediction of binding affinity. ACSF extracts features by utilizing information pertaining to the distances between three atoms and the angles they form. This enables the model to account for the interactions between the three bodies, making it highly advantageous in predicting P-L bond affinity.

Binding affinity predictions are applied in tight collaboration with molecular docking programs. However, many of these prediction systems rely on other docking programs to

predict the most stable conformation, and only a few are capable of evaluating P-L docking. Those systems only predict the binding affinity using the most stable conformation predicted by other docking programs. It is thus imperative to develop a method that can predict both the most stable conformation and the binding affinity of P-L complexes, as few machine learning systems currently exist that can accomplish this task.

One of the reasons why evaluating the stability of P-L complexes is challenging is because differences in binding stability must be predicted from small differences in atomic coordinates. Previously reported machine-learning methods for predicting binding affinity are based on grids of 1 Å increments (Stepniewska-Dziubinska, et al., 2018) or shells of 0.5 Å increments (Wang, et al., 2021; Zheng, et al., 2019), and are unable to recognize small differences in atomic coordinates. As a result, these method are unable to evaluate P-L docking. In this respect, ACSFs can generate high-resolution features that take into account the slightest difference in atomic coordinates because the distance between two atoms is represented by the outputs of the multiple Gaussian functions. Thus, the use of ACSF in predicting P-L binding affinity allows us to evaluate the stability of the P-L complex and even predict the most stable conformation.

In databases such as PDBbind (Wang, et al., 2004), which are commonly employed as training datasets, approximately 20,000 combinations of ligand-protein complex structures and binding affinities are archived. In contrast, while more than 1.5 million training samples are commonly used for image recognition (Kuznetsova, et al., 2020; Russakovsky, et al., 2015) a meager number of samples are available for training in the domain of binding affinity prediction. Thus, there is an urgent need to increase the number of training samples and develop methods for data augmentation.

Previous methods for predicting the binding affinity of P-L complexes using machine learning have basically used only crystal structures registered in databases as training data (Jimenez, et al., 2018; Wang, et al., 2021; Zheng, et al., 2019). However, as mentioned before, the number of complexes registered in the database is limited. This makes it difficult to prepare a sufficient amount of training data by using crystal structures alone. In addition, the VS needs to evaluate not only the crystal structure but also the configurations of the transition process before reaching the most stable configuration. For this reason, it

is not appropriate to train the model for virtual screening only on the most stable configurations. In this study, we propose a new data augmentation method that generates a large number of configurations based on a single crystal structure registered in a database. The challenge here is how to label the generated configurations. To address this challenge, we have devised a method of estimating the change of stability of each generated configuration compared to that of the most stable pose using quantum chemical calculation. By using this method, we were able to generate 900 ~ 1,000 configurations for a single crystal structure and successfully expand the training dataset by labeling each of them.

In labeling the configurations generated by the above method, we used semi-empirical quantum mechanics (SQM)/COSMO to estimate the change of stability from the most stable pose. This method is a scoring function that combines a quantitative SQM description of various non-covalent interactions with an implicit COSMO solvation approach (Ajani, et al., 2017; Pecina, et al., 2017; Pecina, et al., 2016). This is an extremely accurate method of predicting the most stable conformation in P-L docking. In the P-L docking task, where an RMSD of 2 Å or greater from the native binding pose is the criterion for false positives, SQM/COSMO showed a substantially lower number of false positives than classical SFs such as AutoDock Vina and Glide. These outcomes suggest that SQM/COSMO is adept at correctly identifying the native binding pose among decoys for each P-L system.

The energy difference between the generated configurations and the most stable poses estimated by SQM/COSMO was used for labeling. This approach not only serves as a data augmentation method but also as a novel approach to train the model using the P-L quantum energy landscape (P-L QEL) dataset calculated by SQM/COSMO. This study is novel in two ways. The first is the development of a method that applies ACSF to the prediction of P-L binding affinity prediction. Second, quantum mechanics (QM) simulations were used to extend the data and overcome the lack of training data. These two points have enabled us to successfully learn the P-L QEL. As a result, our model was superior to all others evaluated on the Comparative Assessment of Scoring Functions 2016 (CASF -2016) docking power benchmark (Su, et al., 2019).

.

3.2. Methods

3.2.1. Feature Extraction of Protein-Ligand Complexes

Feature extraction was conducted on the three-dimensional structure of the ligand-protein complex. MDTraj (Robert , et al., 2015) was used to read the PDB files and calculate the interatomic distances. The Gaussian function calculation and other processes were implemented using NumPy. The three-dimensional structures were prepared using PDB files that contain both ligand and protein information. The feature extraction method used in this study consisted of two main parts: the radial part, which contains information on the distance between two atoms, and the angular part, which contains information on the distance and angle between three atoms. These two parts are explained below. Figure 3.1 shows a graphical representation of the features.

The reason for using discretized distances with Gaussian basis functions in AQDnet is to treat one complex as one fixed-length vector. AQDnet is a method to characterize protein-ligand interactions on an atomic scale. Protein-ligand interactions are composed of many atomic interactions, and the number of atomic interactions varies greatly depending on the number of atoms in the protein and ligand. If all proteins and ligands had the same number of atoms, it would be possible to generate a fixed-length vector, but in reality, even ligands do not have a fixed number of atoms, so it is not possible to create a fixed-length vector using the atomic interaction distance of each individual atom as a feature. Therefore, instead of directly using the distance between atoms as a feature, the feature of a fixed-length vector is generated by discretizing it. By discretizing with multiple Gaussian functions, the original inter-atomic distance can be estimated to some extent even after the discretization. This device enables the generation of fixed-length vector features while approaching the expressiveness of features that directly incorporate the interatomic distances of individual atoms.

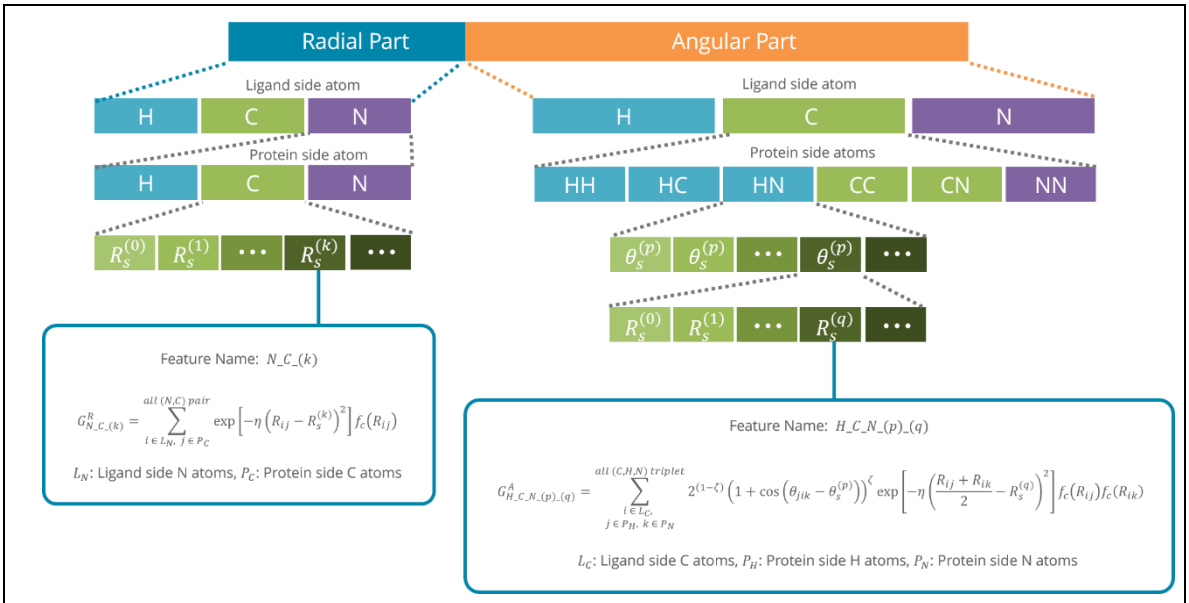


Figure 3. 1. A graphical representation of the AQDnet features.

The features of AQDnet consist of two parts: the radial part and the angular part. In the radial part, each column is defined by a combination of a Ligand atom element, a Protein atom element, and the R_s parameter. In the angular part, each column is defined by a combination of a Ligand atom element, two Protein atom elements, the R_s parameter and the θ_s parameter. In this example, the formula for the radial part with N on the ligand side, C on the protein side, and R_s on the kth value (left side of the panel) and the formula for the angular part with C on the ligand side, H and N on the protein side, θ_s on the p-th and R_s on the q-th value (right side of the panel) are shown.

3.2.1.1 Radial Part

The radial part uses the interatomic distance between one atom on the ligand side and one atom on the protein side as the input of the multiple Gaussian functions, with the output vector being the feature value of the radial part. The feature extraction process is as follows. First, the distance between all ligand atoms and protein atoms is calculated. Next, all pairs with distances below a specified threshold (R_c) are obtained. For each pair, the elements of the ligand atom and the protein atom are retrieved, and the radial symmetry function is calculated using the interatomic distance of the two atoms (R_{ij}) as input.

$$f(R_{ij}) = exp\left[-\eta\left(R_{ij} - R_s^{(k)}\right)^2\right] f_c(R_{ij}) \tag{2}$$

$$f_c(R) = \begin{cases} \frac{1}{2}\left[cos\left(\frac{\pi R}{R_c}\right) + 1\right] & for\ R \leq R_c \\ 0 & for\ R > R_c \end{cases} \tag{3}$$

The cutoff function, f_c includes R_c as the cutoff radius—a hyperparameter that determines which atoms are considered based on their distance. R_s is a hyperparameter that adjusts the peak of the Gaussian functions. In this process, multiple R_s are used to output multi-dimensional features for each pair. Finally, features with the same combinations of ligand-side element, protein element, and R_s are summed to produce the final feature as shown below.

$$G^R_{E_l\_E_p\_k} = \sum_{i \in L_{E_l},\ j \in P_{E_p}}^{all\ (E_l, E_p)\ pairs} exp\left[-\eta\left(R_{ij} - R_s^{(k)}\right)^2\right] f_c(R_{ij}) \tag{4}$$

Here E_l is the element type of the ligand-side atom, E_p is the element type of the protein-side atom, L_{E_l} is the set of all the E_l atoms on the ligand side, P_{E_p} is the set of all the E_p atoms on the protein side and $R_s^{(k)}$ is the k-th R_s. The name of the feature with E_l for the ligand-side atom, E_p for the protein-side atom, and $R_s^{(k)}$ for R_s is defined as $E_l\_E_p\_k$. For example, the feature N_C_3 is characterized by a combination of a ligand-side nitrogen atom, a protein-side carbon atom, and R_s as $R_s^{(3)}$. As a result, a feature vector with $(N^2_{element} \times N_{R_s})$ dimensions is obtained, where $N_{element}$ is the number of elemental species considered, and N_{R_s} is the number of R_s values.

3.2.1.2 Angular Part

The feature value of the angular part is generated using the interatomic distance between a ligand-side atom and two protein-side atoms, as well as the angle between the three atoms. The feature extraction procedure is as follows. First, the distance between all ligand atoms and protein atoms is calculated. Next, all combinations of two protein-side atoms with distances less than a certain threshold (R_c) from any ligand-side atom are obtained. For the resulting triplet of one ligand-side atom and two protein-side atoms, the interatomic distances R_{ij} and R_{ik}, the angle between the three atoms θ_{jik}, and the element type of each atom are obtained. For each triplet, the following angular symmetry function with R_{ij}, R_{ik}, and θ_{jik} as inputs is calculated.

$$f\left(R_{ij}, R_{ik}, \theta_{jik}\right) = 2^{(1-\zeta)}\left(1 + \cos\left(\theta_{jik} - \theta_s^{(p)}\right)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij} + R_{ik}}{2} - R_s^{(q)}\right)^2\right] f_c\left(R_{ij}\right) f_c\left(R_{ik}\right) \quad (5)$$

f_c is the same cutoff function as that of the Radial part. R_s is a hyperparameter that adjusts the peak of the Gaussian function, and θ_s is a hyperparameter that modifies the phase of the Cosine function. In this process, multiple R_s and θ_s are used to output multi-dimensional features for each triplet. Finally, the features with the same combination of ligand side element, protein elements, R_s, and θ_s are summed up to get the final feature value as shown below.

$$G^A_{E_{p1}\_E_l\_E_{p2}\_p\_q} = \sum_{\substack{all\ (E_{p1}, E_l, E_{p2})\ triplets \\ i \in L_{E_l}, \\ j \in P_{E_{p1}},\ k \in P_{E_{p2}}}} 2^{(1-\zeta)}\left(1 + cos\left(\theta_{jik} - \theta_s^{(p)}\right)\right)^{\zeta} \times exp\left[-\eta\left(\frac{R_{ij} + R_{ik}}{2} - R_s^{(q)}\right)^2\right] f_c\left(R_{ij}\right) f_c\left(R_{ik}\right) \quad (6)$$

Here E_l represents the element type of the ligand-side atom, E_{p1} and E_{p2} are the element types of the protein-side atoms. L_{E_l} is the set of all the E_l atoms on the ligand side, $P_{E_{p1}}$ and $P_{E_{p2}}$ are the sets of all the E_{p1} atoms on the protein side and all the E_{p2} atoms on the protein side, respectively. $\theta_s^{(p)}$ is the p-th θ_s value and $R_s^{(q)}$ is the q-th R_s value. The feature comprising E_l for the ligand-side atom, E_{p1} and E_{p2} for the protein-side atoms,

$\theta_s^{(p)}$ for θ_s and $R_s^{(q)}$ for R_s is named as $E_{p1}\_E_l\_E_{p2}\_p\_q$. For example, the feature where the ligand-side atom is carbon, the protein-side atoms are hydrogen and nitrogen, θ_s is $\theta_s^{(1)}$, and R_s is $R_s^{(2)}$ is defined as $H\_C\_N\_1\_2$. As a result, we get a feature vector with $(\frac{1}{2}N_{element}(N_{element}+1)\,N_{R_s}N_{\theta_s})$ dimensions, where $N_{element}$ is the number of elemental species considered, N_{R_s} is the number of R_s values, and N_{θ_s} is the number of θ_s values.

3.2.1.3 Exporting Features to Files

For memory efficiency reasons, the features are exported as TFRecords files.

3.2.1.4 Parameters

For feature extraction, we targeted seven elements for feature extraction, H, C, N, O, P, S, Cl, and Zn, while the remaining elements were collectively represented as Dummy (Du). Consequently, eight element types, H, C, N, O, P, S, Cl, Zn, and Du, were used for E_l and E_p above. The R_c parameters were set to 12 Å for radial part and 6 Å for angular parts. The R_s parameters for the Radial part are [0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 10.5, 11.0, 11.5]. Regarding the angular part, the R_s parameters are [0.5, 2.5, 4.5] and the θ_s parameters are a sequence of numbers from 0 to 2π divided into 8 equal segments [0.0, 0.785, 1. 570, 2.356, 3.142, 3.927, 4.7124, 5.4978].

3.2.2. Dataset Preparation

3.2.2.1 Quantum Docking Simulation

P-L complex structures were obtained from the PDBbind database (Liu, et al., 2015; Wang, et al., 2004). For each complex, hydrogen atoms were added using the Protonate

3D module in the Molecular Operating Environment (MOE) suite (ChemicalComputingGroupULC, 2023) (https://www.chemcomp.com), and the structural data of the complex were intensively augmented by performing the docking simulations as follows. The partial charges of each atom were assigned by mmff94x force field. Over 1,000 configurations were generated from each complex by Smina (Koes, et al., 2013), which is a fork of AutoDock Vina (Trott and Olson, 2010). The configurations were energy-minimized using Amber 18 (http://ambermd.org/) with ff99SBildn force field (Lindorff-Larsen, et al., 2010) for proteins, the second generation of the general AMBER force field (GAFF2) (He, et al., 2020) for ligands. With protein coordinates fixed, the energy of the ligand was minimized. The energy of each configuration was calculated with PM6-D3H4X/COSMO in MOPAC (http://openmopac.net). In order to improve the calculation accuracy, solvation energy was compensated by the effective surface tension coefficient ξ = 0.046,(Kříž and Řezáč, 2019) which was also coupled with our correction scheme depending on the P-L QEL (more detailed descriptions are provided elsewhere).

3.2.2.2 Labeling of the Generated Configurations

In this study, experimentally obtained values are represented as variables without dashes (e.g., pK_a), while values corrected by ΔE are represented as variables with dashes (e.g., pK_a'). From the experimentally measured binding affinity (K_d), ΔG is calculated using Equation 5, where R is the gas constant and T is the temperature.

$$\Delta G = -RT \ln K_d \tag{7}$$

The lowest energy conformation generated from a single ligand is defined as the reference configuration for that ligand. For other configurations, the difference in energy (ΔE) between each configuration and the reference configuration was calculated. The corrected energy label of each configuration, ΔG', is then calculated using Equation 6.

$$\Delta G' = \Delta G + \Delta E \tag{8}$$

The configurations except for the reference configuration were labeled with $\Delta G'$ while the reference configuration was labeled with ΔG. If necessary, pseudo-K_d ($K_d{}'$) and pseudo-pK_a ($pK_a{}'$) corrected by ΔE are obtained using the following equations.

$$\Delta G' = -RT \ln K_d{}' \tag{9}$$

$$pK_a' = -\log_{10} K_d' \tag{10}$$

3.2.2.3 PDB Preparation of CASF-2016 Dataset

Ligand Mol2 files were converted to PDB files by Open Babel. Subsequently, protein PDB files and ligand PDB file were merged to generate complex PDB files. The -log Kd/Ki registered in PDBbind was converted to ΔG using Equation 5 and used as the correct label. In both docking decoy and screening decoy, one ligand Mol2 file contains multiple ligand structures. Therefore, we parsed each structure and created PDB files containing one ligand structure per file. These parsed ligand PDB files were merged with the corresponding protein PDB file to establish the complex PDB file.

3.2.2.4 PDB Preparation of LIT-PCBA Dataset

The smi file of ligand was loaded using RDkit (https://www.rdkit.org) to generate 3D conformation and add hydrogens. The files were then saved as sdf files. All template PDB files were protonated by MOE (ChemicalComputingGroupULC, 2023) (https://www.chemcomp.com). Then, docking of ligand to protein was performed using gnina36. If multiple templates were given in the target, docking was performed on all templates. The docked ligands were then saved as sdf files. Ligand Mol2 files were converted to PDB files by Open Babel. Subsequently, protein PDB files and ligand PDB file were merged to generate complex PDB files.

3.2.3 Evaluation Method

3.2.3.1 Evaluation by CASF-2016

We evaluated scoring power, docking power and screening power of CASF-2016 in order to facilitate comparison with existing methods. Features were extracted from PDB files generated by the method described in PDB preparation of CASF-2016 dataset. The predicted values were formed using the method described in the CASF-2016 update study (Su, et al., 2019), and the docking power, screening power, and scoring power were evaluated. The docking AQDnet, which is a model specialized for the evaluation of docking power, was used for the evaluation of docking power and screening power. Similarly, the scoring AQDnet was used to evaluate scoring power. The only difference between the docking AQDnet and the scoring AQDnet is the training data. The structure and hyperparameter of these two models remain the same. The differences between the two models are discussed in Section 3.2.3.

3.2.3.2 Data Splitting

The data was divided into test set, valid set, and train set with reference to OnionNet and Pufnucy's method. We carried out data augmentation on 1,223 complexes, excluding those in the CASF-2016 and CASF-2013 core sets. During the data expansion process, each complex generated approximately 5,000 conformations. We divided the complexes into trainset, validation set, and test set in the following manner. The trainset for the docking AQDnet consisted of 1,123 complexes from the augmented data. For the scoring AQDnet, we included 16,306 crystal structures that were not present in the validation set, CASF-2016 core set, or CASF-2013 core set. However, we did not add any crystal structures to the docking AQDnet's trainset. The validation set for the docking AQDnet was composed of 100 randomly-selected complexes from the initial 1,223 complexes. Additionally, the scoring AQDnet's valid set contained crystal structures of these 100 complexes and 900 other randomly chosen complexes, totaling 1,000 complexes. Lastly, the test set included the CASF-2016 core set and the CASF-2013 core set. The summary of the aforementioned divisions is presented in Table 3.1. The PDBIDs of the complexes used for the test set,

validation set and training set for both the docking-specific and scoring AQDnet are listed in supporting information of the published paper (link).

3.2.3.3 Data Filtering

We defined ΔE for each configuration as the difference between the minimum energy among the configurations generated for each complex and the energy of the corresponding configuration. For example, if configurations A, B, and C have energies of -7, -5, and -4, respectively, the ΔE values for configurations A, B, and C would be 0, 2, and 3, as configuration A possesses the minimum energy among the three. Given the impracticality of utilizing all generated configurations for training from a computational standpoint, RMSD from each crystal structure and ΔE were used to filter the data. Different filtering criteria were used for the training data of the docking-specific and scoring AQDnet. For the training and validation sets of the docking AQDnet, we used conformations with ΔE less than 30 kcal/mol and RMSD less than 2.5 Å. For the docking AQDnet's training and validation sets, conformations with ΔE values below 30 kcal/mol and RMSD values under 2.5 Å were utilized.

For the scoring AQDnet's training and validation sets, conformations with ΔE values below 2 kcal/mol and RMSD values under 2.0 Å were employed. The validation set for the docking AQDnet contains 100 PDBIDs and a total of 89,740 configurations. The training set for the docking AQDnet comprises 1,123 PDBIDs and a total of 940,038 configurations. The validation set for the scoring AQDnet consists of 1,000 PDBIDs and 20,995 total configurations. The training set for the scoring AQDnet includes 16,306 PDBIDs and 247,393 total configurations.

Table 3. 1 The differences between the training data of the docking AQDnet and that of the scoring AQDnet.

	Docking AQDnet	Scoring AQDnet
Number of expanded complexes (PDBIDs) in validation set	100	100
Number of expanded complexes (PDBIDs) in training set	1,123	1,123
Crystal structures added	No	Yes
Number of crystal structures in validation set	0	1,000
Number of crystal structures in training set	0	16306
Energy filtering	< 30 kcal/mol	< 2 kcal/mol
RMSD filtering	< 2.5 Å	< 2.0 Å
Number of configurations in validation set	89,740	20,995
Number of configurations in training set	940,038	247,393

3.2.4 Neural Network Model

3.2.4.1 Architecture

All of the following models were built and trained in TensorFlow 2.3.2. The architecture of the deep neural network (DNN) is presented in Figure 3.2. The DNN model employed in this project consists of 18 sub DNNs, which process the radial or angular features of each corresponding element, and one output DNN that summarizes the outputs of the 18 sub DNNs. Specifically, 9 of the sub DNNs process radial features and correspond to different elements (H, C, N, O, P, S, Cl, Zn, and Dummy), responsible for processing the features when the atom on the ligand side is the target element. The remaining 9 sub-DNNs process angular features and are responsible for processing features when the atom on the ligand side is the target element, in a similar manner to the sub-DNNs that process radial features above. All 18 sub DNNs share the same structure, and their details are described below.

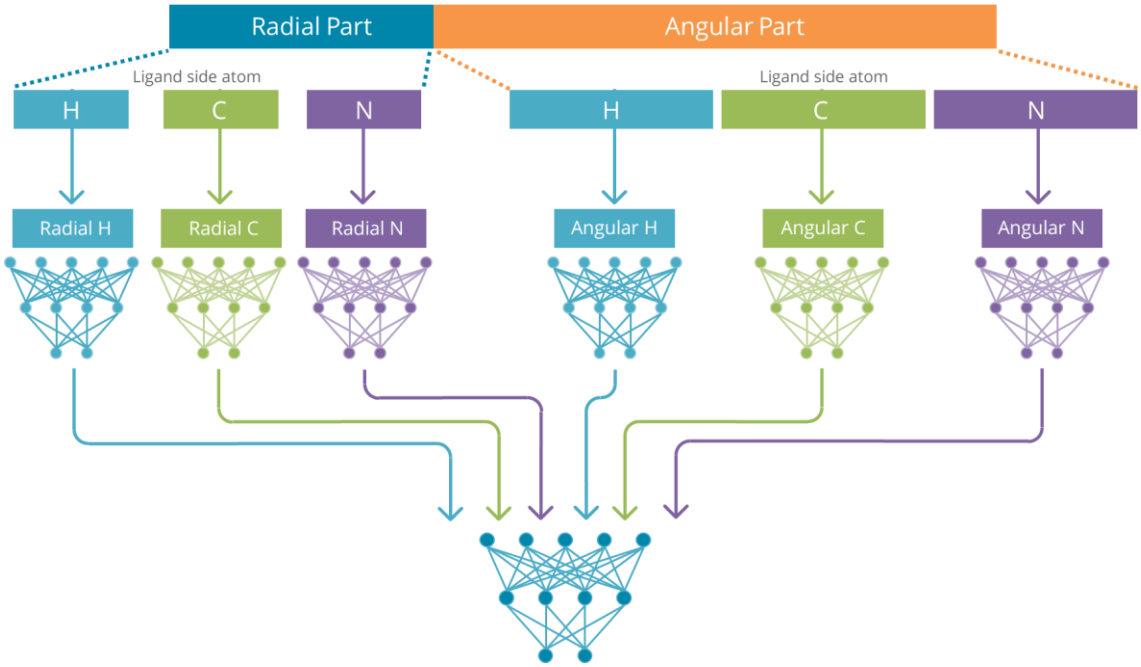


Figure 3. 2. A simplified schematic of the structure of the deep neural network of AQDnet.

The AQDnet features consist of two parts: radial part and angular part. They are further divided by ligand elements, each of which is an input to the corresponding DNN. Each DNN outputs a 10-dimensional tensor, all of which are combined to form the input of the Output DNN, which produces the final output.

