BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/25 pp565–576
DOI: 10.26599/BDMA.2024.9020003
Volume 7, Number 3, September 2024

Exploring Fragment Adding Strategies to Enhance Molecule Pretraining in AI-Driven Drug Discovery

Zhaoxu Meng, Cheng Chen, Xuan Zhang, Wei Zhao*, and Xuefeng Cui*

Abstract: The effectiveness of AI-driven drug discovery can be enhanced by pretraining on small molecules. However, the conventional masked language model pretraining techniques are not suitable for molecule pretraining due to the limited vocabulary size and the non-sequential structure of molecules. To overcome these challenges, we propose FragAdd, a strategy that involves adding a chemically implausible molecular fragment to the input molecule. This approach allows for the incorporation of rich local information and the generation of a high-quality graph representation, which is advantageous for tasks like virtual screening. Consequently, we have developed a virtual screening protocol that focuses on identifying estrogen receptor alpha binders on a nucleus receptor. Our results demonstrate a significant improvement in the binding capacity of the retrieved molecules. Additionally, we demonstrate that the FragAdd strategy can be combined with other self-supervised methods to further expedite the drug discovery process.

Key words: pretraining; information retrieval; drug discovery; virtual screening; molecule property prediction

1 Introduction

Drug discovery is becoming increasingly costly[1, 2]. Research and development of a new medicine can cost anywhere from 944 million to 2826 million US dollars, which has increased exponentially in the last decade[3]. With the growing availability of big data, deep learning is a promising approach to accelerate drug discovery in areas such as compound synthesis, virtual screening, and de novo drug design[4–7]. However, the effectiveness of deep learning depends on the availability of labeled data, which is expensive, time-consuming, and sometimes impractical to obtain[8]. Pretraining can help address this issue by learning background knowledge from a large amount of unlabeled data[9], and this knowledge has been shown to significantly improve the performance of downstream tasks[10].

Recently, the masked language model approach has been widely utilized for pretraining small molecules[11–13]. Infomax[14] was one of the first graph pretraining methods to promote mutual information between local and global representations. Hu et al.[12] then implemented Mask in a molecular graph and discussed the advantages of using local and global-level tasks simultaneously. Grover[11] further advanced the Mask concept by proposing 1-hop Mask augmentation, which queries the model to predict artificial labels at local and graph levels. MolCLR[13] then implemented contrastive learning from computer vision and developed two deletion augmentation methods: bond deletion and subgraph removal, which can further corrupt the

• Zhaoxu Meng is with the School of Life Sciences, Shandong University, Qingdao 266237, China. E-mail: xmeng18@mail.sdu.edu.cn.
• Cheng Chen, Xuan Zhang, and Xuefeng Cui are with the School of Computer Science and Technology, Shandong University, Qingdao 266237, China. E-mail: cchen.bioinfo@mail.sdu.edu.cn; zhangxuan@mail.sdu.edu.cn; xfcui@email.sdu.edu.cn.
• Wei Zhao is with State Key Laboratory of Microbiology Technology, Shandong University, Qingdao 266237, China. E-mail: wei.zhao@sdu.edu.cn.
∗ To whom correspondence should be addressed.

Manuscript received: 2023-11-03; revised: 2023-12-17; accepted: 2024-01-08

molecule.

Although Mask-based pretraining methods have shown some success in small molecule deep learning, it is not ideal for small molecules due to two intrinsic properties: a limited vocabulary size and a non-sequential molecule structure. For example, molecules have a much smaller vocabulary size of less than 20, while university-level English speakers know approximately 10 000 word families on average[15]. If all the masked atoms in the molecules are predicted as carbon, an accuracy of approximately 74% (counted for 1 million molecules) can be achieved. This task is too easy, which prevents the pretraining method from learning useful information. Furthermore, unlike human language, where words are arranged sequentially, molecules have chemical structures that are essential to their properties[5, 16]. Applying a mask to chemical bonds does not cause any changes in the structures of molecules, whereas deleting bonds significantly modifies the properties of the molecule. Consequently, this obstacle prevents the pretraining method from gaining valuable knowledge.