3.2.4.2 Sub DNN Structure

A dropout layer is used after the input layer with a dropout rate of 0.05. Each DNN consists of 6 blocks featuring the residual learning mechanism. As illustrated in Figure 3.3, a block comprises one dense layer with 500 nodes, a batch normalization layer, and a dropout layer with a dropout rate of 0.15. A sub-DNN is a stack of 6 of these blocks, and outputs a 10-dimensional tensor. The output DNN consists of three dense layers with 256 nodes, taking the above mentioned 18 sub DNN's outputs as input and produces a one-dimensional output. In order to prevent the over fit problem, spectral normalization and L2 regularization are implemented in all the above dense layers with the λ parameter of L2 regularization set at 0.1.

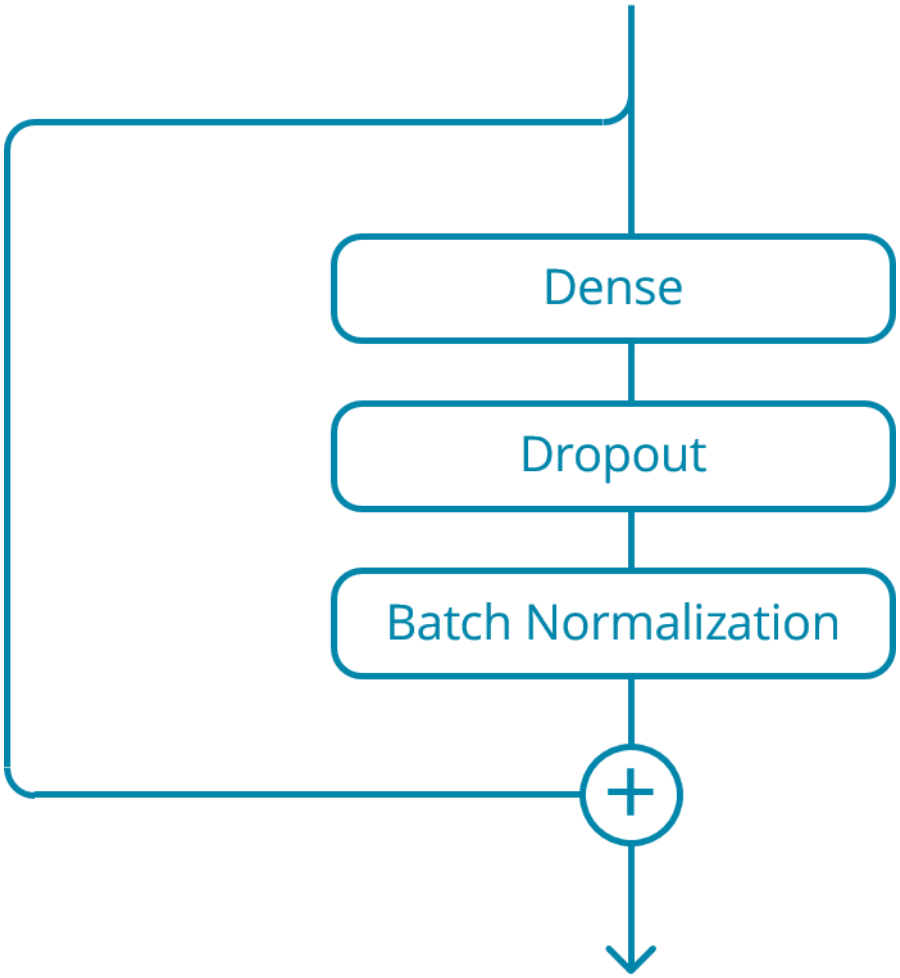


Figure 3. 3. A schematic of residual dense block.

A single residual dense block consists of a 500 nodes dense layer, a Batch Normalization Layer and a Dropout Layer. The output is the addition of what is processed through this block and the input of this block itself.

3.2.4.3 Preprocessing

Feature preprocessing procedures are depicted in Figure 3.2. The features are separated into 18 subsets based on radial or angular features and ligand side element types, which are then input into the corresponding DNN. Radial features are segregated into nine groups using the E_l of the feature name $E_l\_E_p\_k$, while angular features are divided into nine

groups based on the E_l of the feature name $E_{p1}\_E_l\_E_{p2}\_p\_q$. Each segregated feature group is subsequently input into the DNN responsible for processing E_l features.

3.2.4.4 Loss Function (PCC_RMSE)

A loss function combining correlation coefficient R and RMSE, as utilized in OnionNet 11, was adopted. We call this loss function PCC_RMSE. The equation is as follows:

$$loss = \alpha(1 - R) + (1 - \alpha)RMSE \tag{11}$$

R represents the correlation coefficient, RMSE denotes the root mean square error, and α is the coefficient determining the R and RMSE ratio, with values ranging between 0 and 1. All models in this study were trained with α = 0.7.

3.2.4.5 Model Training

Each model underwent 200 epochs of training with early stopping set at 20 epochs. The initial learning rate value was set at 1e-3, and the learning rate was multiplied by 0.2 if the validation loss failed to improve for five epochs.

3.3. Results and Discussion

3.3.1. Overview

We trained two different AQDnet models, depending on the task being evaluated. The docking AQDnet, which is a model specialized for the evaluation of docking power, was used for the evaluation of docking power and screening power. Similarly, the scoring AQDnet was used to evaluate scoring power. The only difference between the docking AQDnet and the scoring AQDnet is the training data. Details are described in Section 3.2.3.

3.3.2. Docking Power

During training the docking AQDnet, the Pearson correlation coefficient (PCC), root mean square error (RMSE), and the loss function (PCC_RMSE) described below were monitored. Finally, the loss function of validation dataset was minimized at epoch 27 (Table 3.2), hence we adopted the model of this epoch.

For the evaluation of our model, we used the CASF-2016 benchmark dataset (Su, et al., 2019). Notably, our model achieved a top 1 success rate of 92.6% in the docking power test, surpassing all other evaluated symmetry functions (SFs) in the CASF-2016 (Figure 3.4a). Additionally, our model demonstrated top 2 and top 3 success rates of 96.5% and 97.2%, respectively, ranking first in both categories (Figure 3.4b-c). These outstanding results are attributed to the successful learning of the P-L QEL by the model, facilitated by our data augmentation method.

Table 3. 2. Transition of loss function during training of docking-specific model

epochs	loss	mse	RMSE	PCC	PCC_RMSE	val_loss	val_mse	val_RMSE	val_PCC	val_PCC_RMSE	lr
1	5122.27	65.03	6.46	0.16	2.53	4675.83	105.27	7.72	0.37	2.77	0.001
2	4265.55	59.62	6.15	0.31	2.33	3861.26	59.73	5.81	0.40	2.17	0.001
3	3522.03	56.98	5.99	0.37	2.24	3188.04	52.63	5.67	0.43	2.11	0.001
4	2908.00	55.31	5.86	0.40	2.18	2632.30	49.97	5.64	0.45	2.08	0.001
5	2401.10	53.56	5.74	0.44	2.11	2173.50	47.89	5.56	0.48	2.04	0.001
6	1982.62	51.57	5.59	0.48	2.05	1794.72	45.96	5.42	0.52	1.97	0.001
7	1637.15	49.54	5.44	0.51	1.98	1482.01	43.94	5.25	0.55	1.90	0.001
8	1351.94	47.79	5.31	0.53	1.92	1223.84	43.19	5.08	0.57	1.83	0.001
9	1116.48	46.44	5.20	0.55	1.88	1010.74	42.24	5.04	0.58	1.81	0.001
10	922.10	45.44	5.12	0.56	1.84	834.82	44.49	4.98	0.58	1.80	0.001
11	761.64	44.78	5.07	0.57	1.82	689.67	51.12	5.21	0.58	1.87	0.001

12	629.16	44.28	5.02	0.58	1.80	569.72	48.78	5.08	0.59	1.82	0.001
13	519.78	43.65	4.97	0.59	1.78	470.77	50.11	5.16	0.58	1.85	0.001
14	429.49	43.49	4.95	0.59	1.77	389.11	54.67	5.35	0.57	1.92	0.001
15	354.95	43.28	4.94	0.59	1.77	321.72	60.32	5.58	0.56	1.99	0.001
16	308.40	43.08	4.89	0.60	1.75	300.65	46.08	4.97	0.62	1.76	0.0002
17	295.09	40.96	4.75	0.62	1.69	289.42	40.37	4.65	0.65	1.65	0.0002
18	284.21	40.34	4.70	0.63	1.67	278.73	39.16	4.54	0.66	1.61	0.0002
19	273.75	39.95	4.67	0.63	1.66	268.43	36.04	4.40	0.68	1.56	0.0002
20	263.67	39.57	4.64	0.64	1.65	258.56	34.03	4.40	0.68	1.55	0.0002
21	253.97	39.24	4.62	0.64	1.64	249.04	35.47	4.35	0.68	1.54	0.0002
22	244.63	38.98	4.60	0.64	1.63	239.88	34.28	4.33	0.69	1.53	0.0002
23	235.64	38.74	4.58	0.65	1.62	231.09	34.56	4.38	0.68	1.55	0.0002
24	226.98	38.43	4.56	0.65	1.61	222.58	34.48	4.32	0.69	1.53	0.0002
25	218.64	38.21	4.54	0.65	1.61	214.42	36.58	4.34	0.68	1.54	0.0002
26	210.61	37.92	4.52	0.66	1.60	206.54	35.77	4.32	0.68	1.53	0.0002
27	202.88	37.74	4.51	0.66	1.59	198.94	34.40	4.27	0.69	1.51	0.0002
28	195.43	37.53	4.49	0.66	1.58	191.70	38.88	4.41	0.68	1.56	0.0002
29	188.26	37.29	4.48	0.66	1.58	184.69	39.32	4.47	0.67	1.58	0.0002
30	181.36	37.15	4.47	0.67	1.57	177.99	44.08	4.68	0.67	1.65	0.0002
31			inf	0.71				inf	0.80		0.0002

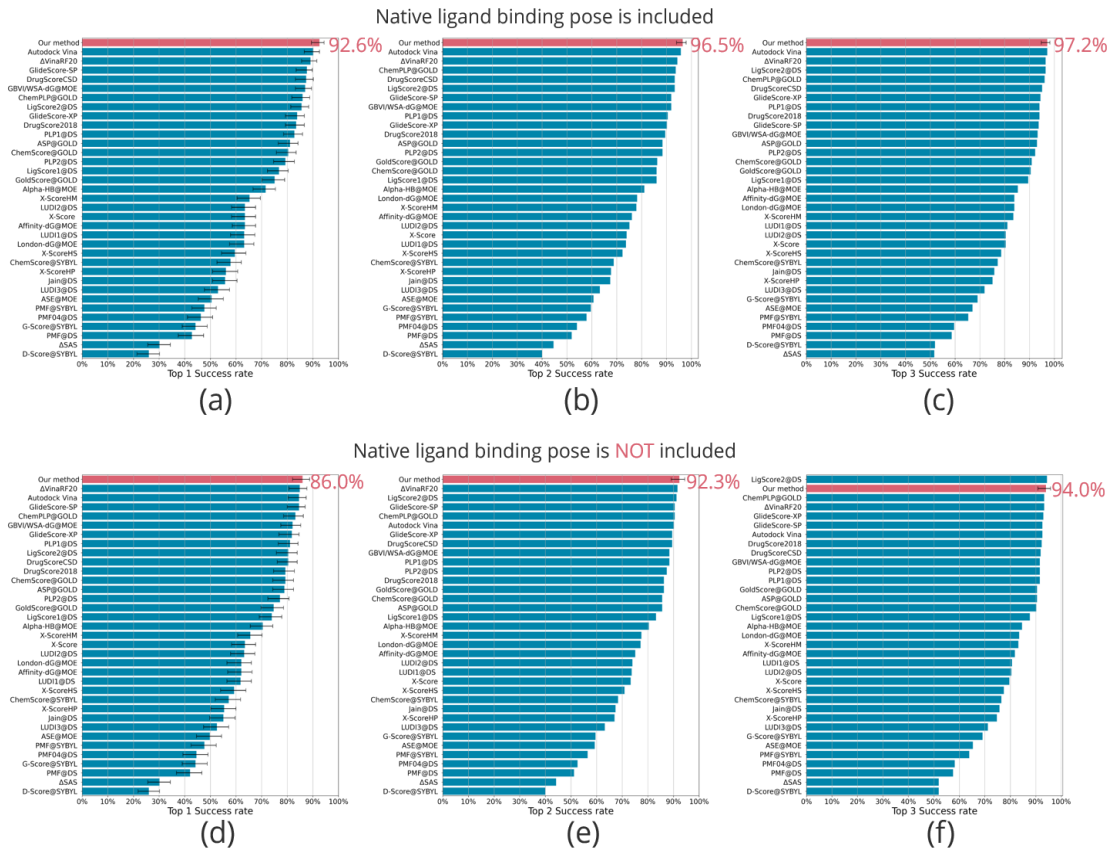


Figure 3. 4. CASF-2016 docking power of AQDnet model.

CASF-2016 docking power test performance (top 1/2/3 success rates) of the AQDnet model (colored pink) and other scoring functions. (a-c) Top 1/2/3 success rates when the native ligand binding pose is included. (d-f) Top 1/2/3 success rates when the native ligand binding pose is not included. Error bars indicate 90% confidence interval obtained from 10000 replicated bootstrapping samples. The 90% confidence intervals are derived from the CASF-2016(Su, et al., 2019) results. However, error bars are not shown for those that do not provide 90% confidence intervals in CASF2016 (e.g., Top2 success rate). Adapted with permission from Su, Minyi et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. Journal of Chemical Information and Modeling, 2019. Copyright 2019 American Chemical Society.

The attempt to expand the data by generating numerous configurations from crystal structure appears clearly beneficial in overcoming the insufficient number of training

samples. However, this has not been achieved so far due to the difficulty of labeling. In this study, we developed a method of labeling the generated configurations by calculating the energy difference from the most stable pose using SQM/COSMO and correcting it using the experimental pKa. This labeling strategy can be employed in various existing machine learning methods for predicting pKa from crystal structures. It is of great significance as it has the potential to substantially enhance the performance of docking tasks, which pose difficulties for many of the current machine learning approaches. Although the proposed data expansion technique has the drawback of necessitating a substantial amount of time for calculation, it is expected to be incorporated into diverse machine learning methods in the future.

DeepBSP (Bao, et al., 2021) uses a simple data augmentation method using configuration generation. This method utilizes the root-mean-square deviations (RMSDs) from crystal structures as labels. While this method is very easy to prepare data for, it is not trained to predict binding affinity and therefore cannot compare binding affinity between ligands. Therefore, it is not possible to evaluate the scoring power and screening power of CASF-2016. In contrast, our method can compare binding affinities among complexes. In fact, we achieved the 4th place in the screening power test of the top 1% enrichment factor, indicating that our method can be used to compare binding affinities between complexes.

In learning the energy landscapes, the amount of high-energy unstable conformation to be included in the training data is very important. To determine the optimal value of this energy threshold, we conducted filtration under three different conditions. As outlined in Section 3.2.2, we designated ΔE as the variation between the lowest energy conformation for each complex and the corresponding conformation's energy. Filtering of the training data was performed under the following three conditions: ΔE <10 kcal/mol, ΔE <20 kcal/mol, and ΔE <30 kcal/mol. For the RMSDs, all the data were filtered under the consistent condition (<2.5 Å). The model was then trained on each of the training datasets. In this case, the number of training data differed across the three conditions. Although the quantity of training data increases as the ΔE threshold elevates, this tendency is also apparent in actual VS, making it valuable to explore the optimal values, including the effect of the increase in the number of training data points.

The results of the docking power evaluation under three different energy thresholds are shown in Figure 3.5. The model trained on filtered training data with ΔE <30 kcal/mol had a Docking power top 1 success rate of 92.6%, the highest success rate among the three conditions. The model trained with training data filtered by ΔE <10 kcal/mol had the lowest docking power top 1 success rate of 71.2%. A success rate of 90.5% was obtained for ΔE <20 kcal/mol, which is not as good as that for ΔE <30 kcal/mol, but still good. From these results, it was observed that docking performance tended to increase as the ΔE filtering threshold was increased.

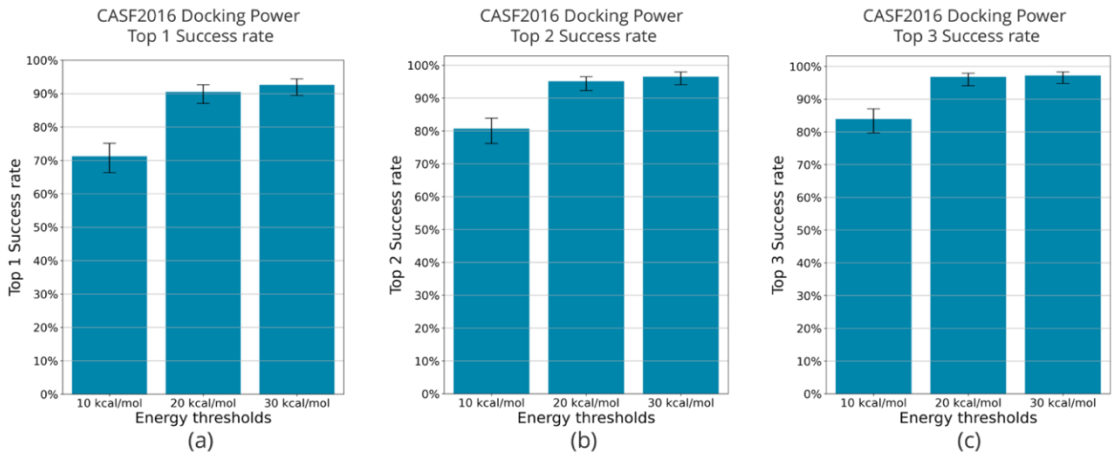


Figure 3. 5. Comparison of CASF-2016 docking power by energy threshold.

The threshold for energy filtering was varied and used to compare CASF-2016 docking power in terms of (a) top 1 success rate, (b) top 2 success rate, and (c) top 3 success rate. Error bars indicate 90% confidence interval obtained from 10000 replicated bootstrapping samples.

3.3.3. Scoring Power

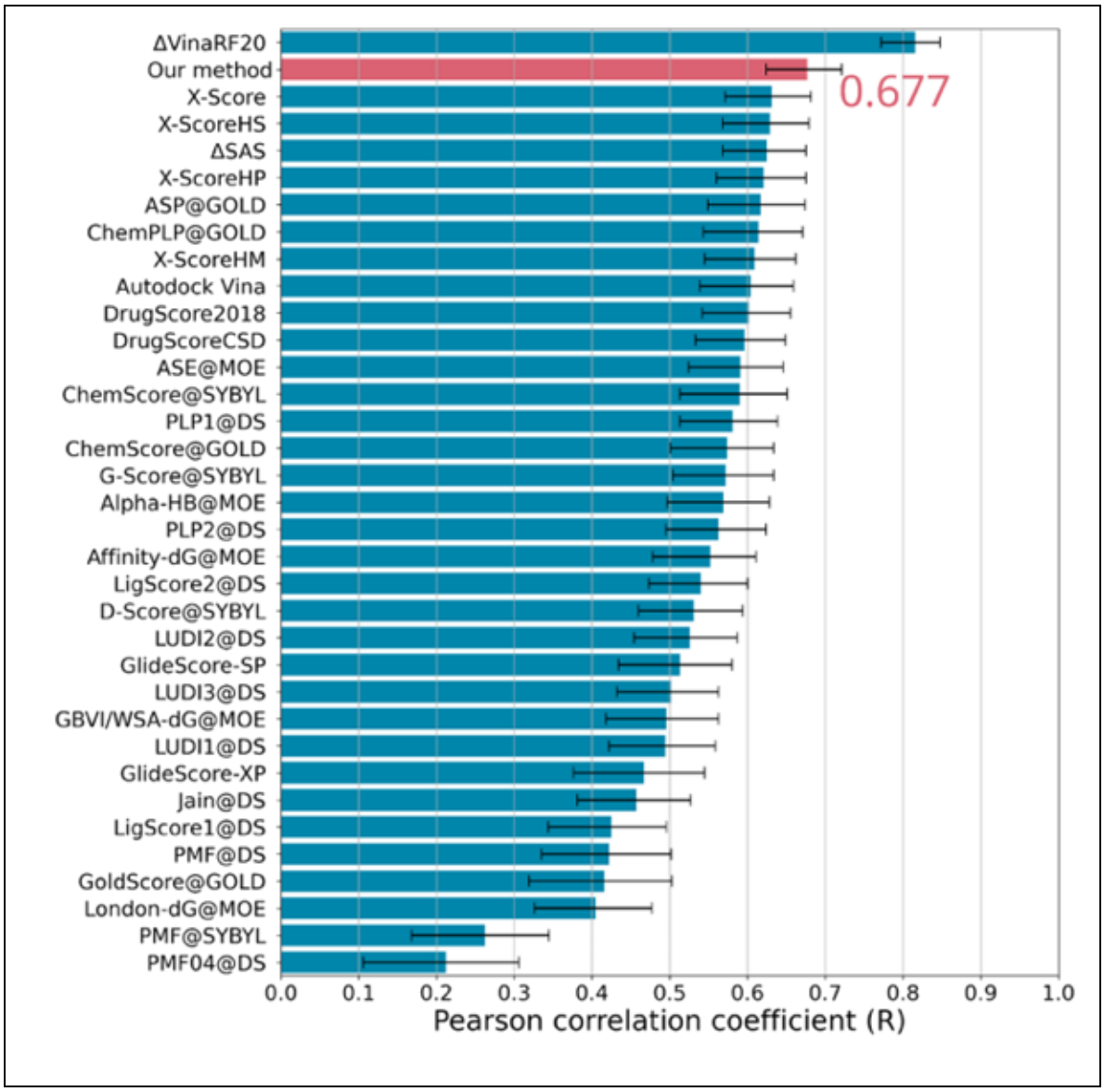
During training the scoring AQDnet, PCC, RMSE and the loss function described below were monitored. Finally, the loss function (PCC_RMSE) of validation dataset was minimized at epoch 33 (Table 3.3), hence we adopted the model of this epoch.

In the scoring power test, our scoring AQDnet achieved 0.677. This is the 2nd best result among the SFs evaluated in the CASF-2016 (Figure 3.6). Note that the best scoring power of the docking AQDnet was about 0.63, while that of the scoring AQDnet was 0.047 higher.

Table 3. 3. Transition of loss function during training of scoring-specific model

epochs	loss	mse	RMSE	PCC	PCC_RMSE	val_loss	val_mse	val_RMSE	val_PCC	val_PCC_RMSE	lr
1	5434.040	61.549	6.306	0.134	2.500	5348.101	1144.897	31.165	0.608	9.624	0.001
2	5239.084	29.788	3.864	0.311	1.649	5101.955	35.447	4.875	0.614	1.733	0.001
3	4990.287	13.467	2.794	0.400	1.266	4847.937	28.745	4.193	0.576	1.555	0.001
4	4738.110	10.752	2.458	0.455	1.127	4599.478	8.079	2.218	0.567	0.969	0.001
5	4494.788	8.493	2.217	0.511	1.015	4362.607	7.473	2.161	0.552	0.962	0.001
6	4263.021	7.800	2.121	0.549	0.959	4137.628	8.974	2.475	0.521	1.078	0.001
7	4042.963	7.127	2.038	0.577	0.914	3924.041	8.006	2.351	0.479	1.070	0.001
8	3834.214	6.718	1.977	0.599	0.881	3721.450	7.688	2.290	0.470	1.058	0.001
9	3636.235	6.306	1.923	0.614	0.854	3529.267	6.884	2.126	0.490	0.995	0.001
10	3448.480	5.991	1.875	0.629	0.828	3347.016	6.416	2.018	0.520	0.942	0.001
11	3270.425	5.718	1.831	0.641	0.807	3174.169	5.901	1.890	0.552	0.881	0.001
12	3101.567	5.500	1.792	0.653	0.786	3010.270	5.586	1.819	0.574	0.844	0.001
13	2941.432	5.256	1.754	0.663	0.768	2854.839	5.263	1.762	0.598	0.810	0.001
14	2789.571	5.089	1.725	0.674	0.751	2707.441	4.995	1.724	0.621	0.782	0.001
15	2645.554	4.922	1.697	0.683	0.736	2567.663	4.795	1.696	0.637	0.763	0.001
16	2508.973	4.739	1.668	0.693	0.721	2435.107	4.637	1.675	0.651	0.747	0.001
17	2379.449	4.581	1.641	0.701	0.707	2309.401	4.516	1.659	0.661	0.735	0.001
18	2256.617	4.477	1.617	0.707	0.696	2190.187	4.400	1.642	0.671	0.723	0.001
19	2140.125	4.347	1.594	0.716	0.682	2077.131	4.305	1.625	0.679	0.712	0.001
20	2029.651	4.201	1.570	0.724	0.669	1969.919	4.248	1.616	0.684	0.706	0.001
21	1924.884	4.088	1.551	0.730	0.659	1868.238	4.158	1.597	0.692	0.695	0.001
22	1825.531	4.047	1.535	0.733	0.652	1771.814	4.115	1.590	0.696	0.690	0.001
23	1731.305	3.927	1.517	0.740	0.642	1680.371	4.082	1.584	0.699	0.686	0.001

24	1641.944	3.813	1.495	0.747	0.630	1593.654	4.077	1.585	0.699	0.686	0.001
25	1557.201	3.725	1.479	0.752	0.622	1511.414	4.060	1.582	0.701	0.684	0.001
26	1476.831	3.634	1.459	0.757	0.612	1433.420	4.038	1.579	0.703	0.681	0.001
27	1400.617	3.589	1.448	0.760	0.606	1359.455	4.024	1.579	0.705	0.680	0.001
28	1328.335	3.505	1.433	0.766	0.598	1289.313	4.036	1.584	0.705	0.682	0.001
29	1259.785	3.427	1.415	0.770	0.589	1222.788	4.006	1.579	0.707	0.679	0.001
30	1194.775	3.358	1.399	0.775	0.581	1159.702	4.006	1.584	0.708	0.679	0.001
31	1133.123	3.287	1.383	0.778	0.574	1099.870	3.984	1.576	0.709	0.676	0.001
32	1074.654	3.224	1.370	0.783	0.567	1043.134	4.005	1.585	0.709	0.679	0.001
33	1019.206	3.177	1.357	0.786	0.561	989.319	3.953	1.568	0.711	0.673	0.001
34	966.618	3.107	1.342	0.791	0.553	938.292	3.992	1.577	0.708	0.677	0.001
35	916.747	3.050	1.328	0.794	0.546	889.895	3.987	1.569	0.707	0.676	0.001
36			inf	0.790				inf	0.800		0.001


ΔVinaRF20
Our method
X-Score
X-ScoreHS
ΔSAS
X-ScoreHP
ASP@GOLD
ChemPLP@GOLD
X-ScoreHM
Autodock Vina
DrugScore2018
DrugScoreCSD
ASE@MOE
ChemScore@SYBYL
PLP1@DS
ChemScore@GOLD
G-Score@SYBYL
Alpha-HB@MOE
PLP2@DS
Affinity-dG@MOE
LigScore2@DS
D-Score@SYBYL
LUDI2@DS
GlideScore-SP
LUDI3@DS
GBVI/WSA-dG@MOE
LUDI1@DS
GlideScore-XP
Jain@DS
LigScore1@DS
PMF@DS
GoldScore@GOLD
London-dG@MOE
PMF@SYBYL
PMF04@DS
0.677
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Pearson correlation coefficient (R)

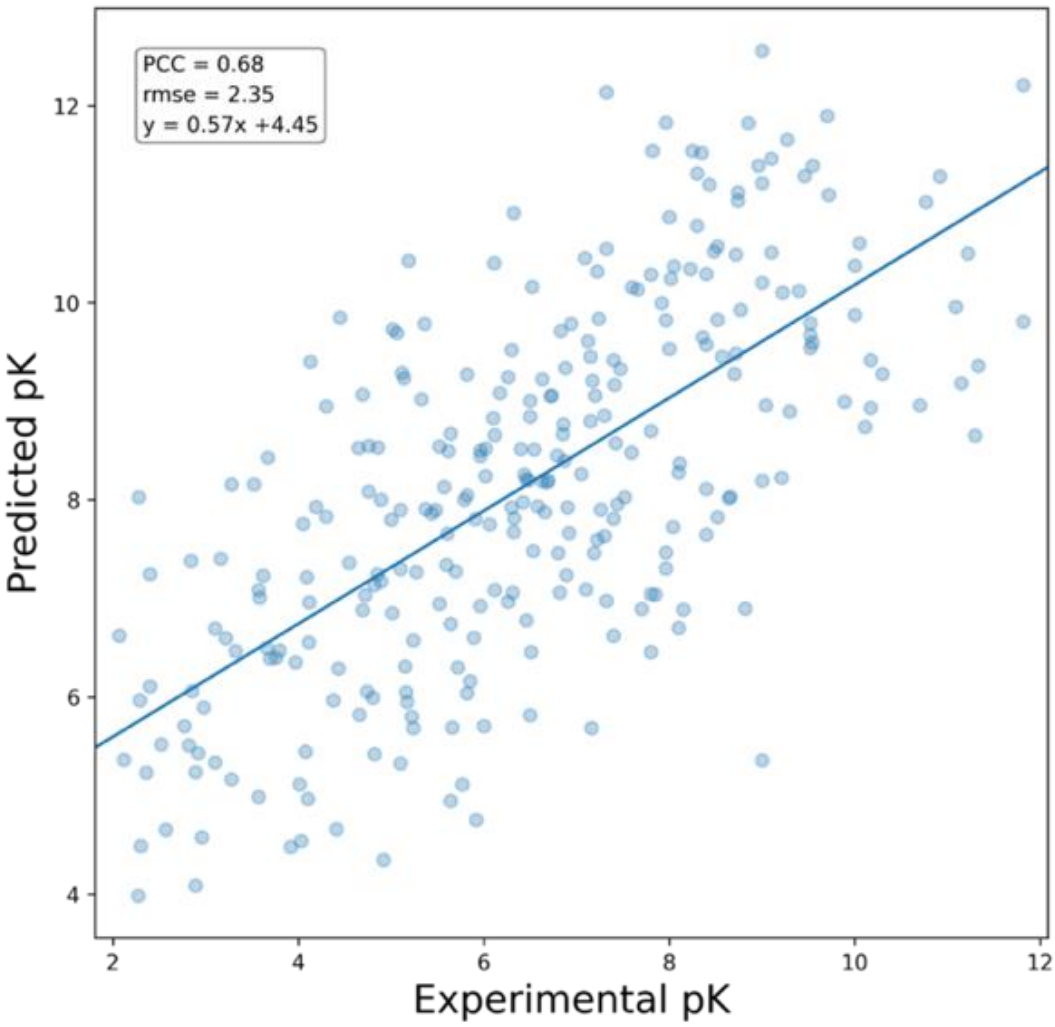


Figure 3. 6. CASF-2016 scoring power of AQDnet model.

Comparison of AQDnet with the scoring functions listed in CASF-2016 (upper panel). Error bars indicate 90% confidence interval obtained from 10000 replicated bootstrapping samples. Scatter plot of CASF-2016 'coreset' experimental values (horizontal axis) and predicted values by AQDnet (vertical axis) (bottom panel). Error bars indicate 90% confidence interval obtained from 10000 replicated bootstrapping samples. The 90% confidence intervals are derived from the CASF-2016(Su, et al., 2019) results. Adapted with permission from Su, Minyi et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. Journal of Chemical Information and Modeling, 2019. Copyright 2019 American Chemical Society.

The scoring AQDnet was trained using only the conformations that were very close to the crystal structure data (ΔE <2 kcal/mol, RMSD <2.5 Å). To improve scoring performance, it is necessary to learn as many different P-L systems as possible, rather than learning many different conformations of the same P-L combination, as in the training data for the docking AQDnet. Comparing the scoring power of the docking AQDnet with that of the scoring AQDnet, which was trained by adding samples with relatively high energy, the scoring power of the scoring AQDnet is higher. A tendency towards better scoring power emerged when filtering with very low ΔE values, in contrast to the docking power, which improved when filtering with higher ΔE values. These results suggest that it is challenging to improve both the docking power results and the scoring power results with a single model at this time. Our future objective is to enhance the scoring power of the docking AQDnet by augmenting the training dataset, enabling a single model to perform superiorly in both docking and scoring evaluations.

In this case, AQDnet's CASF-2016 scoring power was 0.68, which is not competitive with the method that is currently state of the art in CASF-2016 scoring power. After CASF-2016 was published, many machine-learning methods have been evaluated using CASF-2016. For example, Pufnucy (Stepniewska-Dziubinska, et al., 2018), which learns the 3D structure of the P-L complex using a 3DCNN, achieves a scoring power PCC of 0.780, and InteractionGraphNet (IGN)(Jiang, et al., 2021), which learns the structure of the P-L complex as a graph, achieved a scoring power PCC of 0.837. The current best is extended connectivity interaction features (ECIF) (Sanchez-Cruz, et al., 2021), which learns the count values of contacts between atoms represented in detail as features in gradient boosting decision tree (GBDT), achieving a PCC of 0.866.

Our method is apparently not competitive in terms of the scoring power compared to the current state-of-the-art methods. However, it’s important to emphasize that before utilizing such state-of-the-art methods, it's necessary to provide the most stable, or near-stable, structures of the P-L complex, like its crystal structure, for accurately evaluating P-L affinity. Actually, current state-of-the-art methods do not exhibit exceptional docking performance, unlike AQDnet, requiring the "true" structure of the P-L complex for accurately determining P-L affinity with these approaches. In contrast, the AQDnet system exhibit the significant performance for identification of the most stable pose as shown in

docking power result. This was established in the AQDnet system by combining the quantum docking techniques. This is an exclusive feature of our AQDnet system, compared with the current state-of-the-art methods.

3.3.4. Screening Power

Screening power was evaluated using the docking AQDnet. In the forward screening power test, our model's screening power average enrichment factor top 1% was 8.81, , placing it in the 4th position among the SFs evaluated by the updated CASF-2016 (Figure 3.7a).

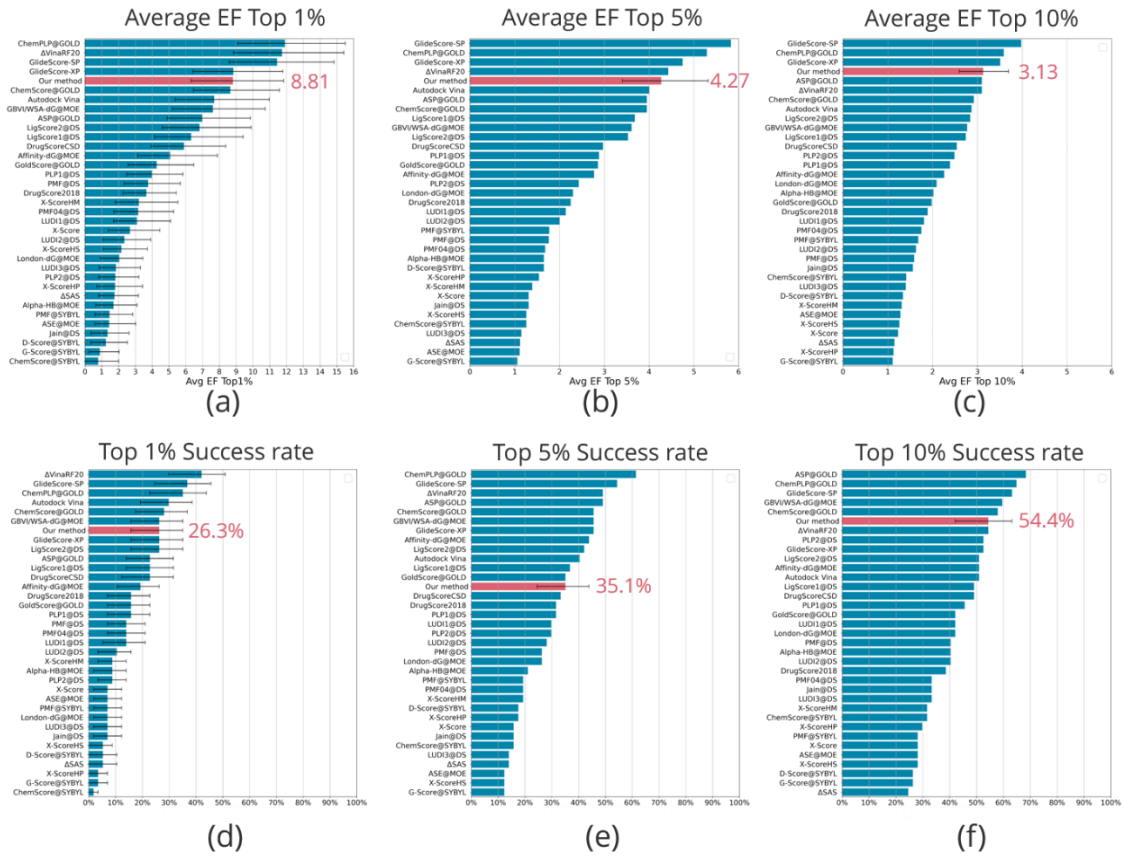
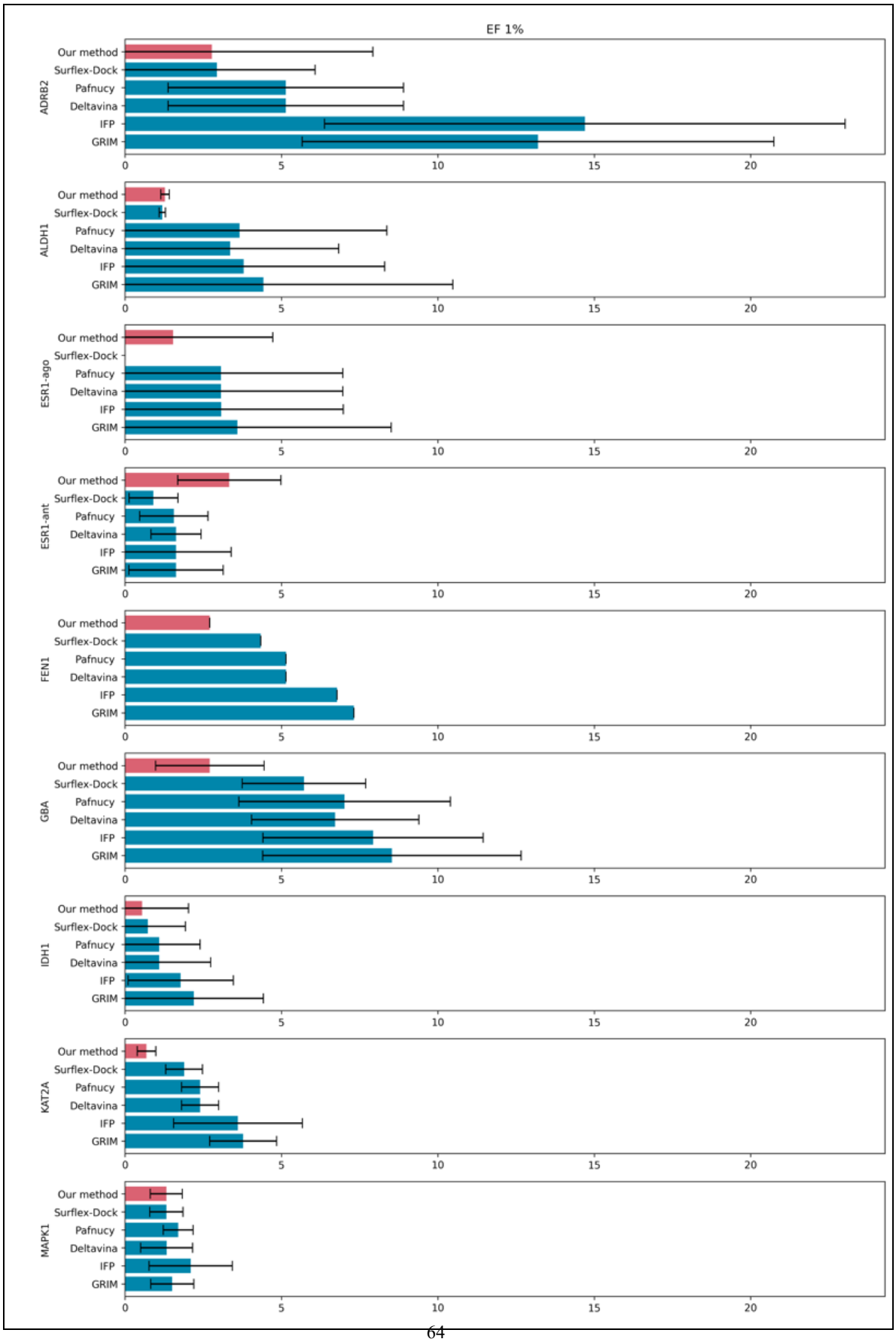


Figure 3. 7. CASF-2016 screening power of AQDnet model.

Average enrichment factor (EF) top 1/5/10 % performance (a-c) and top 1/5/10 % success rate performance (d-f) of the AQDnet model (colored pink) and other scoring functions on the CASF-2016 screening power test. Error bars indicate 90% confidence interval obtained from 10000 replicated bootstrapping samples. Error bars indicate 90% confidence interval obtained from 10000 replicated bootstrapping samples. The 90% confidence intervals are derived from the CASF-2016(Su, et al., 2019) results. However, error bars are not shown for those that do not provide 90% confidence intervals in CASF2016 (e.g., Average EF Top 5%). Adapted with permission from Su, Minyi et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. Journal of Chemical Information and Modeling, 2019. Copyright 2019 American Chemical Society.

The CASF2016 screening power test sample was generated by cross docking 285 ligands against 95 proteins, and it has not been experimentally shown whether the decoys are truly inactive. It is also reported that there are serious biases in the dataset using decoy, but it is unclear how CASF-2016 addresses this issue. Therefore, as another reliable indicator, we conducted a validation with LIT-PCBA(Tran-Nguyen, et al., 2020) dataset. In order to compare our method with the state of the art machine-learning based method in screening power, we compared the results of the five scoring functions (Surflex-Dock (Jain, 2007), Pafnucy (Stepniewska-Dziubinska, et al., 2018), Deltavina (Wang and Zhang, 2017), IFP (Marcou and Rognan, 2007) and GRIM (Desaphy, et al., 2013)) evaluated in Viet-Khoa et al (Tran-Nguyen, et al., 2021) . Details of the method are described in Section 3.2.3 AQDnet predictions were made for all 15 targets included in the LIT-PCBA and compared using enrichment factor (EF) of 1% as metric. The results are shown in Figure 3.8. Although it fell short of the state-of-the-art machine-learning based methods Pafnucy and deltavina for 11 of the 15 targets, it outperformed the other 5 SFs for ESR1-antagonist, MTORC1 and TP53. Both evaluation results of AQDnet's CASF-2016 and LIT-PCBA are not competitive with current state of the art methods, but they show a reasonable screening performance.


EF 1%
ADRB2
ALDH1
ESR1-ago
ESR1-ant
FEN1
GBA
IDH1
KAT2A
MAPK1
Our method
Surflex-Dock
Pafnucy
Deltavina
IFP
GRIM
0
5
10
15
20

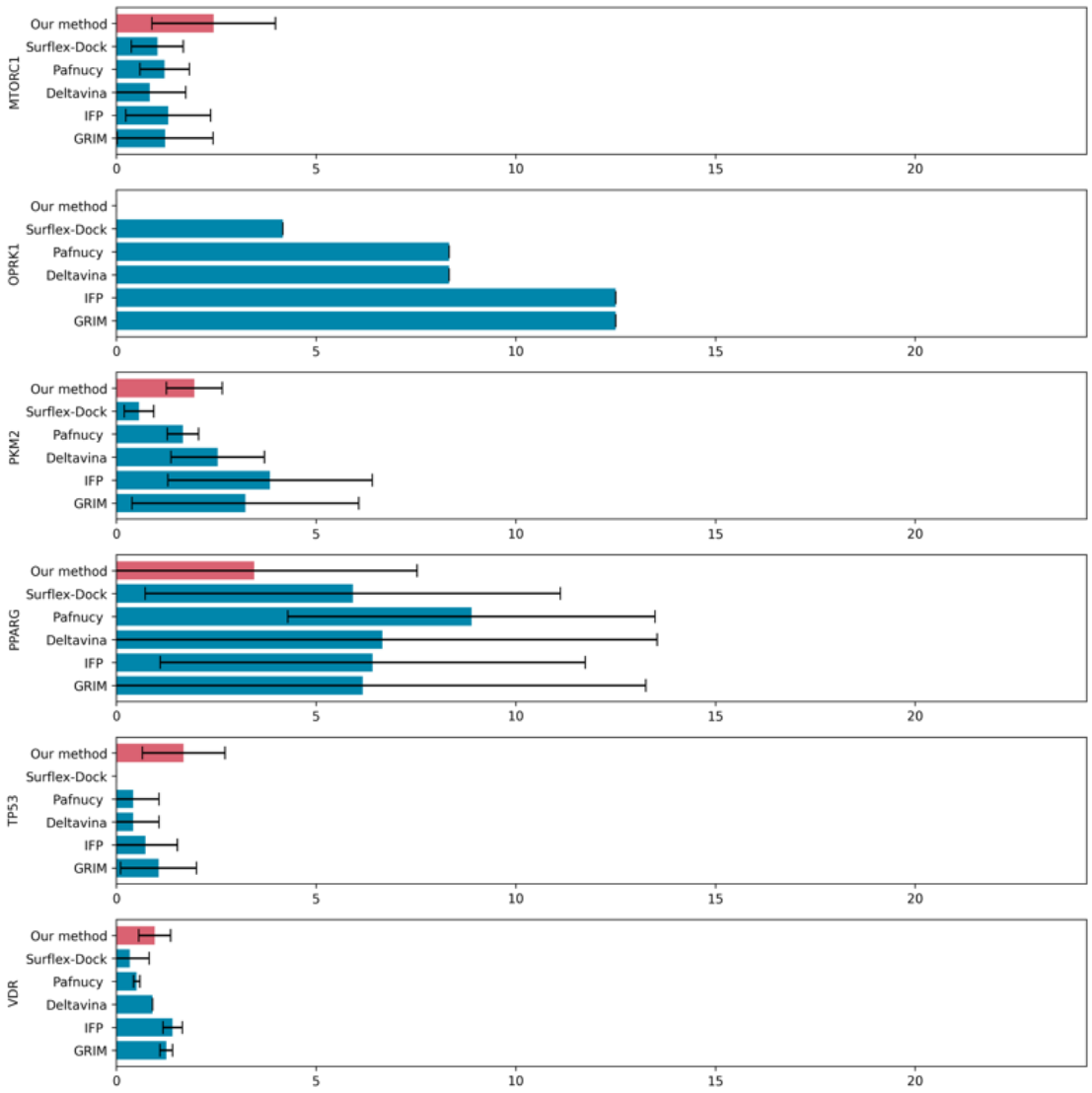


Figure 3. 8. Comparison of AQDnet performance on the LIT-PCBA data set.

If multiple templates existed for a target, docking was performed for all of them. Bars indicate the mean of their Enrichment Factor 1% (EF1%), and error bars indicate standard deviation (SD). If only one template is given for a target (e.g. FEN1), the width of the error bar is 0. Adapted with permission from Tran-Nguyen, Viet-Khoa et al. True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better. Journal of Chemical Information and Modeling, 2021. Copyright 2021 American Chemical Society.

We discuss below how the methodology can be improved from a developmental standpoint and its prospective applications. There are three ways to improve this methodology.

First, a small number of complexes was currently used to generate the training data. Most of the existing methods that use PDBbind as the source of training data use about 12,000 complexes. In contrast, we used only 1,123 complexes herein as training data due to computational reasons. By creating a training dataset based on a larger number of complexes, we expect to achieve improved results in scoring power and screening power.

Second, our method presently neglects the energy difference between the free and bound states of the ligand. To accurately compare ligand binding affinities, it is crucial to consider the energy differences, specifically requiring an understanding of the ligand's topology to ascertain its distortion. However, the presented algorithm in AQDnet solely considers the distance between the protein atoms and the ligand atoms, without accounting for the covalent bonding of the ligand atoms. This limitation could potentially hinder the scoring power and screening power of the model, highlighting the need for a feature extraction approach that incorporates the topology of ligands to further enhance the methodology. This work is actually ongoing in our lab.

Third, our method does not recognize the exposed moiety of the ligand molecule from the protein pocket. Feature extraction is performed for protein atoms within 12 Å of each atom of the ligand and regions lacking protein atoms within 12 Å of ligand atoms are disregarded. This is a major challenge in predicting the binding affinity of relatively large molecules such as peptides. For example, even if two ligands are in the same conformation in a protein pocket, the stability of one that fits completely in the protein pocket and another with a large hydrophobic segment exposed outside the pocket may differ significantly. However, AQDnet predicts that the two would have the same binding affinity. Therefore, implementing measures such as increasing the cutoff distance is necessary when dealing with large ligand molecules.

3.4. Summary

In this study, we devised a novel approach for predicting the P-L binding affinity by applying ACSF, which is suitable for describing the energy of a single molecule. This method takes into account not only two-body interactions but also three-body interactions and generates high-resolution features that clearly represent changes in atomic coordinates of 0.1 Å or smaller. In these two respects, this method proves valuable for VS applications. Additionally, we have devised a data augmentation technique that leverages configuration generation and QM calculations, thereby overcoming the shortage of training data for P-L binding affinity prediction. We believe that this method has great value in that it can be easily integrated into existing machine learning methods and enhance their performance.

To date, machine learning has not been used for predicting P-L binding affinity by assigning energy labels to decoy poses generated during configuration. Doubling the data in this context is already considered innovative. The data expansion method used in AQDnet can expand data 900-1000 times. This method can be easily applied to existing P-L binding affinity prediction methods, potentially significantly improving docking performance. This is considered both novel and important.

Our method was evaluated by the CASF-2016 dataset and ranked first in the docking power test of the top 1 success rate and fourth in the screening power test of the top 1% enrichment factor. Note here that our method achieves the above results by creating training data based on the crystal structures of only 1123 complexes. Actually, this is the smallest number of training data used by any machine learning method, which suggests that our method can learn features originated from the QM-based first-principles (*i.e.*, the quantum field) found in P-L interactions. Accordingly, further increase in the number of training data will enable us to effectively obtain higher performance.

Moreover, the presented method does not take ligand characteristics into account and predicts affinities only with information on P-L interactions, which is unfavorable for scoring and screening performance. Nevertheless, the obtained achievements were within an acceptable level even for the actual use of the presented system. Thus, our method is also promising in terms of improvements of the scoring and screening performance by

incorporating a ligand energy feature, which leads to involvement of how much the docked ligand is destabilized from the most stable conformation of the free ligand.

Note herein that prior to performing most of current P-L docking methods including state-of-the-art ones, we need to provide the most stable structures (or one that is close to be native) of the P-L complexes, such as the crystal structures of the complexes, for the appropriate evaluation of the P-L affinity. In fact, present state-of-the-art methods for the P-L docking task do not exhibit the excellent docking performance (whereas the AQDnet system does, as described in this report), and thus do require the “true” (native) structures of the P-L complexes for obtaining the appropriate P-L affinity. In contrast, the AQDnet system discriminates the most appropriate pose among other many poses, as shown in the presented docking score data. This is an exclusive feature of our AQDnet system in the actual VS workflow, compared with those of the other methods.