In contrast to the existing pretraining strategies that involve reducing or eliminating information through the use of masks, we introduce a novel approach called FragAdd, which involves the addition of a chemically implausible molecular fragment to the input molecule. This strategy is intended to provide structural variation and prevent the collapse of the molecular structure. To learn rich local information while producing a meaningful molecular representation, we designed a series of experiments to explore how the adding strategy can be implemented. The fragments used in the strategy were taken from a fragment database created using pretraining data.

We conducted experiments to assess the downstream performance, components, and molecular representations of FragAdd, and to explore its potential to improve a drug discovery application. To compare FragAdd with other pretraining frameworks, we tested their ability on a benchmark of eight molecular properties. After validating four components of FragAdd, which can influence the diversity and difficulty of pretraining tasks, we examined whether the produced molecular representations can contain important chemical information. Furthermore, we explored how to apply our pretrained model in the virtual screening of small molecules for drug discovery. This work can motivate future research on the addition strategy for pretraining small molecules and also illustrates a possible application scenario in virtual screening.

2 Method

2.1 FragAdd framework

We created FragAdd to pretrain small molecules and use the pretrained model for downstream objectives such as property prediction and virtual screening. Pretraining provides Artificial Intelligence (AI) systems with a basic understanding of the data by learning the patterns in small molecule data[9]. As a small molecule pretraining framework, FragAdd introduces novel augmentation and training objectives to process molecule graphs and update parameters. After pretraining with unlabeled data, the model is further refined on supervised tasks, for example, predicting the toxicity of molecules.

Inspired by the modular nature of small molecules, FragAdd changes the molecular structure to provide diversity and avoids predicting molecular vocabulary to increase the difficulty of pretraining tasks. Diversity describes the number of chemical forms generated from the augmentation, and difficulty indicates how challenging the task is for an intelligent system to complete. Focusing on these two aspects may corrupt the molecules' structure to increase diversity and adjust the difficulty level by multiple operations. Molecules have a modular nature, which regards molecules as a collection of molecular fragments generated by addition reactions. Pharmacists use this idea to optimize the quality of drug candidates by adding or deleting parts from molecules. Based on this idea, FragAdd attaches a fragment outside the input molecule to imitate the process of the natural addition reaction.

During the augmentation process, FragAdd generates a chemically invalid fragment and adds it to the input molecule, as shown in Fig. 1. We generated a fragment database from all molecules in the pretraining dataset. To sample a fragment from the database, we designed a two-step approach: first, choosing a subgroup based on size (the number of atoms in a fragment[17]), and then randomly sampling one fragment from the chosen group (fragments larger than 20 atoms are placed into one group). Further, we corrupted the sampled fragment by atom mutation and ring break so as to