Chapter 4

Multi-Shelled ECIF: Improved Extended Connectivity Interaction Features for Accurate Binding Affinity Prediction

4.1. Background

Extended connectivity interaction features (ECIF) is a method developed to predict P-L binding affinity, allowing for detailed atomic representation. It performed very well in terms of Comparative Assessment of Scoring Functions 2016 (CASF-2016) scoring power test. However, ECIF has the limitation of not being able to adequately account for interatomic distances.

To investigate what kind of distance representation is effective for P-L binding affinity prediction, we have developed two algorithms that improved ECIF's feature extraction method to take distance into account. One is multi-shelled ECIF, which takes into account the distance between atoms by dividing the distance between atoms into multiple layers. The other is weighted ECIF, which weights the importance of interactions according to the distance between atoms. A comparison of these two methods shows that multi-shelled ECIF outperforms weighted ECIF and the original ECIF, achieving a CASF-2016 scoring

power Pearson correlation coefficient of 0.877. All the codes and data are available on GitHub (https://github.com/koji11235/MSECIFv2). Supplementary data are available at GitHub (https://github.com/koji11235/MSECIFv2)

Prediction of protein-ligand (P-L) binding affinity plays a very important role in virtual screening (VS). The docking-based VS is the large-scale application of the docking methodology. The components of the docking method are a search algorithm that generates poses within the binding site, scoring functions that quantify the quality of the docking poses, and one or more scoring functions to predict binding affinity. Examples of this scoring function include a classifier for active/inactive classes, a regressor for the absolute value of the binding free energy, and a compound ranking system that sorts compounds according to a certain score. (Gilson and Zhou, 2007; Kimber, et al., 2021; Liu and Wang, 2015; Maia, et al., 2020; Pason and Sotriffer, 2016; Yang, et al., 2022) For accurate and fast VS, various methods for affinity prediction have been developed, including physics-based (Abel, et al., 2017; Khalak, et al., 2021; King, et al., 2021) and machine learning-based methods (Jiang, et al., 2021; Jimenez, et al., 2018; Moon, et al., 2022; Sanchez-Cruz, et al., 2021; Stepniewska-Dziubinska, et al., 2018; Wojcikowski, et al., 2019; Zhang, et al., 2023; Zhang, et al., 2023; Zheng, et al., 2019).

Comparative Assessment of Scoring Functions 2016 (CASF-2016) (Li, et al., 2014; Su, et al., 2019) is a benchmark dataset created with the concept of evaluating scoring performance and docking performance separately. Four metrics are provided at CASF-2016: scoring power, which evaluates the linear correlation between predicted and experimental binding affinity values given a crystal structure; ranking power, which evaluates the accuracy of the binding affinity rank prediction for a given target protein; docking power, which evaluates the accuracy of the prediction of the native binding pose from 100 generated ligand configurations; and screening power, which predicts the binding ligand for a given protein. The performance of many machine learning methods developed in recent years has been evaluated by CASF-2016 and reported. Especially in terms of scoring power, methods using machine learning have been very successful, and their performance far exceeds that of physics-based methods. For example, the following have demonstrated very good performance in terms of CASF-2016 scoring power. PLEC-nn (Wojcikowski, et al., 2019) is a model with fingerprint-based features trained on a neural

network, and it has achieved a Pearson Correlation Coefficient (Pearson's R) of 0.820. Kdeep (Jimenez, et al., 2018) and Pafnucy (Stepniewska-Dziubinska, et al., 2018) are convolutional neural networks (CNN) trained on a 3D voxel representation of the P-L complex, with Pearson's Rs of 0.82 and 0.78, respectively. InteractionGraphNet (IGN) (Jiang, et al., 2021) represents the P-L complex as a graph and is trained by a graph neural network, with a Pearson's R of 0.837. OnionNet (Zheng, et al., 2019) is a multiple-layer inter-molecular contact-based feature that has been trained using a CNN, and it achieved a Pearson's R of 0.816. Extended connectivity interaction features (ECIF) (Sanchez-Cruz, et al., 2021) achieved a Pearson's R of 0.866 in terms of CASF-2016 scoring power, the best performance reported to date. ECIF is unique in its ability to represent atoms in very fine detail. While many methods rely solely on elemental species to represent an atom, ECIF takes into account five additional factors in its representation: explicit valence, the number of attached heavy atoms, the number of attached hydrogens, aromaticity, and ring membership. The number of interactions between atoms represented by this method that exist within a certain threshold distance is defined as a feature value. A Gradient boosted decision tree (GBDT) was used for the model and trained with the features created above. For the training dataset, the PDBbind (Wang, et al., 2004) v2016 plus a part of PDBbind v2019 was used. For more details, please refer to the original publication.

While feature extraction using the above method has the great advantage of being able to represent the atoms in detail, it has the problem of not being able to take into account any differences in distance within 6 Å. This is because the count value of how many interactions are within 6 Å is used as the feature value. For example, hydrogen bonds are generally reported to be 2.5 to 3.5 Å away and play a very important role in P-L interactions. The magnitude of the contribution to the P-L interaction is likely to be different for interactions at this distance and for more distant interactions. Therefore, we have attempted to improve the performance of ECIF by modifying it to take into account the interatomic distance. Despite the availability of multiple methods for quantifying distances, no systematic comparison has been conducted to determine the most effective approach. In this study, we prepared two methods for expressing distances that can be freely combined with methods of expressing atoms qualitatively. By comparing them, we seek to gain insight into the usefulness of various distance considerations.

The first method is to divide the distance into multiple shells. We call this method multi-shelled ECIF. With it, the distance can be taken into account by explicitly representing the count value region, such as 0~2.5Å, 2.5~3.5Å, 3.5~6.0Å, etc. The second method is to apply weights that are the inverse of the square of the distance and make the sum of the weights the feature value. We call this method weighted ECIF. It was created based on the hypothesis that interactions at close distances are more important.

A comparison of multi-shelled ECIF, weighted ECIF, and ECIF shows that the multi-shell ECIF significantly outperformed the weighted ECIF and the original ECIF, achieving a scoring Pearson's R of 0.877 for the CASF-2016. On the other hand, weighted ECIF was inferior to the original ECIF concerning CASF-2016, but outperformed the original ECIF on evaluation with the LIT-PCBA dataset. Both multi-shelled ECIF and weighted ECIF are freely available on GitHub (https://github.com/koji11235/MSECIFv2)

.

4.2. Methods

4.2.1 Feature Extraction

All code used in this study was developed in Python 3.9.0. RDkit version 2022.09.5 was used. Both the multi-shelled ECIF and weighted ECIF used in this study were developed based on ECIF; see the original publication for more information on ECIF. An overview of ECIF is given below. ECIF represents each atom by six elements: an atom's symbol, explicit valence, number of attached heavy atoms, number of attached hydrogens, aromaticity, and ring membership, and joined by a semicolon. For example, nitrogen is represented as N;3;2;1;0;0 if it has a valence of 3, two attached heavy atoms, one attached hydrogen, no aromaticity, and is not contained in a ring. The pairs of protein side atoms and ligand side atoms expressed in this way are then joined by hyphens. For example, the interaction between N;3;2;1;0;0 on the protein side and O;2;1;1;0;0 on the ligand side is expressed as N;3;2;1;0;0-O;2;1;1;0;0, which is the name of the feature. The count value of how many of the relevant interactions are present in the P-L complex within the distance

threshold is then used as the feature value. The above process is then applied to all ECIF-type representation combinations to generate features of 1,540 dimensions.

We found that the information on aromaticity in the original ECIF is almost the same as that of ring membership, and its contribution to the prediction is small. Therefore, we removed the aromaticity entry and represented each atom with five elements: an atom's symbol, explicit valence, number of attached heavy atoms, number of attached hydrogens, and ring membership. For example, nitrogen is represented as N;3;2;1;0 if it has a valence of 3, two attached heavy atoms, one attached hydrogen, and is not contained in a ring. Both of the following two methods are based on this feature extraction method with modifications to allow distance to be accounted for.

Regarding the treatment of tautomers, each atom on the protein side is treated uniformly by dictionary-based mapping, and other tautomers are not considered (e.g., HIS-NE2 is treated only as N;3;2;1;1). For ligands, only the states described in the sdf file are considered and other tautomers are not considered. For the PDBbind dataset, only the states in the provided ligand sdf are considered, and for LIT-PCBA, only the states expressed by the SMILES described in the smi file are considered. The protonation/tautomeric state is an important element in accurately characterizing the details of biological systems. The data set used in this study is based on standard protonation/tautomeric states using automated procedures. In particular, histidine residues are treated without a hydrogen atom bonded to the nitrogen. Furthermore, models deposited in the PDB often do not include curated flips of histidine, asparagine, and glutamine residues. This can lead to the phenomenon of P-L complexes adopting non-optimal protonation states at the binding site and problems with hydrogen bond optimization not being taken into account.

Table 4. 1. Correspondence between protein side atoms in PDB and ECIF atom type.

PDB residue	PDB atom	ECIF atom type
ALA	ALA-C	C;4;3;0;0
ALA	ALA-CA	C;4;3;1;0
ALA	ALA-CB	C;4;1;3;0
ALA	ALA-N	N;3;2;1;0
ALA	ALA-O	O;2;1;0;0
ALA	ALA-OXT	O;2;1;0;0
ARG	ARG-C	C;4;3;0;0
ARG	ARG-CA	C;4;3;1;0
ARG	ARG-CB	C;4;2;2;0
ARG	ARG-CD	C;4;2;2;0
ARG	ARG-CG	C;4;2;2;0
ARG	ARG-CZ	C;6;3;0;0
ARG	ARG-N	N;3;2;1;0
ARG	ARG-NE	N;4;2;1;0
ARG	ARG-NH1	N;4;1;2;0
ARG	ARG-NH2	N;4;1;2;0
ARG	ARG-O	O;2;1;0;0
ARG	ARG-OXT	O;2;1;0;0
ASN	ASN-C	C;4;3;0;0
ASN	ASN-CA	C;4;3;1;0
ASN	ASN-CB	C;4;2;2;0
ASN	ASN-CG	C;4;3;0;0
ASN	ASN-N	N;3;2;1;0
ASN	ASN-ND2	N;3;1;2;0
ASN	ASN-O	O;2;1;0;0
ASN	ASN-OD1	O;2;1;0;0
ASN	ASN-OXT	O;2;1;0;0
ASP	ASP-C	C;4;3;0;0
ASP	ASP-CA	C;4;3;1;0
ASP	ASP-CB	C;4;2;2;0
ASP	ASP-CG	C;5;3;0;0
ASP	ASP-N	N;3;2;1;0
ASP	ASP-O	O;2;1;0;0
ASP	ASP-OD1	O;2;1;0;0
ASP	ASP-OD2	O;2;1;0;0
ASP	ASP-OXT	O;2;1;0;0
CYS	CYS-C	C;4;3;0;0
CYS	CYS-CA	C;4;3;1;0
CYS	CYS-CB	C;4;2;2;0
CYS	CYS-N	N;3;2;1;0
CYS	CYS-O	O;2;1;0;0
CYS	CYS-OXT	O;2;1;0;0
CYS	CYS-SG	S;2;1;1;0
GLN	GLN-C	C;4;3;0;0
GLN	GLN-CA	C;4;3;1;0
GLN	GLN-CB	C;4;2;2;0
GLN	GLN-CD	C;4;3;0;0
GLN	GLN-CG	C;4;2;2;0
GLN	GLN-N	N;3;2;1;0
GLN	GLN-NE2	N;3;1;2;0
GLN	GLN-O	O;2;1;0;0
GLN	GLN-OE1	O;2;1;0;0
GLN	GLN-OXT	O;2;1;0;0
GLU	GLU-C	C;4;3;0;0
GLU	GLU-CA	C;4;3;1;0
GLU	GLU-CB	C;4;2;2;0
GLU	GLU-CD	C;5;3;0;0
GLU	GLU-CG	C;4;2;2;0
GLU	GLU-N	N;3;2;1;0
GLU	GLU-O	O;2;1;0;0
GLU	GLU-OE1	O;2;1;0;0
GLU	GLU-OE2	O;2;1;0;0
GLU	GLU-OXT	O;2;1;0;0
GLY	GLY-C	C;4;3;0;0
GLY	GLY-CA	C;4;2;2;0
GLY	GLY-N	N;3;2;1;0
GLY	GLY-O	O;2;1;0;0
GLY	GLY-OXT	O;2;1;0;0
HIS	HIS-C	C;4;3;0;0
HIS	HIS-CA	C;4;3;1;0
HIS	HIS-CB	C;4;2;2;0
HIS	HIS-CD2	C;4;2;1;1
HIS	HIS-CE1	C;4;2;1;1
HIS	HIS-CG	C;4;3;0;1
HIS	HIS-N	N;3;2;1;0
HIS	HIS-ND1	N;3;2;0;1
HIS	HIS-NE2	N;3;2;1;1
HIS	HIS-O	O;2;1;0;0
HIS	HIS-OXT	O;2;1;0;0
ILE	ILE-C	C;4;3;0;0
ILE	ILE-CA	C;4;3;1;0
ILE	ILE-CB	C;4;3;1;0
ILE	ILE-CD1	C;4;1;3;0

ILE	ILE-CG1	C;4;2;2;0	MET	MET-OXT	O;2;1;0;0	THR	THR-CB	C;4;3;1;0
ILE	ILE-CG2	C;4;1;3;0	MET	MET-SD	S;2;2;0;0	THR	THR-CG2	C;4;1;3;0
ILE	ILE-N	N;3;2;1;0	PHE	PHE-C	C;4;3;0;0	THR	THR-N	N;3;2;1;0
ILE	ILE-O	O;2;1;0;0	PHE	PHE-CA	C;4;3;1;0	THR	THR-O	O;2;1;0;0
ILE	ILE-OXT	O;2;1;0;0	PHE	PHE-CB	C;4;2;2;0	THR	THR-OG1	O;2;1;1;0
LEU	LEU-C	C;4;3;0;0	PHE	PHE-CD1	C;4;2;1;1	THR	THR-OXT	O;2;1;0;0
LEU	LEU-CA	C;4;3;1;0	PHE	PHE-CD2	C;4;2;1;1	TRP	TRP-C	C;4;3;0;0
LEU	LEU-CB	C;4;2;2;0	PHE	PHE-CE1	C;4;2;1;1	TRP	TRP-CA	C;4;3;1;0
LEU	LEU-CD1	C;4;1;3;0	PHE	PHE-CE2	C;4;2;1;1	TRP	TRP-CB	C;4;2;2;0
LEU	LEU-CD2	C;4;1;3;0	PHE	PHE-CG	C;4;3;0;1	TRP	TRP-CD1	C;4;2;1;1
LEU	LEU-CG	C;4;3;1;0	PHE	PHE-CZ	C;4;2;1;1	TRP	TRP-CD2	C;4;3;0;1
LEU	LEU-N	N;3;2;1;0	PHE	PHE-N	N;3;2;1;0	TRP	TRP-CE2	C;4;3;0;1
LEU	LEU-O	O;2;1;0;0	PHE	PHE-O	O;2;1;0;0	TRP	TRP-CE3	C;4;2;1;1
LEU	LEU-OXT	O;2;1;0;0	PHE	PHE-OXT	O;2;1;0;0	TRP	TRP-CG	C;4;3;0;1
LYS	LYS-C	C;4;3;0;0	PRO	PRO-C	C;4;3;0;0	TRP	TRP-CH2	C;4;2;1;1
LYS	LYS-CA	C;4;3;1;0	PRO	PRO-CA	C;4;3;1;1	TRP	TRP-CZ2	C;4;2;1;1
LYS	LYS-CB	C;4;2;2;0	PRO	PRO-CB	C;4;2;2;1	TRP	TRP-CZ3	C;4;2;1;1
LYS	LYS-CD	C;4;2;2;0	PRO	PRO-CD	C;4;2;2;1	TRP	TRP-N	N;3;2;1;0
LYS	LYS-CE	C;4;2;2;0	PRO	PRO-CG	C;4;2;2;1	TRP	TRP-NE1	N;3;2;1;1
LYS	LYS-CG	C;4;2;2;0	PRO	PRO-N	N;3;3;0;1	TRP	TRP-O	O;2;1;0;0
LYS	LYS-N	N;3;2;1;0	PRO	PRO-O	O;2;1;0;0	TRP	TRP-OXT	O;2;1;0;0
LYS	LYS-NZ	N;4;1;3;0	PRO	PRO-OXT	O;2;1;0;0	TYR	TYR-C	C;4;3;0;0
LYS	LYS-O	O;2;1;0;0	SER	SER-C	C;4;3;0;0	TYR	TYR-CA	C;4;3;1;0
LYS	LYS-OXT	O;2;1;0;0	SER	SER-CA	C;4;3;1;0	TYR	TYR-CB	C;4;2;2;0
MET	MET-C	C;4;3;0;0	SER	SER-CB	C;4;2;2;0	TYR	TYR-CD1	C;4;2;1;1
MET	MET-CA	C;4;3;1;0	SER	SER-N	N;3;2;1;0	TYR	TYR-CD2	C;4;2;1;1
MET	MET-CB	C;4;2;2;0	SER	SER-O	O;2;1;0;0	TYR	TYR-CE1	C;4;2;1;1
MET	MET-CE	C;4;1;3;0	SER	SER-OG	O;2;1;1;0	TYR	TYR-CE2	C;4;2;1;1
MET	MET-CG	C;4;2;2;0	SER	SER-OXT	O;2;1;0;0	TYR	TYR-CG	C;4;3;0;1
MET	MET-N	N;3;2;1;0	THR	THR-C	C;4;3;0;0	TYR	TYR-CZ	C;4;3;0;1
MET	MET-O	O;2;1;0;0	THR	THR-CA	C;4;3;1;0	TYR	TYR-N	N;3;2;1;0

TYR	TYR-O	O;2;1;0;0	VAL	VAL-CA	C;4;3;1;0	VAL	VAL-N	N;3;2;1;0
TYR	TYR-OH	O;2;1;1;0	VAL	VAL-CB	C;4;3;1;0	VAL	VAL-O	O;2;1;0;0
TYR	TYR-OXT	O;2;1;0;0	VAL	VAL-CG1	C;4;1;3;0	VAL	VAL-OXT	O;2;1;0;0
VAL	VAL-C	C;4;3;0;0	VAL	VAL-CG2	C;4;1;3;0			

4.2.1.1 Multi-Shelled ECIF

Multi-shelled ECIF is a subdivision of the feature counts of the original ECIF by dividing them into several distance regions. The feature name is defined by appending the upper limit distance of each region to the end of the original ECIF feature, together with a hyphen. The lower limit is identical to the upper limit of one previous feature but is not explicitly shown. For example, if we divide N;3;2;1;0-O;2;1;1;0 into three steps of 0~2.5Å, 2.5~4.5Å, and 4.5~6.0Å, then N;3;2;1;0-O;2;1;1;0 is represented by dividing it into N;3;2;1;0-O;2;1;1;0-2.5, N;3;2;1;0-O;2;1;1;0-4.5, and N;3;2;1;0-O;2;1;1;0-6.0. The number of relevant interactions in each region is then used as the feature. For example, if there are three N;3;2;1;0-O;2;1;1;0 interactions within the distance threshold and the distances are 2Å, 3Å, and 4Å, respectively, then the value of N;3;2;1;0-O;2;1;1;0-2.5 is 1, N;3;2;1;0-O;2;1;1;0-4.5 is 2, and the value of N ;3;2;1;0-O;2;1;1;0-6.0 is 0.

4.2.1.2 Weighted ECIF

Instead of treating interactions between atoms at any distance as one count, as in the original ECIF, the following feature extraction was performed to reflect the intuition that interactions at closer distances are more important. Weights were assigned as the inverse of the inter-atomic distance or the inverse of the square of the inter-atomic distance, and the weighted sum of the weights was used as the feature. The formulations are as follows.

$$f_{X\text{-}Y;} = \sum_{\substack{i \in P_X \\ j \in L_Y}}^{all\ (X,Y)\ pair} \frac{1}{d_{ij}} \tag{12}$$

$$f_{X\text{-}Y;} = \sum_{\substack{i \in P_X \\ j \in L_Y}}^{all\ (X,Y)\ pair} \frac{1}{{d_{ij}}^2} \tag{13}$$

d_{ij} denotes interatomic distance. X and Y indicate ECIF-style atom representations such as N;3;2;1;0;0. P_X and L_Y designate all X atoms on the protein side and all Y atoms on the ligand side, respectively. The procedure is as follows. Interaction is expressed in the similar

way as ECIF, as in N;3;2;1;0-O;2;1;1;0. Atom pairs that exist at distances within a threshold are obtained. The weights above are calculated using d_{ij} as the distance between atoms in each pair. The sum of the weights is calculated for each interaction and is defined as the feature value. For example, if three interactions of N;3;2;1;;0-O;2;1;1;0 are within the distance threshold and the distances are 2Å, 3Å, and 4Å, respectively, then the value of N;3;2;1;0-O;2;1;1;0 is 1/4+1/9+1/16=0.4236. The resulting feature has the same 1,540 dimensions as the original ECIF.

4.2.1.3 Ligand Descriptor

To compare the performance with the original ECIF in CASF-2016, we used the same ligand descriptor as in the original work. A summary is given below. Of the 200 molecular descriptors available in the 'Descriptors' module of RDkit (2020.03.1), those with zero variance, null values, and extreme values for the entire data set were removed. As a result, 170 molecular descriptors were used. All the ligand descriptors used are listed in the Table 4.2. The 170-dimensional ligand descriptors generated were added to the multi-shelled ECIF and weighted ECIF features to train the model. The ligand descriptor for the LIT-PCBA dataset was calculated using RDkit (2022.09.5)

Table 4. 2. RDkit ligand descriptors used for training along with multi-shelled ECIF and weighted ECIF features.

MaxEStateIndex, MinEStateIndex, MaxAbsEStateIndex, MinAbsEStateIndex, qed, MolWt, HeavyAtomMolWt, ExactMolWt, NumValenceElectrons, FpDensityMorgan1, FpDensityMorgan2, FpDensityMorgan3, BalabanJ, BertzCT, Chi0, Chi0n, Chi0v, Chi1, Chi1n, Chi1v, Chi2n, Chi2v, Chi3n, Chi3v, Chi4n, Chi4v, HallKierAlpha, Kappa1, Kappa2, Kappa3, LabuteASA, PEOE_VSA14, SMR_VSA1, SMR_VSA10, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7, SMR_VSA9, SlogP_VSA1, SlogP_VSA10, SlogP_VSA11, SlogP_VSA12, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, TPSA, EState_VSA1, EState_VSA10, EState_VSA11, EState_VSA2, EState_VSA3, EState_VSA4, EState_VSA5, EState_VSA6, EState_VSA7, EState_VSA8, EState_VSA9, VSA_EState1, VSA_EState10, VSA_EState2, VSA_EState3, VSA_EState4, VSA_EState5, VSA_EState6, VSA_EState7, VSA_EState8, VSA_EState9, FractionCSP3, HeavyAtomCount, NHOHCount, NOCount, NumAliphaticCarbocycles, NumAliphaticHeterocycles, NumAliphaticRings, NumAromaticCarbocycles, NumAromaticHeterocycles, NumAromaticRings, NumHAcceptors, NumHDonors, NumHeteroatoms, NumRotatableBonds, NumSaturatedCarbocycles, NumSaturatedHeterocycles, NumSaturatedRings, RingCount, MolLogP, MolMR, fr_Al_COO, fr_Al_OH, fr_Al_OH_noTert, fr_ArN, fr_Ar_N, fr_Ar_NH, fr_Ar_OH, fr_COO, fr_COO2, fr_C_O, fr_C_O_noCOO, fr_C_S, fr_HOCCN, fr_Imine, fr_NH0, fr_NH1, fr_NH2, fr_N_O, fr_Ndealkylation1, fr_Ndealkylation2, fr_Nhpyrrole, fr_SH, fr_aldehyde, fr_alkyl_carbamate, fr_alkyl_halide, fr_allylic_oxid, fr_amide, fr_amidine, fr_aniline, fr_aryl_methyl, fr_azo, fr_barbitur, fr_benzene, fr_bicyclic, fr_dihydropyridine, fr_epoxide, fr_ester, fr_ether, fr_furan, fr_guanido, fr_halogen, fr_hdrzine, fr_hdrzone, fr_imidazole, fr_imide, fr_isocyan, fr_isothiocyan, fr_ketone, fr_ketone_Topliss, fr_lactam, fr_lactone, fr_methoxy, fr_morpholine, fr_nitrile, fr_nitro, fr_nitro_arom, fr_nitroso, fr_oxazole, fr_oxime,

fr_para_hydroxylation, fr_phenol, fr_phenol_noOrthoHbond, fr_piperdine, fr_piperzine, fr_priamide, fr_pyridine, fr_quatN, fr_sulfide, fr_sulfonamd, fr_sulfone, fr_term_acetylene, fr_tetrazole, fr_thiazole, fr_thiocyan, fr_thiophene, fr_urea

4.2.2 Training Data

To compare the performance with the original ECIF, we used the same training data as in the original study. In brief, the 'core set' (n = 285) from the PDBbind 2016 was used as the test set first. Then, what remained from the PDBbind 2016 ‘refined set’ (n = 4,057) minus the 'core set' was set as the primary source of training data. In addition to this, a 'general set' of the PDBbind 2019 that meets the following criteria was added to the training data. (i) structures resolved by X-ray crystallography with a resolution better or equal than 3.0 Å, (ii) binding data reported accurately (not ‘>’, ‘<’ or ‘∼’) as inhibition constant (Ki) or dissociation constant (Kd) with a range from 1 pM to 10 mM, (iii) atom types of the ligand already included in the ‘refined set’ and (iv) structures not containing any P-L atom pair at a distance of 2.0 Å or less. Three complexes (PDB ID: 2YLC, 3O7U, 3ZNR) with ligands containing atomic types that appear only once in the data set were discarded. In summary, the training set used in this study consists of 9,299 P-L complexes, and the test set consists of 285 structures from CASF-2016. A list of all complexes included in the training and test sets is in the supplementary data of the published paper (link). All protein structures were used without any other processing. On the other hand, Standardizer, JChem 22.6.0 was used for protonation and aromatization of the ligand.

4.2.3 Model

To compare the performance with the original ECIF, we also used GBDT. It was implemented using the Scikit-learn Python library (1.0.1). All the models were trained to predict the binding affinity of the P-L complex denoted as pK, which is the logarithm of the negative base 10 of Ki or Kd. Before the hyperparameter optimization, the GBDT model was built using the same hyperparameters described in ECIF(Sanchez-Cruz, et al., 2021). The specific parameters are 20,000 boosting stages, a maximum depth of 8, a learning rate of 0.005, least squares regression as the loss function to optimize, 0.7 as the fraction of samples that fit individual learners, and "sqrt" as the fraction of features to look at for the optimal split. All remaining parameters were set to default. The best GBDT hyperparameter for multi-shelled ECIF is as follows: 20,000 boosting stages, a maximum depth of 10, a learning rate of 0.005, least squares regression as the loss function to optimize, 0.6 as the fraction of samples that fit individual learners, min sample split of 3, and "sqrt" as the fraction of features to look at for the optimal split. The best GBDT hyperparameter for weighted ECIF is as follows: 30,000 boosting stages, a maximum depth of 10, a learning rate of 0.005, least squares regression as the loss function to optimize, 0.6 as the fraction of samples that fit individual learners, min sample split of 2, and "sqrt" as the fraction of features to look at for the optimal split.

4.2.4 Cross-Validation

The hyperparameters of the multi-shelled ECIF and weighted ECIF features and the GBDT hyperparameters were adjusted by 10-fold cross-validation over the training set. We performed 10-fold cross-validation over the training set with 10 trials each with different random seeds in each condition. The average of the 10 trials was used to compare the conditions. In the cross-validation for the hyperparameters of the multi-shelled ECIF and weighted ECIF features, the GBDT hyperparameters were fixed to the same hyperparameters described in ECIF (Sanchez-Cruz, et al., 2021). The specific parameters are 20 000 boosting stages, a maximum depth of 8, a learning rate of 0.005, least squares regression as the loss function to optimize, 0.7 as the fraction of samples that fit individual learners, and "sqrt" as the fraction of features to look at for the optimal split. All remaining

parameters were set to default. For computational convenience, the optimization of the GBDT parameters was performed in two stages. In the first stage, n_estimators, learning_rate, and max_depth were examined, and in the second stage, min_sample_split, max_features, and subsample were examined. In the first stage, parameters other than n_estimators, learning_rate, and max_depth used values reported in ECIF. In the second phase, we fixed n_estimators, learning_rate, and max_depth, which were optimal in the first phase and examined the target parameters.

4.2.5 Evaluation Method

We used CASF-2016 and LIT-PCBA dataset as independent test sets. For evaluation by CASF-2016, Model performance was evaluated using the Pearson's R and the root mean square error (RMSE) of CASF-2016 scoring power. For evaluation by LIT-PCBA, all 15 targets of LIT-PCBA were predicted by the indicated model and evaluated with EF1%. For those that were given more than one protein template, evaluation was performed for all templates. Preparation of the protein pdb and ligand sdf for the LIT-PCBA dataset was done as follows. The smi file of ligand was loaded using RDkit (https://www.rdkit.org) to generate 3D conformation and add hydrogens. The files were then saved as sdf files. All template PDB files were protonated by MOE (ChemicalComputingGroupULC, 2023) (https://www.chemcomp.com). Then, docking of ligand to protein was performed using GNINA (McNutt, et al., 2021). If multiple templates were given in the target, docking was performed on all templates. The docked ligands were then saved as sdf files. Standardizer, JChem 22.6.0 was used for protonation and aromatization of the ligand. Additionaly, CASF-2007, 2013, 2016 and 2019 defined by Orhobor et al were evaluated using the training and test sets provided by them. Ligands were provided as mol files and used as is. Proteins were provided as mol2 files and were converted to pdb files using openbabel. It should be noted here that the training set differs between CASF-2016 as defined by Orhobor et al. and CASF-2016 by Sanchez-Cruz et al. mentioned above. The hyperparameters for the features and model used were those obtained in the 10-fold cross validation described above.

4.2.6 Permutation Feature Importance

A feature importance analysis was conducted on the best model of multi-shelled ECIF. The feature importance was calculated by the permutation_importance function of the Scikit-learn Python library using 30 repeats with different random seeds.

4.3. Results

4.3.1 Results of Multi-Shelled ECIF

4.3.1.1 CASF-2016 Scoring Power

We first investigated the method of shell segmentation, which is a hyperparameter of multi-shelled ECIF features. Next, we examined the GBDT hyperparameters. We then trained 5000 models with different random seeds using the best conditions obtained above and examined the distribution of CASF-2016 scoring power, Pearson's R and RMSE. The best multi-shelled ECIF model achieved a Pearson's R of 0.877 and an RMSE of 1.152. (Figure 4.1) In the following, we describe the results of our examination of the hyperparameters of multi-shelled ECIF features and GBDT hyperparameters, and finally compare multi-shelled ECIF with other methods.

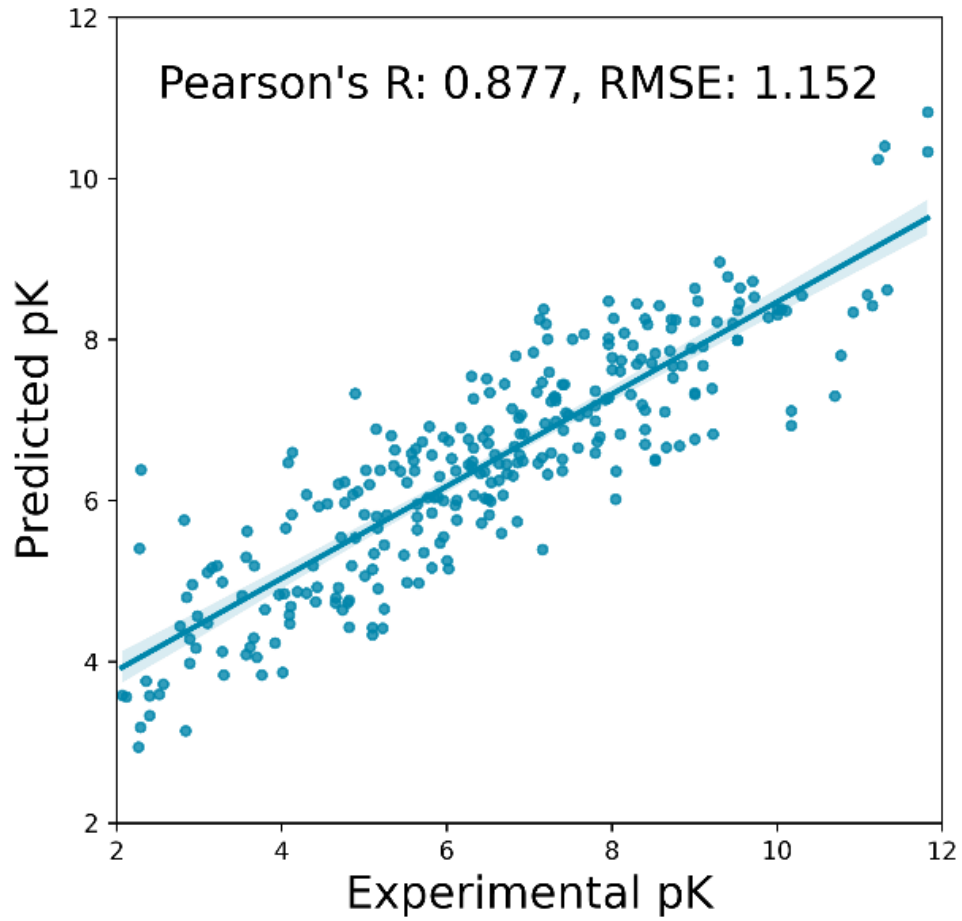


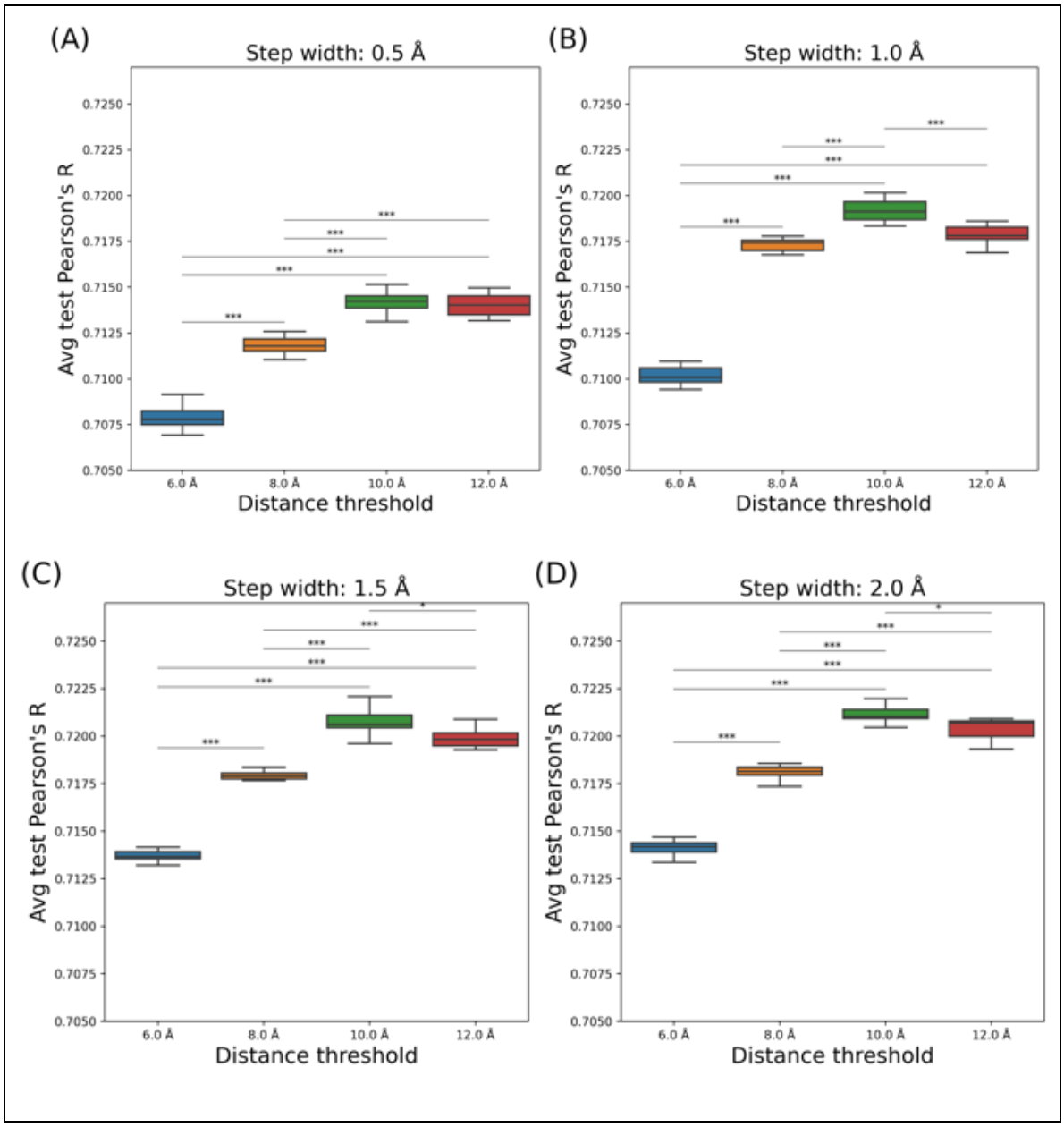
Figure 4. 1. The best result of multi-shelled ECIF.

Scatterplot of experimental values of CASF-2016 'core set' (n=285) and predicted values by the best model of multi-shelled ECIF.

4.3.1.2 Exploration of the Multi-Shelled ECIF Feature Parameters

Since the shell-splitting method potentially affects performance, we searched for the optimal shell-splitting method. To systematically study the approximate optimal shell-splitting method, we chose to split the shell with a constant step width. We performed a conditional study on the distance threshold and the step width. We checked the distribution of interatomic distances between ligand proteins and found that the shortest interatomic distance was about 2.0 Å and that interactions of 2.5 Å or less existed in only 0.03 % of the total interactions existing within 6.0 Å. For this reason, we fixed the minimum shell at 2.5 Å for our study. To align the distance thresholds, the step width may be different for

the terminal portion of each partitioning method than for the other portions. The conditions for the distance threshold and the step width were examined by 10-fold cross-validation over the training set. 10-fold cross-validation was performed in each condition with 10 trials each with 10 different random seeds. A Bonferroni-corrected independent t-test was also used to check whether the differences between conditions were statistically significant. Initially, to determine the maximum distance at which interactions should be taken into consideration, we varied the distance threshold under conditions of 6 to 12 Å with a constant step width. Four different step widths were examined: 2.0 Å, 1.5 Å, 1.0 Å, and 0.5 Å. Comparisons of distance thresholds within the same step width showed the maximum performance at a distance threshold of 10 Å for all conditions. (Figure 4.2)


(A)
Step width: 0.5 Å
(B)
Step width: 1.0 Å
(C)
Step width: 1.5 Å
(D)
Step width: 2.0 Å
Avg test Pearson's R
Distance threshold
6.0 Å
8.0 Å
10.0 Å
12.0 Å
0.7250
0.7225
0.7200
0.7175
0.7150
0.7125
0.7100
0.7075
0.7050
***
*

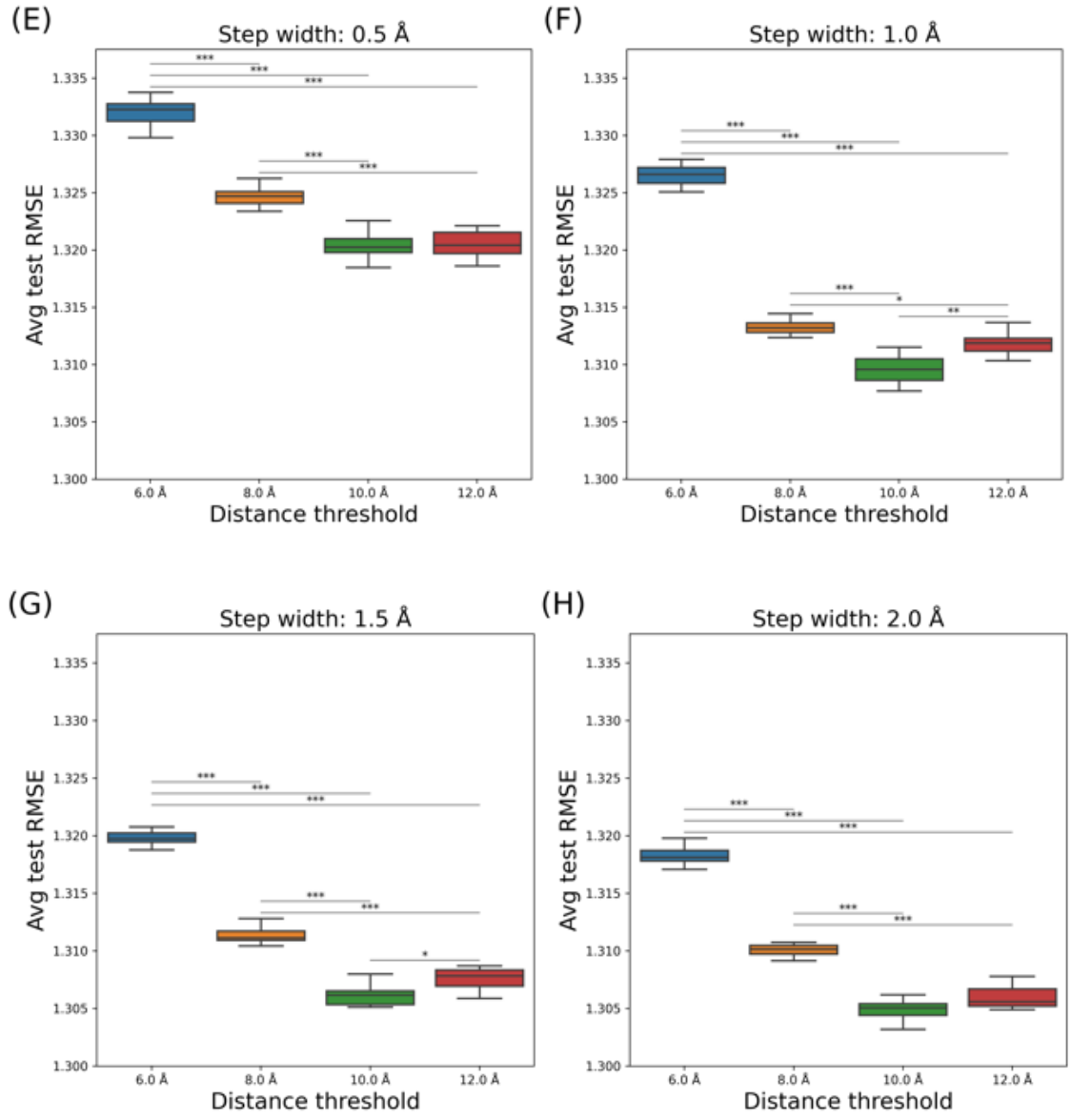
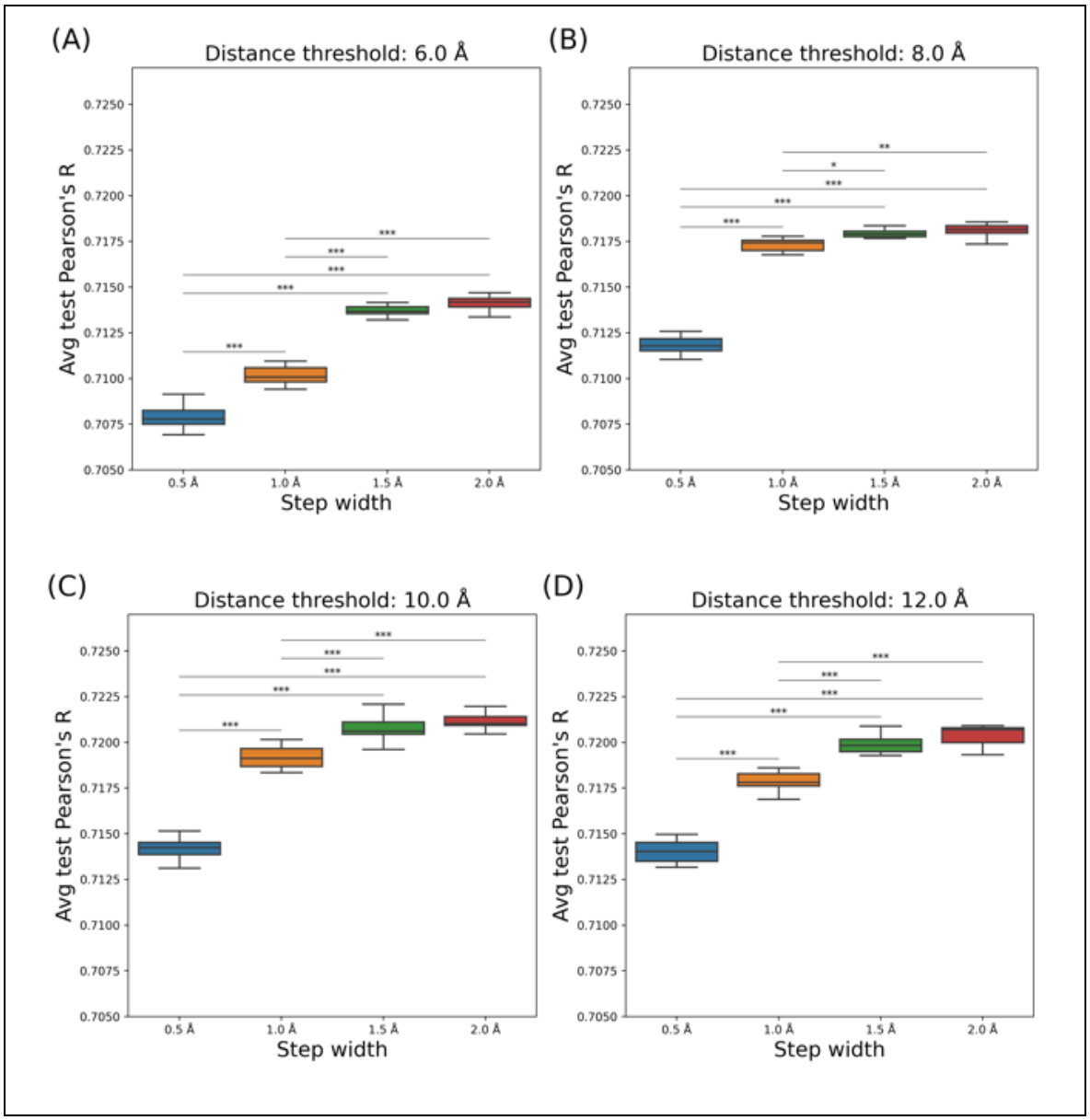


Figure 4. 2. Exploration of the optimal distance threshold of multi-shelled ECIF.

Each boxplot represents the results of 10 runs of 10-fold cross-validation over the training set. The upper four panels (A~D) show the average Pearson's R of 10 runs, and the bottom four panels (E~H) show the average RMSE. The step width is indicated at the top of each plot. The horizontal axis represents the distance threshold. Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

A similar study was conducted for step width. With the distance threshold fixed, a comparison was made for different step widths of 0.5 Å, 1.0 Å, 1.5 Å, and 2.0 Å. Comparisons of step widths showed the maximum performance at step width 2.0 Å for all conditions. (Figure 4.3)


(A)
Distance threshold: 6.0 Å
(B)
Distance threshold: 8.0 Å
(C)
Distance threshold: 10.0 Å
(D)
Distance threshold: 12.0 Å
Avg test Pearson's R
Step width
0.5 Å
1.0 Å
1.5 Å
2.0 Å

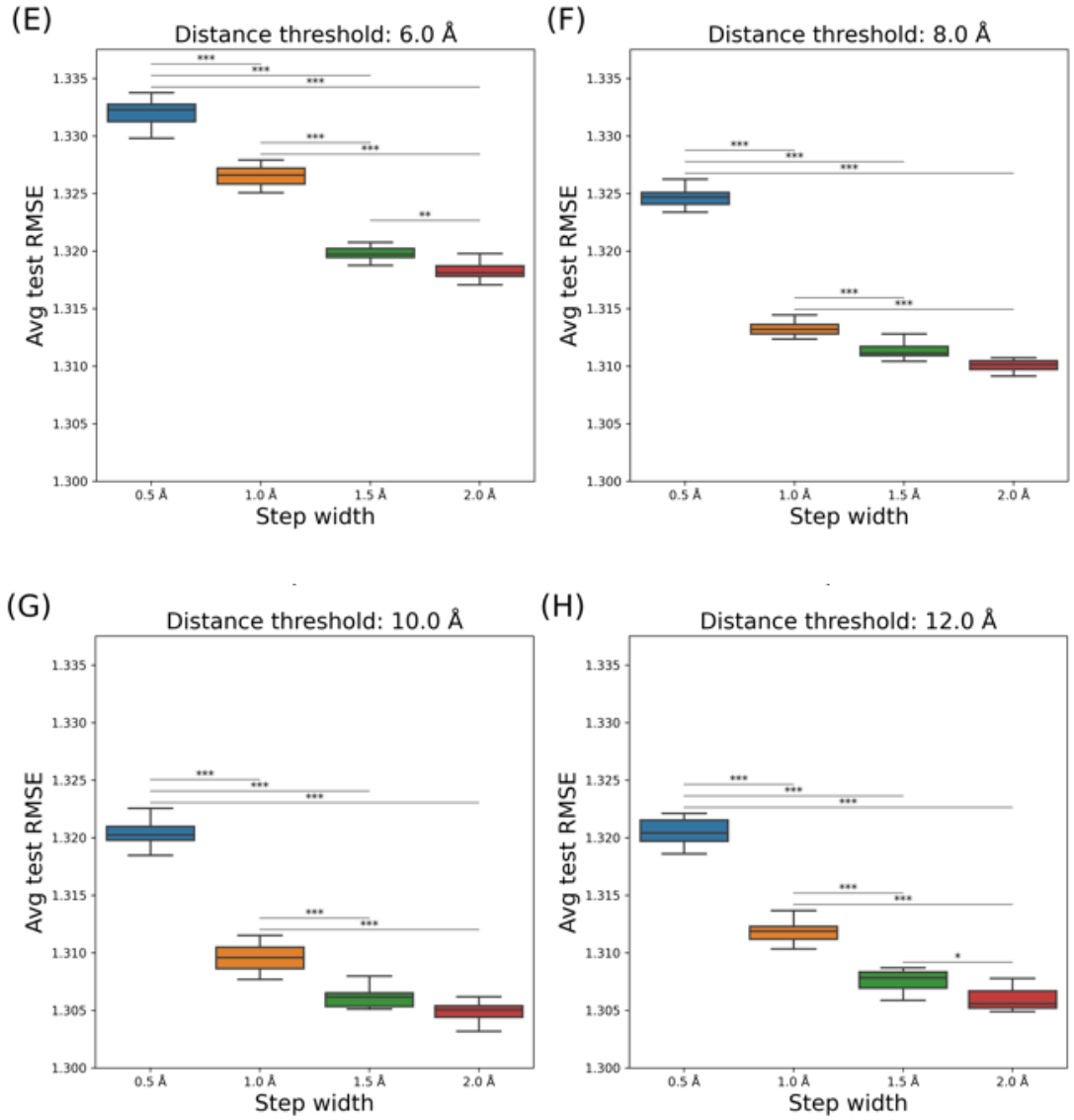


Figure 4. 3. Exploration of the optimal step width of multi-shelled ECIF.

Each boxplot represents the results of 10 runs of 10-fold cross-validation over the training set. The upper four panels (A~D) show the average Pearson's R of 10 runs, and the bottom four panels (E~H) show the average RMSE. The distance threshold is indicated at the top of each plot. The horizontal axis represents the distance threshold. Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

Although there was no significant difference between 1.5 Å and 2.0 Å for almost all results, we decided to use 2.0 Å for step width because 2.0 Å showed higher performance under all conditions. Based on the above considerations, we decided to use 10.0 Å for the distance threshold and 2.0 Å for the step width. A heatmap listing all results is shown in Figure 4.4.

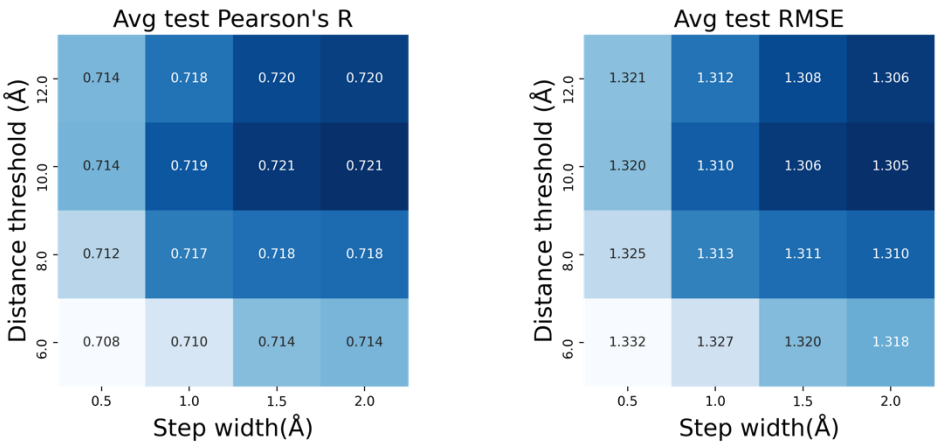


Figure 4. 4. Result of the multi-shelled ECIF feature parameter exploration.

Each cell of the heat map represents the average Pearson's R (left) and average RMSE (right) of 10 runs of 10 fold cross-validation under each condition. The horizontal axis of the heat map shows step width, and the vertical axis shows the distance threshold.

4.3.1.3 GBDT Parameters Optimization for Multi-Shelled ECIF

The distance threshold and step width of multi-shelled ECIF were fixed at 10 Å and 2.0 Å, respectively, and the hyperparameter of GBDT was optimized by 10-fold cross-validation over the training set. As well as exploration of the shell partitioning method of multi-shelled ECIF, 10-fold cross-validation was performed in each condition with 10 runs each with 10 different random seeds. Details are described in Section 4.2.4. The results are shown in Figure 4.5. The best GBDT hyperparameter for multi-shelled ECIF is shown in Section 4.2.3.

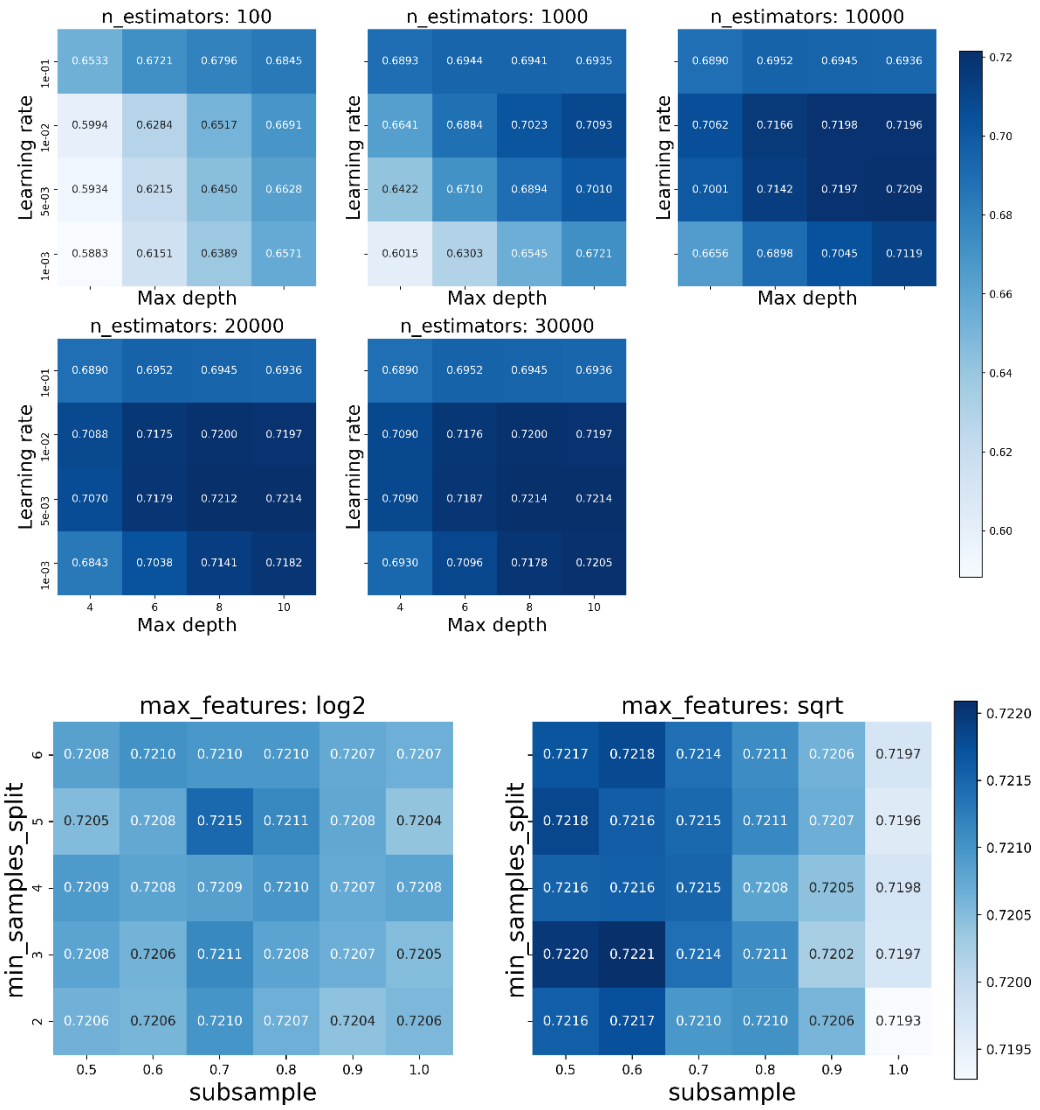


Figure 4. 5. Result of the GBDT parameters optimization of the multi-shelled ECIF.

Each cell in the heat map represents the average Pearson's R of the 10 runs of 10 fold cross-validation trained under each condition. The horizontal axis of the heat map shows step width, and the vertical axis shows the distance threshold. The upper five panels show the optimization results for n_estimators, learning_rate, and max_depth, while the lower two panels show the optimization results for min_sample_split, max_features, and subsample.

4.3.2 Result of Weighted ECIF

4.3.2.1 CASF-2016 Scoring Power

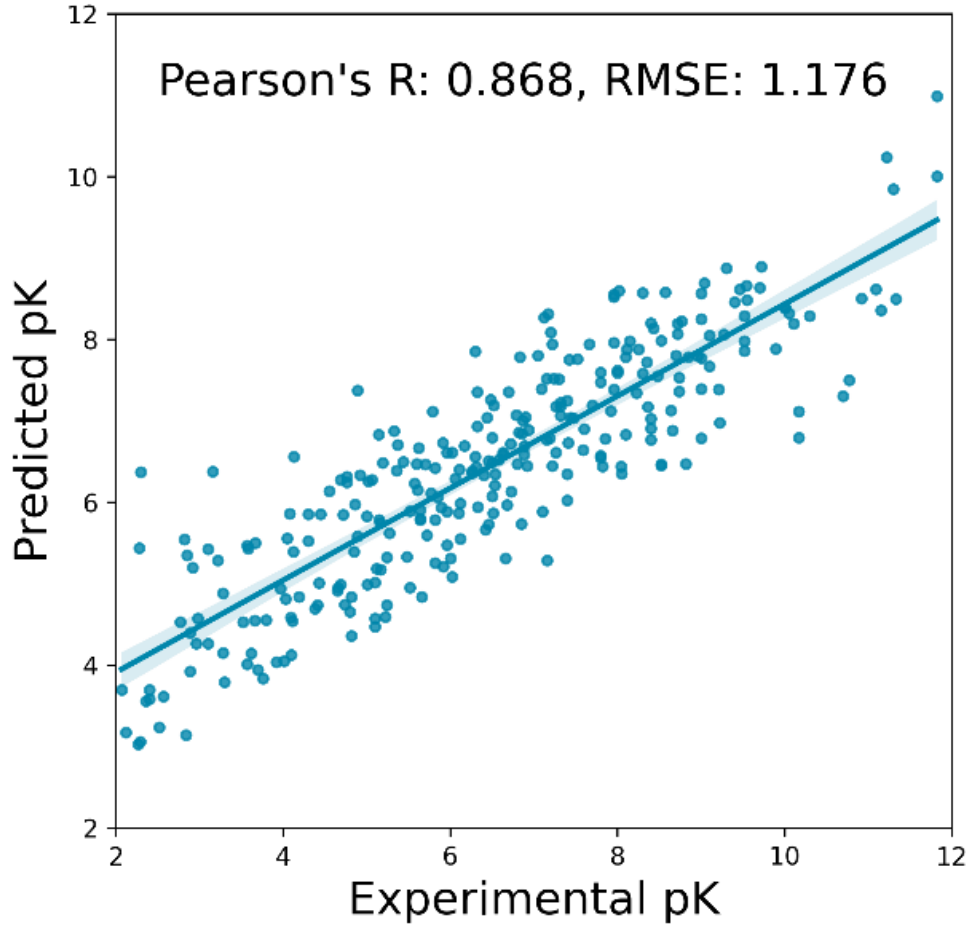


Figure 4. 6. The best result of weighted ECIF.

Scatterplot of experimental values of CASF-2016 'core set' (n=285) and predicted values by the best model of weighted ECIF.

As well as multi-shelled ECIF, We first investigated the hyperparameter of weighted ECIF features. Next, we examined the GBDT hyperparameters. We then trained 5000 models with different random seeds using the best conditions obtained above and examined the distribution of two metrics of CASF-2016 scoring power, Pearson's R and RMSE. The

best weighted ECIF model achieved a Pearson's R of 0. 868 and an RMSE of 1.176. (Figure 4.6) This is not as good as multi-shelled ECIF.

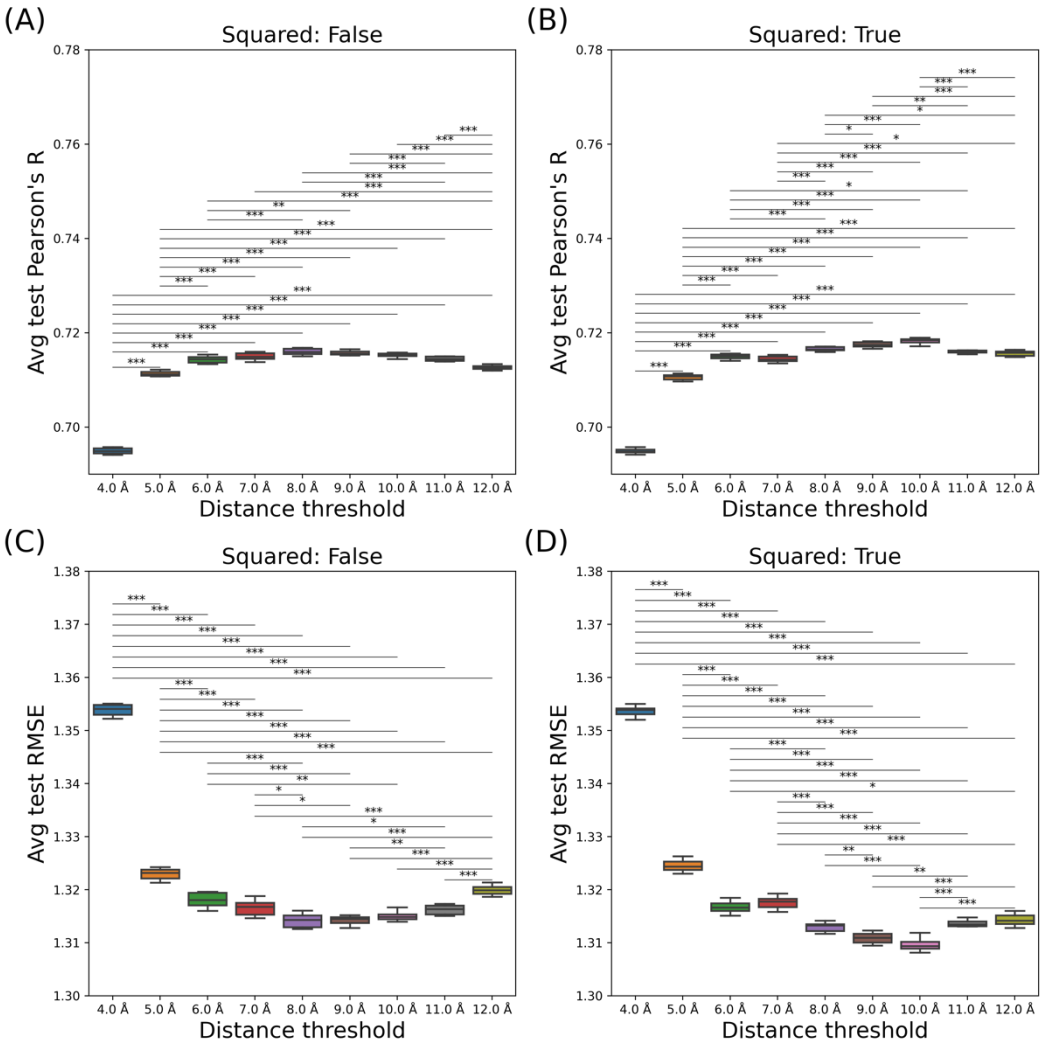


Figure 4. 7. Exploration of the optimal distance threshold of weighted ECIF.

Each boxplot represents the results of 10 runs of 10-fold cross-validation over the training set. The upper four panels (A, B) show the average Pearson's R of 10 runs, and the bottom four panels (C, D) show the average RMSE. The "squared" parameter is indicated at the top of each plot. The horizontal axis represents the distance threshold. Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

4.3.2.2 Exploration of the Weighted ECIF Feature Parameters

Weighted ECIF has two hyperparameters: distance threshold and squared. The distance threshold is the threshold to which extent P-L interactions are considered, and "squared" is the choice of whether the weights are assigned as the inverse of the inter-atomic distance or the inverse of the square of the inter-atomic distance. As well as multi-shelled ECIF, the conditions for the distance threshold and "squared" were examined by 10-fold cross-validation over the training set. 10-fold cross-validation was performed in each condition with 10 trials each with 10 different random seeds. Bonferroni-corrected independent t-tests were also used to check whether the differences between conditions were statistically significant. First, the search for the optimal distance threshold was conducted with "squared" fixed. To investigate the optimal distance threshold, we examined it under the conditions of 4.0~12.0 Å. When "squared" is false, the best performance is obtained when the distance threshold is 8.0 Å, and when "squared" is true, the best performance is obtained when the distance threshold is 10.0 Å (Figure 4.7). Next, we checked whether squared was True or False for the same distance threshold to see which performed better. The performance was significantly better when "squared" was True for all distance thresholds greater than 8 Å. (Figure 4.8)

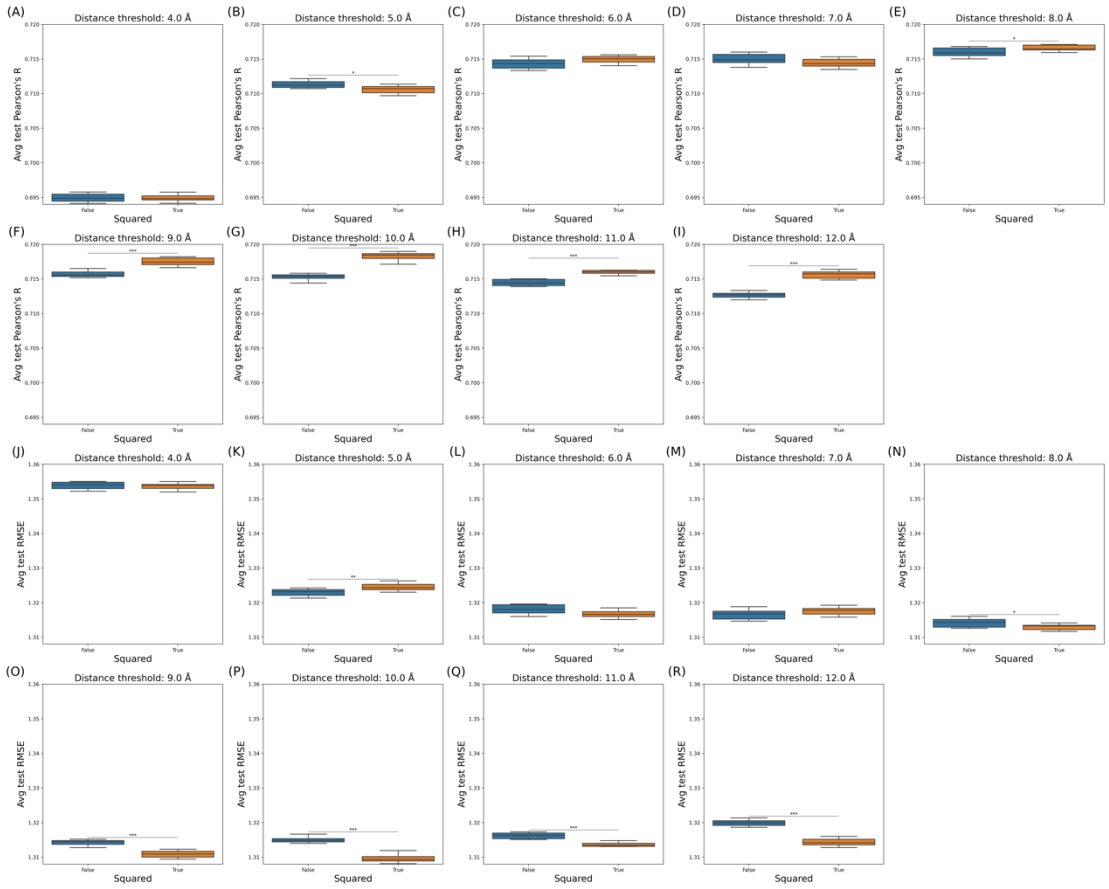


Figure 4. 8. Exploration of the optimal "squared" parameter of weighted ECIF.

Each boxplot represents the results of 10 runs of 10-fold cross-validation over the training set. The upper four panels (A~I) show the average Pearson's R of 10 runs, and the bottom four panels (J~R) show the average RMSE. The distance threshold is indicated at the top of each plot. The horizontal axis represents the "squared" parameter. Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

As a result, the distance threshold 10 Å, squared True had the best performance, thus these values were used. The results are shown in Figure 4.9.

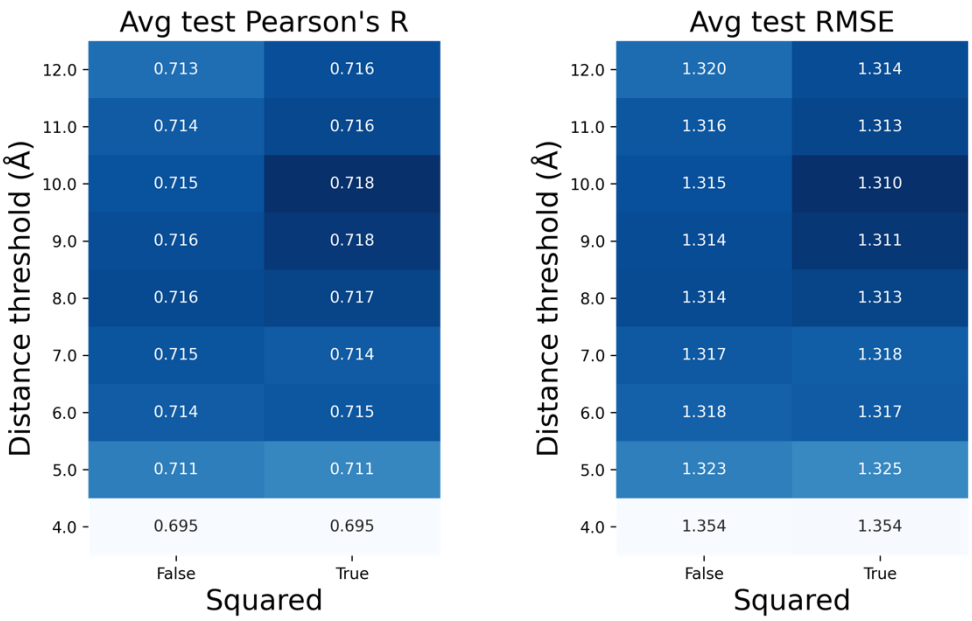


Figure 4. 9. Result of the weighted ECIF feature parameter exploration.

Each cell in the heatmap represents the average Pearson's R (left) and average RMSE (right) of 10 runs of 10 fold cross-validation under each condition. The horizontal axis of the heat map shows the "squared" parameter, and the vertical axis shows the distance threshold.

4.3.2.3 GBDT Parameters Optimization for Weighted ECIF

The distance threshold and "squared" of weighted ECIF were fixed at 10 Å and True respectively, and the hyperparameter of GBDT was optimized by 10-fold cross-validation over the training set. As well as the exploration of GBDT parameters for multi-shelled ECIF, 10-fold cross-validation was performed in each condition with 10 trials each with 10 different random seeds. The results are shown in Figure 4.10. The best GBDT hyperparameter for weighted ECIF is shown in Section 4.2.3.

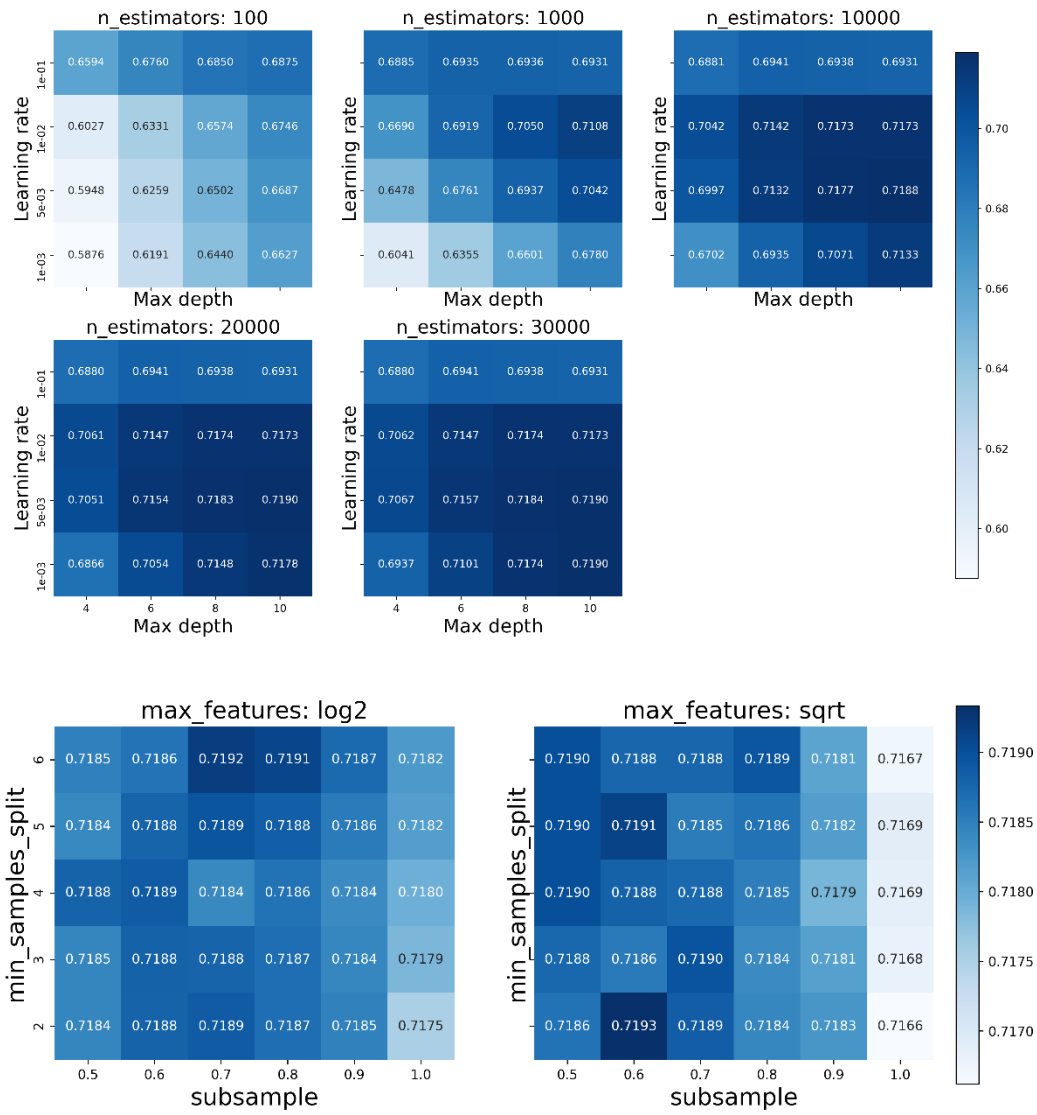


Figure 4. 10. Result of the GBDT parameters optimization of the weighted ECIF.

Each cell in the heat map represents the average Pearson's R of the 10 runs of 10 fold cross-validation trained under each condition. The horizontal axis of the heat map shows step width, and the vertical axis shows the distance threshold. The upper five panels show the optimization results for n_estimators, learning_rate, and max_depth, while the lower two panels show the optimization results for min_sample_split, max_features, and subsample.

4.3.3 Comparing Performance through Statistical Testing

To compare the quality of features rather than the performance of the best models, we trained 5000 models with different random seeds using the best parameters obtained above and compared their CASF-2016 Pearson's R and RMSE distributions for multi-shelled ECIF, weighted ECIF, and original ECIF. Then, a Bonferroni-corrected independent t-test was used to check whether the difference was statistically significant. The results showed that multi-shelled ECIF performed significantly better than the original ECIF and weighted ECIF in Pearson's R. (Figure 4.11) The t-statistic is 75.53, and the p-value is smaller than the smallest value that can be represented in Python (approximately 2.2e-308). The effect size for this difference was very large, with a Cohen's d of 1.51. On the other hand, the original ECIF was significantly lower than multi-shelled ECIF for RMSE. The Cohen's d was 1.08, indicating a large effect. In VS, ranking binding affinities is more important than accurately predicting binding affinity values. Therefore, multi-shelled ECIF with higher Pearson's R is more useful in VS. In addition to the difference between the averages of the 5000 models, the performance of the best models of each of the multi-shelled ECIF and ECIF was compared by Mann-Whitney U-test at 10,000 bootstrapped Pearson's R and RMSE. The Pearson's R of the best multi-shelled ECIF model is 0.877 and the RMSE is 1.152. The Pearson's R of the best ECIF model is 0.874 and the RMSE is 1.151. The result showed that the best multi-shelled ECIF model was significantly higher than the best ECIF model for Pearson's R. (p-value=4.9e-56) The effect size, as measured by Cliff's delta, was 0.128, indicating a small effect. On the other hand, no significant difference was found for RMSE (p-value=0.147). The Cliff's delta was 0.00856, indicating that there was almost no difference between the groups.

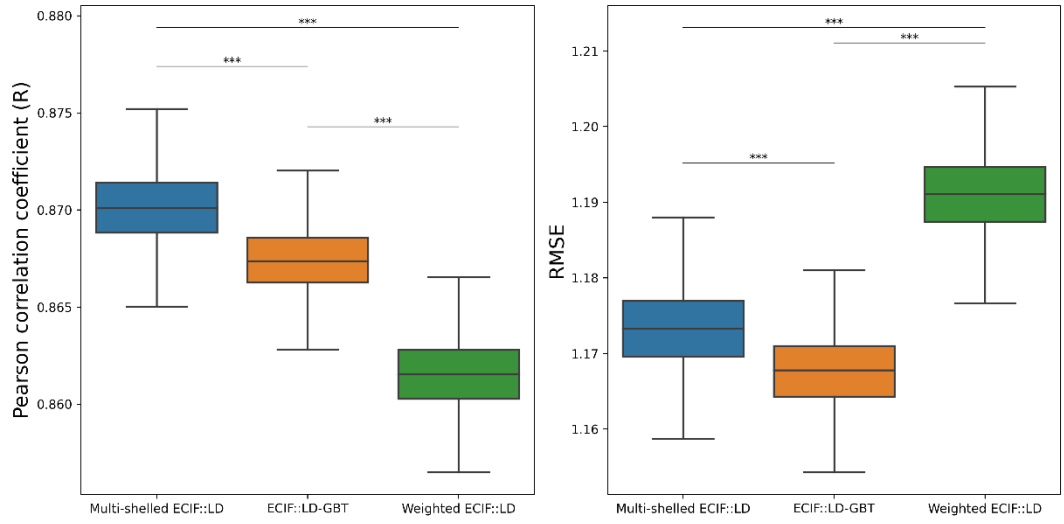
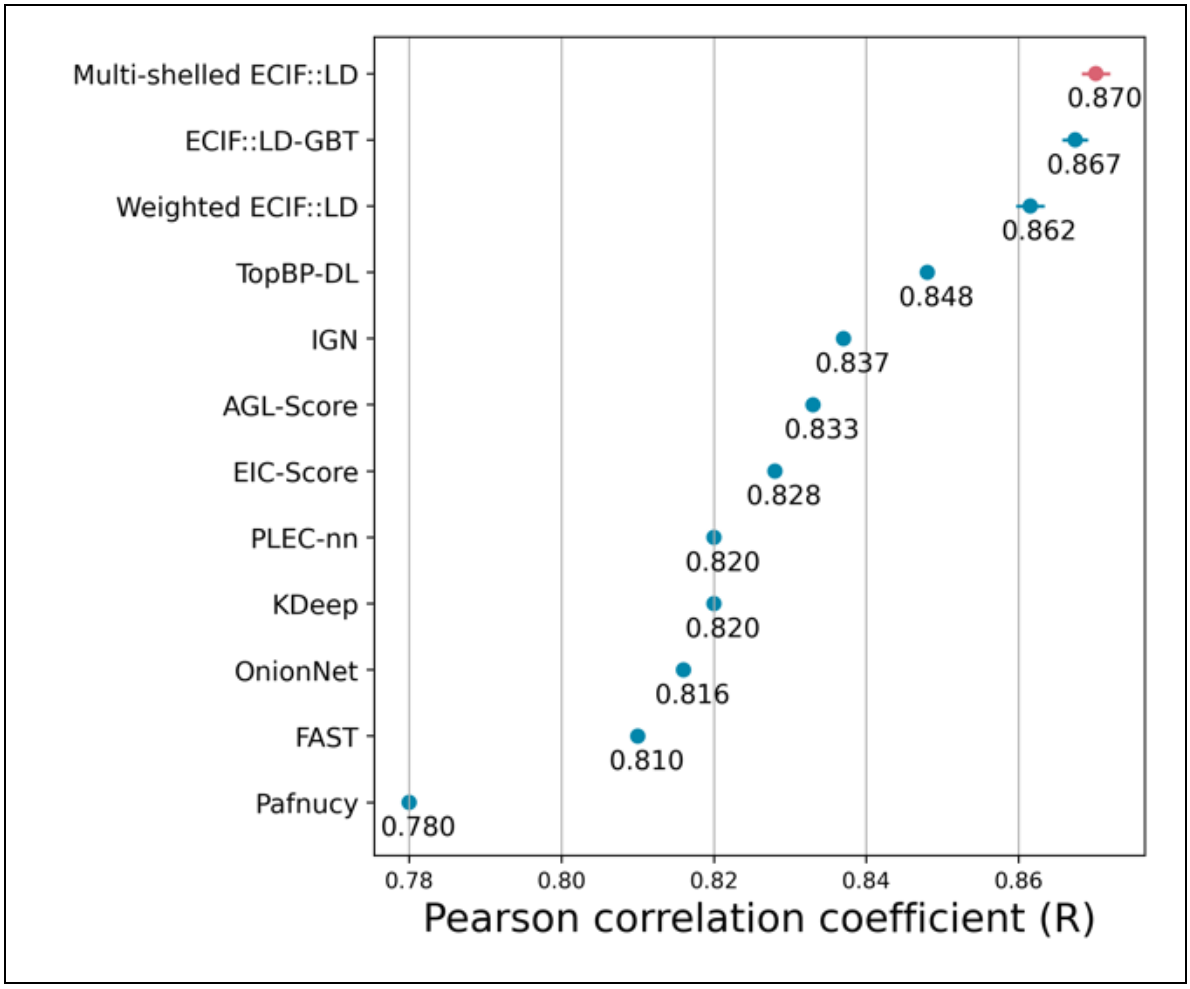


Figure 4. 11. Comparing performance through statistical testing.

Each boxplot represents the results of 5,000 models trained with different random seeds and evaluated with CASF-2016 'core set' Pearson's R (left) and RMSE (right). Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

4.3.4 Comparison with Other Reported Scoring Functions

A comparison of our results with the evaluation results by CASF-2016 of previously reported methods is shown in Figure 4.12. Multi-shelled ECIF achieved the best performance in terms of average Pearson's R for CASF-2016 scoring power among the methods reported to date. The results show that the distance consideration improves the performance of ECIF.


Multi-shelled ECIF::LD
ECIF::LD-GBT
Weighted ECIF::LD
TopBP-DL
IGN
AGL-Score
EIC-Score
PLEC-nn
KDeep
OnionNet
FAST
Pafnucy
0.870
0.867
0.862
0.848
0.837
0.833
0.828
0.820
0.820
0.816
0.810
0.780
0.78
0.80
0.82
0.84
0.86
Pearson correlation coefficient (R)

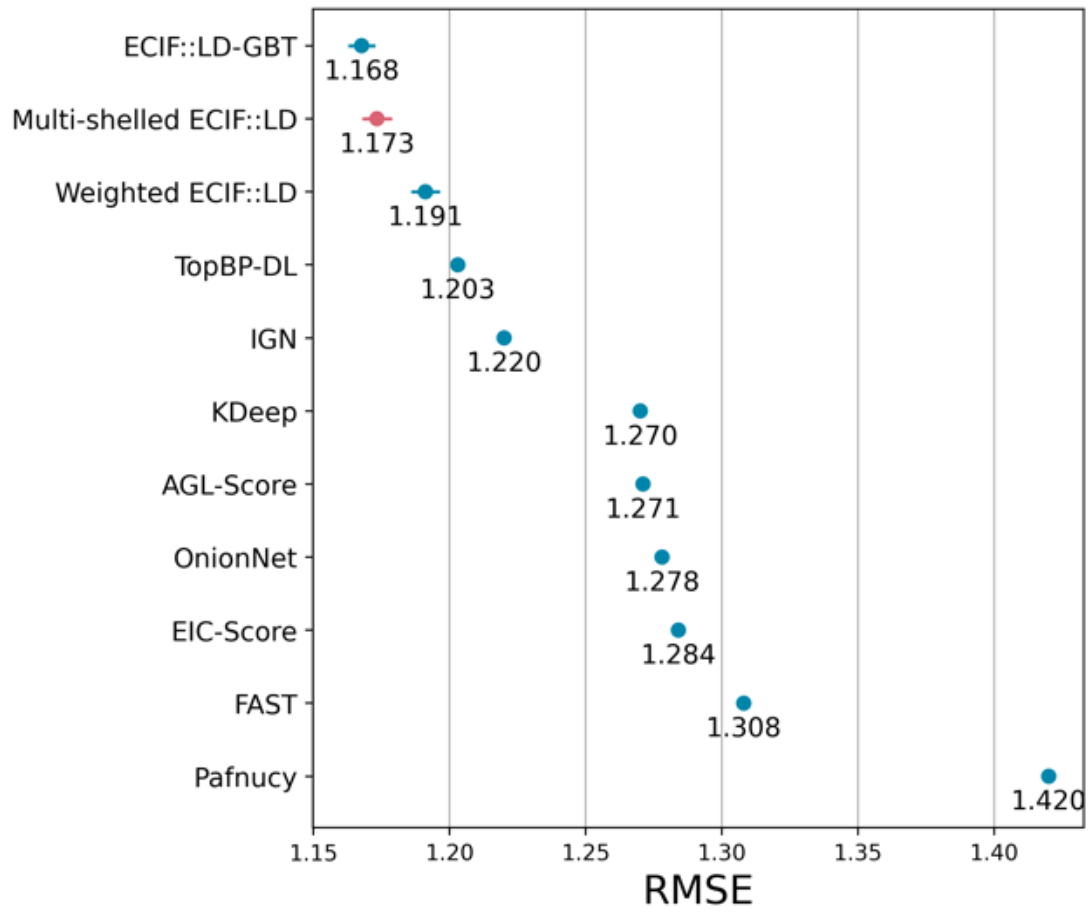


Figure 4. 12. Comparison with other reported scoring functions.

Comparison of reported evaluation results by CASF-2016 Pearson's R (left panel) and RMSE (right panel). Multi-shell ECIF results are highlighted in pink. For multi-shelled ECIF, weighted ECIF, and ECIF, the mean of 5000 models is displayed and the standard deviation is indicated by error bars. (Cang, et al., 2018; Jiang, et al., 2021; Jimenez, et al., 2018; Jones, et al., 2021; Nguyen and Wei, 2019; Nguyen and Wei, 2019; Sanchez-Cruz, et al., 2021; Stepniewska-Dziubinska, et al., 2018; Wojcikowski, et al., 2019; Zheng, et al., 2019)

4.3.5 Evaluation by Other CASF Dataset

A previous report on a modification to consider distance in ECIF was made by (Orhobor, et al., 2022). They defined four datasets, CASF-2007, CASF-2013, CASF-2016 and CASF-2019, for a comprehensive comparison between ECIF and their developed method, pair distance ECIF (PDECIF). It should be noted here that the training set differs between CASF-2016 as defined by Orhobor et al. and CASF-2016 by Sanchez-Cruz et al. mentioned above. As PDECIF is a similar method to our multi-shelled ECIF, we also evaluated multi-shelled ECIF on the four CASF datasets used by Orhobor et al. and compared their results. The results are presented in Table S4; as Orhobor et al reported all the results for PDECIF and ECIF for several distance hyperparameters, only the best ones for each dataset are extracted and cited. For those without ligand descriptors, multi-shelled ECIF showed the best Pearson's R for all four datasets. For those containing Ligand descriptors, multi-shelled ECIF::LD showed the best results, except for CASF-2013, where PDECIF::LD showed the best results. While Orhobor adjusted GBDT hyperparameters using CV for each dataset, we achieved great results with the above parameters without individual adjustments.

Table 4. 3 Evaluation result of other CASF datasets.

Those with the notation "::LD" include the ligand descriptor as input. Orhobor et al report the results of several distance hyperparameters for PDECIF and ECIF, so only the best from each dataset is cited. For each dataset, the best Pearson's R and RMSE for "+ ligand" and non "+ ligand" each are marked in bold.

dataset	representation	R	RMSE
CASF-2007	ECIF	0.812	1.472
CASF-2007	ECIF::LD	0.815	1.472
CASF-2007	Multi-shelled ECIF	**0.824**	**1.44**
CASF-2007	Multi-shelled ECIF::LD	**0.828**	1.423

CASF-2007	PDECIF	0.816	1.455
CASF-2007	PDECIF::LD	0.827	**1.418**
CASF-2013	ECIF	0.781	1.464
CASF-2013	ECIF::LD	0.801	1.419
CASF-2013	Multi-shelled ECIF	**0.803**	**1.442**
CASF-2013	Multi-shelled ECIF::LD	0.81	1.414
CASF-2013	PDECIF	0.792	1.449
CASF-2013	PDECIF::LD	**0.811**	**1.405**
CASF-2016	ECIF	0.815	1.32
CASF-2016	ECIF::LD	0.844	1.245
CASF-2016	Multi-shelled ECIF	**0.843**	1.28
CASF-2016	Multi-shelled ECIF::LD	**0.851**	**1.252**
CASF-2016	PDECIF	0.833	**1.277**
CASF-2016	PDECIF::LD	0.843	**1.252**
CASF-2019	ECIF	0.832	1.284
CASF-2019	ECIF::LD	0.856	1.211
CASF-2019	Multi-shelled ECIF	**0.868**	**1.199**
CASF-2019	Multi-shelled ECIF::LD	**0.875**	**1.171**
CASF-2019	PDECIF	0.854	1.235

CASF-2019	PDECIF::LD	0.859	1.217

4.3.6 Evaluation by LIT-PCBA Dataset

Validation was performed on the LIT-PCBA dataset to confirm that multi-shelled ECIF's performance is not due to a bias present in CASF-2016, but rather that it performs well against other datasets. Predictions were made using the best models of multi-shelled ECIF, weighted ECIF, and ECIF respectively for the 15 targets of LIT-PCBA and evaluated with enrichment factor (EF) of 1%. As a result, in 9 of 15 targets, multi-shelled ECIF performed as well or better than ECIF in terms of mean EF1%. (Figure 4.13) Results from CASF-2016 and LIT-PCBA show that multi-shelled ECIF, modified to account for interatomic distances, outperforms ECIF in VS. On the other hand, weighted ECIF outperformed original ECIF in 11 out of 15 LIT-PCBA targets in terms of mean EF1%, even though weighted ECIF was inferior to original ECIF for CASF-2016. Depending on the data set, weighted ECIF may be a worthwhile choice.

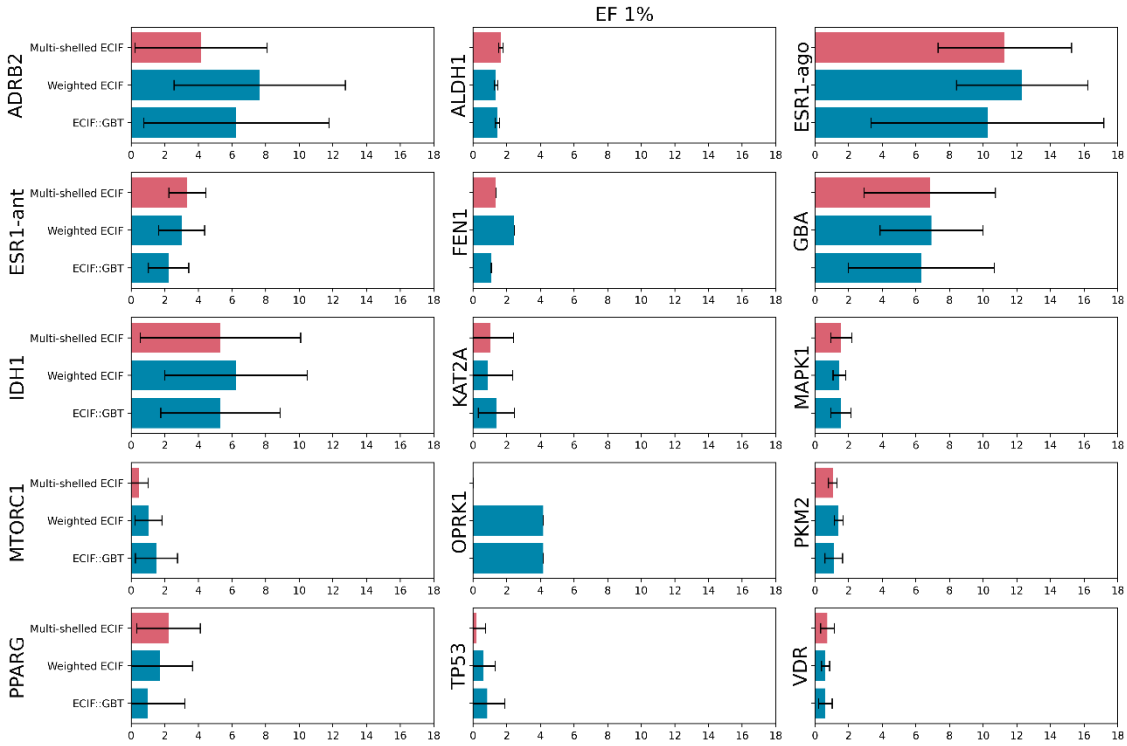


Figure 4. 13. Result of the evaluation by LIT-PCBA dataset.

The bars indicate the mean value of the Enrichment Factor 1% (EF1%) and the error bars indicate the standard deviation (SD). If only one template is provided for a given target (i.e., FEN1), the width of the error bar is 0.

4.3.7 Feature Importance

To investigate the performance improvement of adding our features to the ligand descriptors, we compared the performance of our model with that of a model trained with only ligand descriptors as features. (Figure 4.14) Ten models trained with only ligand descriptors were evaluated in CASF-2016, with an average Pearson's R of about 0.76, which is 0.1 lower than when multi-shell ECIF or weighted ECIF features were added (about 0.87). The results show that there is a significant performance improvement by adding our features.

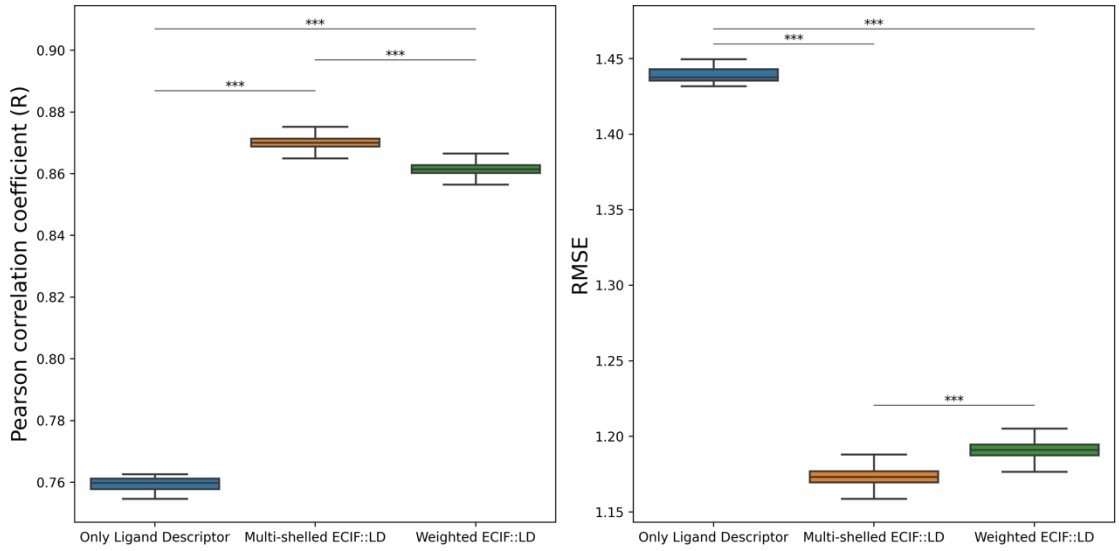


Figure 4. 14. Comparison with the models trained by only ligand descriptors.

Each boxplot represents the average Pearson's R (left) and average RMSE (right) of the models for each method trained with different random seeds and evaluated on the CASF 2016 "core set". The boxplot of "Only Ligand Descriptor" shows the average of 10 models and the boxplots of Multi-shelled ECIF and Weighted ECIF show the average of 5000 models. The vertical axis shows the distance range of the features. Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

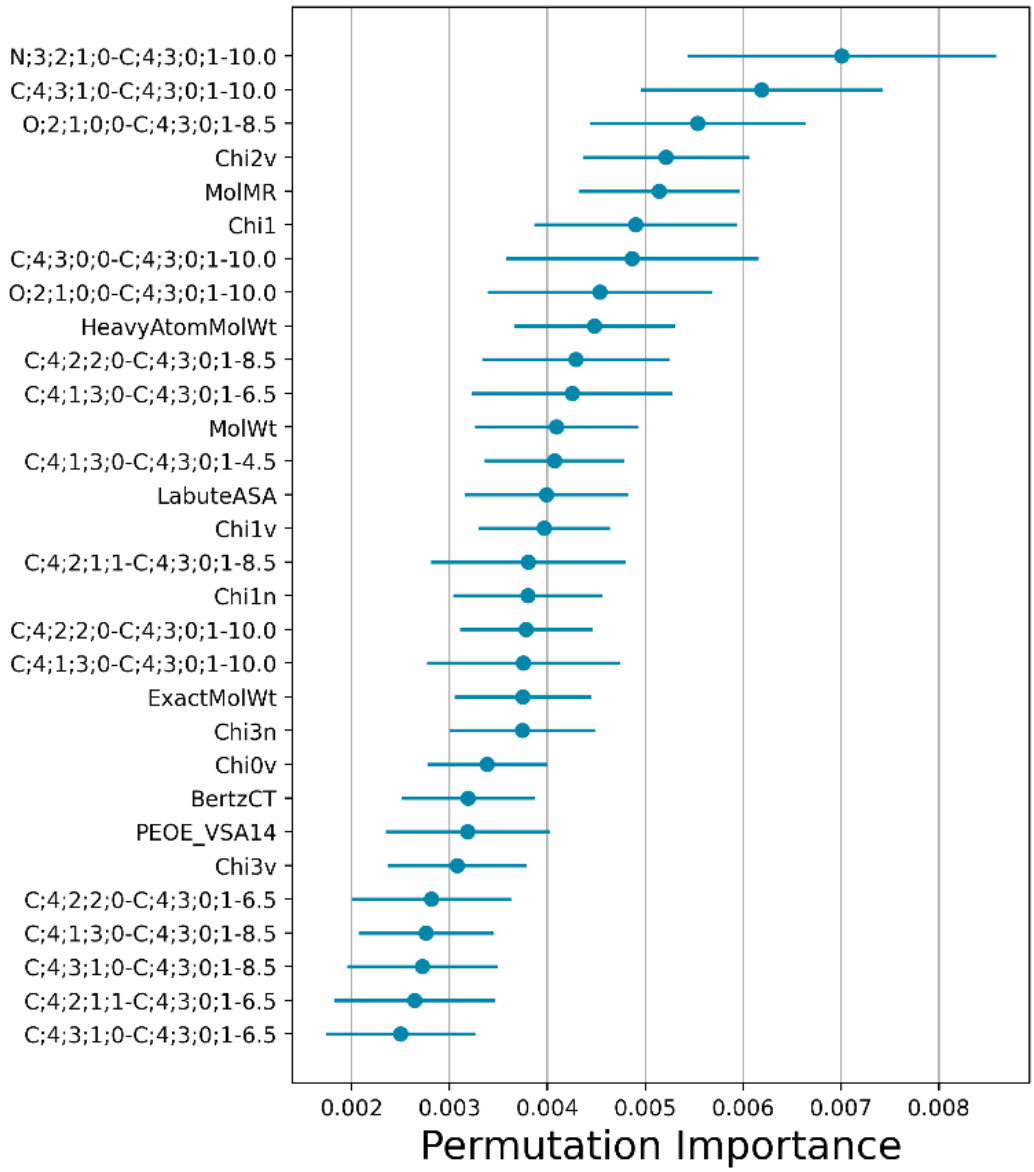


Figure 4. 15. Feature importance of the best multi-shelled ECIF model.

The top 30 permutation feature importance of the best model of multi-shelled ECIF is shown. All non-multi-shelled ECIF descriptors are ligand descriptors. For more details about each ligand descriptor, please refer to RDkit (https://www.rdkit.org/).

We conducted a feature importance analysis on the best model of multi-shelled ECIF. The top 30 features with the highest feature importance are shown in Figure 4.15. Multi-shelled ECIF features appear at the top of the list, indicating their substantial contribution to prediction. Of the 71 features in the top 1% of the total 7,100 features, 43 features were from multi-shelled ECIF. For the ligand side, those containing C;4;3;0;1, which are derived from aromatic rings, are in the top positions. For the protein side, N;3;2;1;0, O;2;1;0;0, and C;4;3;1;0, were found in the top positions. As for N;3;2;1;0;0 and O;2;1;0;0, they are derived from the peptide bond, while C;4;3;1;0, is derived from the alpha carbon. This finding suggests that the interaction between these atoms on the protein side and the aromatic ring of the ligand is crucial. Contrary to our intuition, the highest levels of permutation feature importance included many interactions at relatively distant distances of 10 Å and 8.5 Å. Features at 2~4 Å were predicted to be important, where hydrogen bonding and ionic bonding are dominant, but only 3 of the 71 features in the top 1% contained features of this range of distances. Therefore, to compare the importance of features by distance, we split the multi-shelled ECIF features by distance and compared the performance of models trained only with features at specific distances. Ligand descriptors were not used to simply compare the importance of the features only. We trained 10 models each with different random seeds in each condition and compared them with the average CASF-2016 Pearson's R and RMSE. As a result, the model trained with only 4.5~6.5 Å features had the best performance. (Figure 4.16) The results show that the features at distances 4.5~6.5 Å have the highest contribution to the prediction. Considering the permutation feature importance results, it seems that the 4.5~6.5 Å distance features play a major role in the prediction, with some particularly important interactions above 6.5 Å (such as the hydrophobic interaction with alpha carbon) making the prediction more accurate.

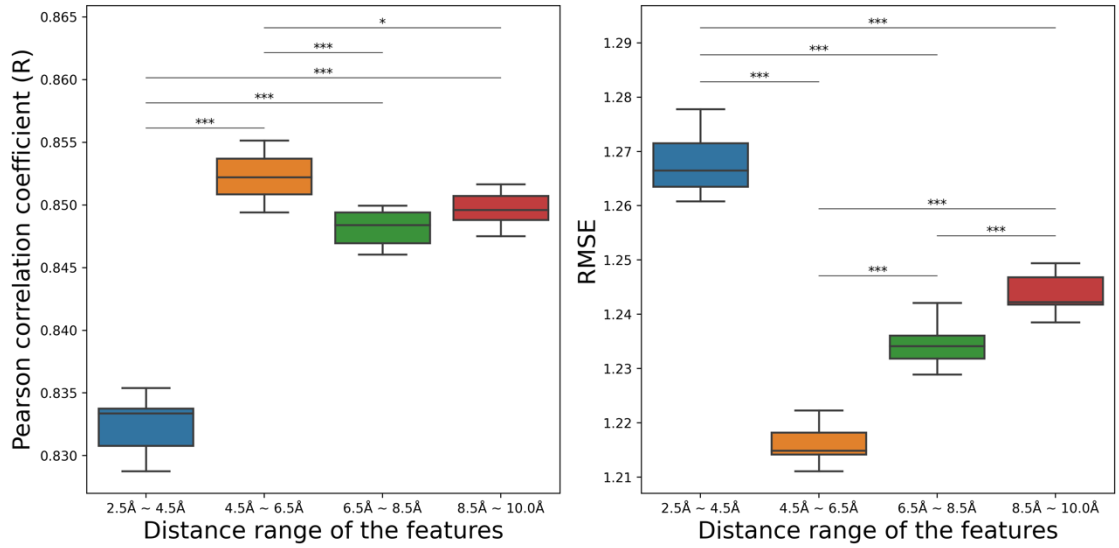


Figure 4. 16. Comparison with the models trained by the features at specific distance range.

Each boxplot represents the average Pearson's R (left) and average RMSE (right) of 10 models trained under each condition with different random seeds and evaluated with CASF2016 'core set'. The vertical axis shows the distance range of the features. Combinations that are statistically significant by Bonferroni-corrected independent t-test are marked with *. (*** : $p < 0.001$, ** : $0.001 <= p < 0.01$, * : $0.01 <= p < 0.05$.)

4.4. Summary

We have made several modifications to ECIF to allow it to account for distance. One is multi-shelled ECIF, in which several virtual shells are created by dividing the inter-atomic distance into several regions, and the count values of interactions in each shell are used as the feature. The other is weighted ECIF, in which the features are weighted sums of the squared inverse of the interatomic distances. The above two methods and the original ECIF were compared in terms of CASF-2016 scoring power. The results showed no improvement from the original ECIF for weighted ECIF, but significant improvement for multi-shelled ECIF. This indicates that the multi-shelled type is more effective in considering interatomic distances. For multi-shelled ECIF, a Pearson's R of 0.877 and RMSE of 1.152 were achieved in terms of CASF-2016 scoring power. Weighted ECIF was not as good as original ECIF in CASF-2016, but was superior to the original ECIF in the evaluation on the LIT-PCBA dataset. Multi-shelled ECIF is a method that can describe P-L interactions more precisely as the distance threshold is set farther away and as the step width is set smaller. As more experimental data become available in the future, it is expected that multi-shelled ECIF will be trainable using thresholds at farther distances and smaller step widths, thus further improving its performance. Our method is highly dependent on the number of hydrogen atoms and explicit valences. Therefore, the present method using an automatic procedure and standard protonation/tautomeric states has the limitations described in Section 4.2.1. Therefore, the preparation of P-L complexes that take into account the optimized structure of the hydrogen bonding network at the desired pH may improve the accuracy of the model.

Chapter 5

Conclusion and Future Work

In this dissertation, two novel studies in the field of protein-ligand (P-L) binding affinity prediction are presented.

The first study, "AQDnet: Deep Neural Network for P-L Docking Simulation," presents the development of AQDnet, a novel system for predicting P-L binding affinity using the three-dimensional structure of P-L complexes.

AQDnet stands out due to its ability to utilize the three-dimensional structures of P-L complexes and expand training datasets significantly by generating numerous ligand configurations. This is coupled with quantum chemistry computation to estimate the binding energy of these configurations. A key feature of AQDnet is the incorporation of the atom-centered symmetry function (ACSF), which is instrumental in learning the P-L quantum energy landscape (P-L QEL).

The system has shown exceptional results, achieving a top 1 success rate of 92.6% in the Comparative Assessment of Scoring Functions 2016 (CASF -2016) docking power benchmark, surpassing all other models evaluated. This remarkable performance underscores the model's capability in accurately predicting P-L interactions, a critical aspect in drug discovery and development.

The second study, "Multi-Shelled ECIF: Improved Extended Connectivity Interaction Features for Accurate Binding Affinity Prediction," tackles the shortcomings of Extended Connectivity Interaction Features (ECIF) in providing sufficient accounting for interatomic distances. The primary focus is on enhancing the Extended Connectivity Interaction Features (ECIF) methodology, particularly addressing its inadequacies in effectively capturing interatomic distances—a crucial element for accurate binding affinity prediction in molecular interactions.

The research introduces two innovative algorithms: multi-shelled ECIF and weighted ECIF, both aiming to refine ECIF's feature extraction process by integrating distance considerations. Among these, the multi-shelled ECIF emerges as particularly groundbreaking. It ingeniously segments the interatomic space into multiple layers or 'shells,' thus furnishing a nuanced and detailed atomic representation. This multi-layered approach does not merely add complexity but significantly enriches the informational depth of the interaction features, leading to a more accurate and reliable prediction of binding affinities.

Empirical results validate the superiority of the multi-shelled ECIF, as it notably surpasses both its weighted counterpart and the original ECIF framework. This is quantitatively evidenced by its impressive performance metric, achieving a Pearson correlation coefficient of 0.877 on the CASF-2016 scoring power benchmark. Such a high correlation coefficient is indicative of the model's precision and its potential as a robust tool in the drug discovery process, providing researchers with a more accurate understanding of molecular interactions.

In conclusion, this study marks a significant step forward in the field of computational drug discovery, with the multi-shelled ECIF setting a new benchmark for binding affinity prediction. The meticulous consideration of interatomic distances and the introduction of a layered analytical approach pave the way for more precise and reliable drug efficacy predictions, potentially accelerating the drug development process and opening new avenues for therapeutic innovation.

Both studies contribute significantly to the field of virtual screening (VS) and drug discovery, offering innovative methods for predicting P-L binding affinity with high

accuracy. The integration of quantum chemistry computation and enhanced feature extraction methods marks a significant advancement in the development of predictive models for P-L interactions, holding promise for future applications in drug discovery and development. The integration of AQDnet and multi-shelled ECIF into the drug discovery process is poised to significantly enhance the efficiency and accuracy of VS and binding affinity prediction. By advancing beyond traditional computational methods, these models offer a more refined and accurate approach to drug discovery.

The position of this work in the study of affinity prediction using machine learning is that it proposes novel methods that show state-of-the-art results in the Docking and Scoring tasks respectively. For AQDnet, in addition to the above, we propose a data extension method that can improve the performance of the Docking task. This will revolutionize the training of machine learning models for docking.

Next, the position of this research towards "complete virtual screening" is described below. Tentatively, 'complete virtual screening' is defined as 'virtual screening that identifies drug candidates by simulation alone, without relying on wet experiments'. In this case, there is currently no method that can achieve 'complete virtual screening' in the field of P-L binding affinity prediction. Even with the two methods proposed in this study, it is considered difficult to achieve 'complete virtual screening'. However, it is thought that there is sufficient potential to train models with sufficient prediction accuracy in practice by using experimental data obtained in actual drug discovery as new training data and fine-tuning existing models. Both of the two methods proposed in this study are designed for such an operation.

In the following, the contributions of this research to machine learning are described. We believe that the most significant contribution of this work to machine learning is that it shows that 'simulation-based labelling is effective' and that 'it is very important to generate features that contain information needed by the target to be learnt'.

A novel data extension method was proposed in AQDnet that can train neural networks with the results of quantum chemical calculations. We have also proposed a novel feature extraction method that is good at P-L interaction energy representation in order to train the data generated by this method. These two innovations have successfully trained the

Quantum Energy Landscape, resulting in a model capable of performing very accurate docking. This data extension method enables the labelling of the generated pose, which has been a major challenge in the field of P-L binding affinity prediction, by using the results of quantum chemical simulations. The significant improvement in docking performance of the models trained with this data extension method shows that the method is fully effective. This suggests that simulation-based labelling is very effective and has the potential to significantly improve the performance of a model when attempting to reproduce in machine learning the task that its simulation is trying to perform. AQDnet was an example of a QM simulation, which is not limited to the field of P-L binding affinity prediction, but could be transferable to many other fields (e.g. pharmacokinetic simulation).

However, we believe that the successful learning of the AQDnet was not only due to the fact that effective labelling was achieved. Another factor for the successful learning of this could be that "the features generated contain the information required by the target to be learnt (in this case, the results of quantum chemical calculations)". Quantum chemical calculations rely very strongly on small changes in interatomic distances, and in addition take into account not only two-body interactions but also many-body interactions involving three or more atoms. However, many machine learning methods reported in recent years fail to take these into account adequately. For example, 3DCNN, which uses a grid of 1 Å, and feature extraction by rounding inter-atomic distances using a shell with 0.5 Å increments cannot recognize small differences in inter-atomic distances, making it very difficult to learn the results of quantum chemical calculations. On the other hand, AQDnet uses features that can sensitively detect small changes in interatomic distances and many-body interactions and can take into account three-body interactions in the angular part in order to represent small changes in interatomic distances and many-body interactions. These results indicate that it is very important to generate features that contain sufficient information required by the target to be learnt, and this finding is an example worth referring to not only in the field of drug discovery, but also in many other fields.

For Multi-shelled ECIFs, the importance of incorporating information on interatomic distances into the features was demonstrated. Traditionally, chemists have placed great importance on the interatomic distances of P-L interactions when designing compounds that bind to proteins. The fact that incorporating this domain knowledge into the features

produced a model with high performance is another example of the critical importance of generating features that contain sufficient information required by the target to be learnt, similar to the findings obtained for AQDnet.

In the following, future work is discussed.

One key issue with AQDnet is the current limitation in the number of training data samples, which is essential for further enhancing the model's accuracy and generalizability. Future plans include expanding the training dataset with more complex structures to improve both scoring and screening power. Additionally, refining the model to take into account the energy differences between free and bound states of ligands is seen as a crucial step towards more precise affinity predictions. These advancements are expected to further solidify AQDnet's role as a groundbreaking tool in the field of computational drug discovery. This research represents a major leap forward in the application of deep neural networks to P-L docking and has promising implications for the future of pharmaceutical research.

The current issues with Multi-shelled ECIF are primarily to optimize the protonation state of the P-L complex structures used in the training data, to further enhance the feature extraction process, and to optimize the algorithm to increase computational efficiency without compromising prediction accuracy. Future plans include optimizing the treatment of protonated states of the P-L complex, integrating additional molecular features into the ECIF framework, exploring machine learning models that can complement the multi-shell approach, and validating the algorithm across more diverse data sets to ensure generality and robustness in a variety of molecular scenarios.

The following is applicable to both AQDnet and Multi-shelled ECIF. To better understand the practical implications of these models, future studies could focus on their application in real-world drug discovery scenarios. This might involve case studies where the models are applied to the discovery and development of new therapeutic molecules, analyzing the time and cost efficiency compared to traditional methods. Additionally, it would be beneficial to explore the models' adaptability and performance in predicting binding affinity for a wide range of target proteins and ligands, as this would provide a more comprehensive view of their potential impact on the pharmaceutical industry.

Publication List

Chapter 3 is based on a paper in *ACS Omega*:

AQDnet: Deep Neural Network for Protein–Ligand Docking Simulation

Koji Shiota, Akira Suma, Hiroyuki Ogawa, Takuya Yamaguchi, Akio Iida, Takahiro Hata, Mutsuyoshi Matsushita, Tatsuya Akutsu, and Masaru Tateno

ACS Omega, 2023 8 (26), 23925-23935

https://doi.org/10.1021/acsomega.3c02411

Chapter 4 is based on a paper in *Bioinformatics Advances*:

Multi-shelled ECIF: improved extended connectivity interaction features for accurate binding affinity prediction

Koji Shiota, Tatsuya Akutsu

Bioinformatics Advances, Volume 3, Issue 1, 2023, vbad155,

https://doi.org/10.1093/bioadv/vbad155

References

Abel, R.*, et al.* Advancing drug discovery through enhanced free energy calculations. *Accounts of Chemical Research* 2017;50(7):1625-1632.

Ajani, H.*, et al.* Superior performance of the SQM/COSMO scoring functions in native pose recognition of diverse protein-ligand complexes in cognate docking. *ACS Omega* 2017;2(7):4022-4029.

Bao, J., He, X. and Zhang, J.Z.H. DeepBSP-a machine learning method for accurate prediction of protein-ligand docking structures. *Journal of Chemical Information and Modeling* 2021;61(5):2231-2240.

Bartlett, R.J. and Musiał, M. Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics* 2007;79(1):291-352.

Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry* 2015;115(16):1032-1050.

Behler, J. and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters* 2007;98(14):146401.

Blay, V.*, et al.* High-throughput screening: today's biochemical and cell-based approaches. *Drug Discovery Today* 2020;25(10):1807-1821.

Cang, Z., Mu, L. and Wei, G.W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Computational Biology* 2018;14(1):e1005929.

Cavasotto, C.N., Adler, N.S. and Aucar, M.G. Quantum chemical approaches in structure-based virtual screening and lead optimization. *Frontiers in Chemistry* 2018;6:188.

Chemical Computing Group ULC. Molecular Operating Environment (MOE). 2023.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. p. 785-794.

Cheng, T.*, et al.* Comparative assessment of scoring Functions on a diverse test set. *Journal of Chemical Information and Modeling* 2009;49(4):1079-1093.

Desaphy, J.*, et al.* Encoding protein–ligand interaction patterns in fingerprints and graphs. *Journal of Chemical Information and Modeling* 2013;53(3):623-637.

Dhakal, A.*, et al.* Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions. *Briefings in Bioinformatics* 2022;23(1).

Feinberg, E.N.*, et al.* PotentialNet for molecular property prediction. *ACS Central Science* 2018;4(11):1520-1530.

Friedman, J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 2001;29(5):1189-1232, 1144.

Gao, X.*, et al.* TorchANI: A free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. *Journal of Chemical Information and Modeling* 2020;60(7):3408-3415.

Gilson, M.K. and Zhou, H.X. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure* 2007;36:21-42.

Gimeno, A.*, et al.* The light and dark sides of virtual screening: What Is there to know? *International Journal of Molecular Sciences* 2019;20(6).

Hansel, C.S.*, et al.* Advancing automation in high-throughput screening: Modular unguarded systems enable adaptable drug discovery. *Drug Discovery Today* 2022;27(8):2051-2056.

He, K.*, et al.* Deep residual learning for image recognition. In, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. p. 770-778.

He, X.*, et al.* A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics* 2020;153(11):114502.

Huang, N., Shoichet, B.K. and Irwin, J.J. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* 2006;49(23):6789-6801.

Hughes, J.P.*, et al.* Principles of early drug discovery. *British Jornal of Pharmacology* 2011;162(6):1239-1249.

Jain, A.N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *Journal of Computer-Aided Molecular Design* 2007;21(5):281-306.

Jiang, D.*, et al.* InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions. *Journal of Medicinal Chemistry* 2021;64(24):18209-18232.

Jimenez, J., *et al.* K(DEEP): Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of Chemical Information and Modeling* 2018;58(2):287-296.

Jones, D., *et al.* Improved protein-ligand binding affinity prediction with structure-based deep fusion inference. *Journal of Chemical Information and Modeling* 2021;61(4):1583-1592.

Ke, G., *et al.* LightGBM: A highly efficient gradient boosting decision tree. In, *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2017. p. 3149–3157.

Khalak, Y., *et al.* Alchemical absolute protein-ligand binding free energies for drug design. *Chemical Science* 2021;12(41):13958-13971.

Kimber, T.B., Chen, Y. and Volkamer, A. Deep Learning in virtual screening: recent applications and developments. *International Journal of Molecular Sciences* 2021;22(9).

King, E., *et al.* Recent developments in free energy calculations for drug discovery. *Frontiers in Molecular Biosciences* 2021;8:712085.

Koes, D.R., Baumgartner, M.P. and Camacho, C.J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling* 2013;53(8):1893-1904.

Kohn, W. and Sham, L.J. Self-consistent equations including exchange and correlation effects. *Physical Review* 1965;140(4A):A1133-A1138.

Kříž, K. and Řezáč, J. Reparametrization of the COSMO solvent model for semiempirical methods PM6 and PM7. *Journal of Chemical Information and Modeling* 2019;59(1):229-235.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. ImageNet classification with deep convolutional neural networks. In, *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc.; 2012. p. 1097–1105.

Kuznetsova, A.*, et al.* The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* 2020;128(7):1956-1981.

LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. *Nature* 2015;521(7553):436-444.

Lee, I., Keum, J. and Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLOS Computational Biology* 2019;15(6):e1007129.

Li, Y.*, et al.* Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of Chemical Information and Modeling* 2014;54(6):1717-1736.

Li, Y.*, et al.* Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *Journal of Chemical Information and Modeling* 2014;54(6):1700-1716.

Lindorff-Larsen, K., *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* 2010;78(8):1950-1958.

Liu, J. and Wang, R. Classification of current scoring functions. *Journal of Chemical Information and Modeling* 2015;55(3):475-482.

Liu, Z., *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 2015;31(3):405-412.

Lorenz, M.M. and Dejan, B. Novel trends in high-throughput screening. *Current Opinion in Pharmacology* 2009;9(5):580-588.

Maia, E.H.B., *et al.* Structure-based virtual screening: From classical to artificial intelligence. *Frontiers in Chemistry* 2020;8:343.

Marcou, G. and Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling* 2007;47(1):195-207.

McNutt, A.T., *et al.* GNINA 1.0: Molecular docking with deep learning. *Journal of Cheminformatics* 2021;13(1):43.

Moon, S., *et al.* PIGNet: A physics-informed deep learning model toward generalized drug-target interaction predictions. *Chemical Science* 2022;13(13):3661-3673.

Mysinger, M.M., *et al.* Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* 2012;55(14):6582-6594.

Natekin, A. and Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* 2013;7.

Nguyen, D.D. and Wei, G.W. AGL-Score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *Journal of Chemical Information and Modeling* 2019;59(7):3291-3304.

Nguyen, D.D. and Wei, G.W. DG-GL: Differential geometry-based geometric learning of molecular datasets. *International Journal for Numerical Methods in Biomedical Engineering* 2019;35(3):e3179.

Orhobor, O.I.*, et al.* A simple spatial extension to the extended connectivity interaction features for binding affinity prediction. *Royal Society Open Science* 2022;9(5):211745.

Ozturk, H., Ozgur, A. and Ozkirimli, E. DeepDTA: Deep drug-target binding affinity prediction. *Bioinformatics* 2018;34(17):i821-i829.

Pason, L.P. and Sotriffer, C.A. Empirical scoring functions for affinity prediction of protein-ligand complexes. *Molecular Informatics* 2016;35(11-12):541-548.

Pecina, A.*, et al.* SQM/COSMO scoring function at the DFTB3-D3H4 level: Unique identification of native protein-ligand poses. *Journal of Chemical Information and Modeling* 2017;57(2):127-132.

Pecina, A.*, et al.* The SQM/COSMO filter: Reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation. *Chemical Communications* 2016;52(16):3312-3315.

Pérot, S., *et al.* Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. *Drug Discovery Today* 2010;15(15-16):656-667.

Prokhorenkova, L., *et al.* CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* 2018;31.

Ragoza, M., *et al.* Protein-ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling* 2017;57(4):942-957.

Robert , T.M., *et al.* MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal* 2015;109(8):1528-1532.

Rohrer, S.G. and Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of Chemical Information and Modeling* 2009;49(2):169-184.

Russakovsky, O., *et al.* ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* 2015;115(3):211-252.

Sanchez-Cruz, N., *et al.* Extended connectivity interaction features: Improving binding affinity prediction through chemical description. *Bioinformatics* 2021;37(10):1376-1382.

Smith, J.S., Isayev, O. and Roitberg, A.E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* 2017;8(4):3192-3203.

Stepniewska-Dziubinska, M.M., Zielenkiewicz, P. and Siedlecki, P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics* 2018;34(21):3666-3674.

Su, M.*, et al.* Comparative assessment of scoring functions: The CASF-2016 update. *Journal of Chemical Information and Modeling* 2019;59(2):895-913.

Torng, W. and Altman, R.B. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling* 2019;59(10):4131-4149.

Tran-Nguyen, V.-K., Jacquemard, C. and Rognan, D. LIT-PCBA: An unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling* 2020;60(9):4263-4273.

Tran-Nguyen, V.K., Bret, G. and Rognan, D. True accuracy of fast scoring functions to predict high-throughput screening data from docking poses: The simpler the better. *Journal of Chemical Information and Modeling* 2021;61(6):2788-2797.

Tran-Nguyen, V.K., Jacquemard, C. and Rognan, D. LIT-PCBA: An unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling* 2020;60(9):4263-4273.

Tripos, L. Tripos Mol2 file format. *St. Louis, MO: Tripos* 2007.

Trott, O. and Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* 2010;31(2):455-461.

Vaswani, A., *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* 2017;30.

Wang, C. and Zhang, Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of Computational Chemistry* 2017;38(3):169-177.

Wang, R., *et al.* The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* 2004;47(12):2977-2980.

Wang, Z., *et al.* OnionNet-2: A convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Frontiers in Chemistry* 2021;9:753002.

Weibo, L., *et al.* A survey of deep neural network architectures and their applications. *Neurocomputing* 2017;234:11-26.

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 1988;28(1):31-36.

Wojcikowski, M., *et al.* Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* 2019;35(8):1334-1341.

Yang, C., Chen, E.A. and Zhang, Y. Protein-ligand docking in the machine-learning era. *Molecules* 2022;27(14).

Yu, W. and MacKerell, A.D. Computer-aided drug design methods. In: Sass, P., editor, *Antibiotics: Methods and Protocols*. New York, NY: Springer New York; 2017. p. 85-106.

Zhang, C.*, et al.* Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM* 2021;64(3):107-115.

Zhang, S.*, et al.* SS-GNN: A simple-structured graph neural network for affinity prediction. *ACS Omega* 2023;8(25):22496-22507.

Zhang, X.*, et al.* PLANET: A multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling* 2023.

Zheng, L., Fan, J. and Mu, Y. OnionNet: A multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction. *ACS Omega* 2019;4(14):15956-15965.

Zheng, X.*, et al.* Pocket-based drug design: Exploring pocket space. *An Official Journal of the American Association of Pharmaceutical Scientists* 2013;15(1):228-241.