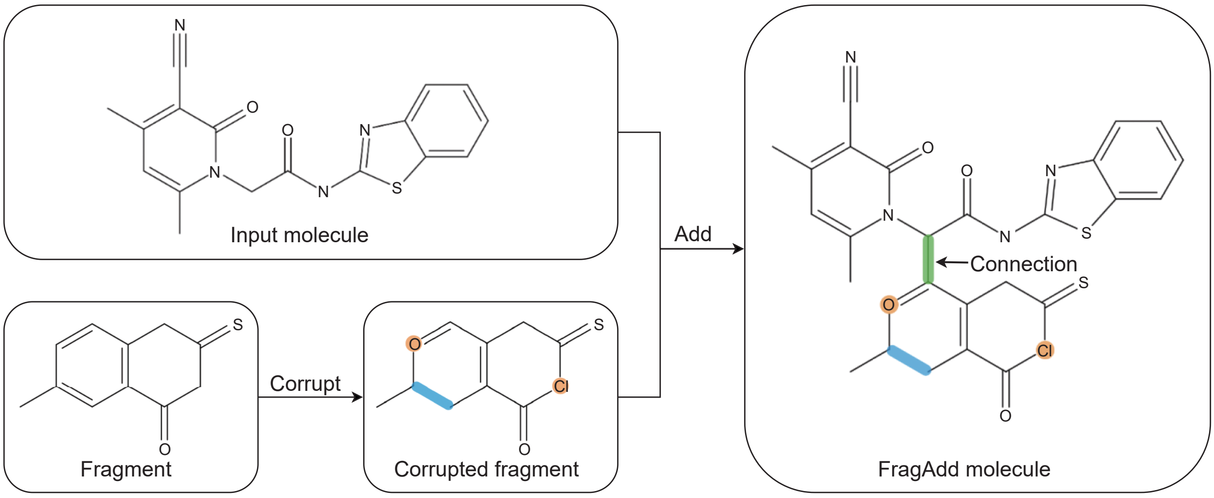


Fig. 1 Illustration of FragAdd pretraining. The molecule shown in the upper left corner serves as the initial input. A fragment is randomly chosen from the generated fragment database, as depicted in the lower left corner. This selected fragment is then corrupted by substituting highlighted orange atoms and removing highlighted blue bonds. To add the corrupted fragment to the input molecule, a carbon-carbon bond highlighted in green is utilized. The model is subsequently trained to identify the corrupted fragment at both local and global levels.

avoid problems in distinguishing the fragment from the original molecule. Atom mutation replaces some atoms with a different atom type, and ring break deletes a bond in a ring from a molecule if a ring exists. The ratio of mutation and break can be adjusted for a suitable difficulty level. To attach the damaged fragment to the input molecule, we connected two randomly sampled carbons from the two pieces. If no carbon exists for connection, atoms indexed zero in the molecule graph can be chosen. The FragAdd augmentation corrupts a dynamic region that depends on the size of the added fragment instead of a fixed local region by Mask-like methods.

For pretraining objectives, FragAdd locally classifies whether each atom belongs to the extra fragment while globally summing up the number of added atoms. In fact, previous work has proved the effectiveness of pretraining small molecules at both local and global levels[11, 12]. Locally, FragAdd predicts a binary classification for each atom so that the model learns to decompose molecules into fragments and determine which fragment is chemically unreasonable. Globally, FragAdd predicts the number of added atoms to summarize the chemical knowledge into molecular representation by pooling. Both levels of training objectives are vital for effective pretraining.

2.2 Data preparation

We transformed small molecules from SMILES to molecular graphs by computing node and edge features. SMILES is not exclusive to a single molecule and necessitates treating molecules as texts. We determined atom number and chirality as node features. Additionally, we incorporated bond features with chirality and bond type selected from single, double, triple, and aromatic. We hypothesized that atom type, bond type, and chirality are sufficient to differentiate one molecule from another.

The fragment database was generated from the pretraining dataset with the molecule decomposition algorithm BRICS. BRICS algorithm breaks molecules in positions where synthetic reaction could happen[18]. We implemented BRICS with RDKit[17], and by default, the decomposition process undergoes multiple rounds until no synthetically accessible bonds exist in fragments. We saved the output fragments in SMILES format to organize fragment data and removed duplicates. Additionally, the fragments were tagged with the number of atoms in that fragment, categorizing the database by fragment size.

2.3 Graph neural networks

A molecule graph is represented as $G=(V,E)$ with a node feature X_v for each $v \in V$. Graph Neural Networks (GNNs)[19] use a message-passing approach, where the representations of the neighboring nodes of node v are combined to iteratively update the representation of node v. After k rounds of aggregation, the representation of node v captures the structural information within its k-hop neighborhood. Formally, the k-th layer of a GNN is expressed as follows:

$$\begin{aligned} a_v^{(k)} &= \text{AGGREGATE}^{(k)}((h_v^{(k-1)}, h_u^{(k-1)}) : u \in \mathcal{N}(v)), \\ h_v^{(k)} &= \text{COMBINE}^{(k)}(h_v^{(k-1)}, a_v^{(k)}), \end{aligned} \quad (1)$$

where $h_v^{(k)}$ is the feature vector of node v at k-th layer, and $\mathcal{N}(v)$ is a set of neighbors of node v.

We implemented Graph Isomorphism Network (GIN)[20] as our model. GIN is the most expressive of the GNNs for the representation learning of graphs. Moreover, GIN uses Multi-Layer Perceptrons (MLPs) as the aggregation function, proving that it satisfies the conditions for a maximally powerful GNN. For the pretraining of molecular graphs, GIN is the most recognized architecture.

When setting up the GIN model, all hyperparameters stay the same as in the previous work to exclude the model’s influence during comparison. Five GIN layers were used to process molecule graphs. Nodes were embedded into 300-dimensional units, and no dropout are used. Only node features of the last layer were considered when model outputs and mean pooling were used to read out global representations. We used a linear layer to predict the training objective for all the pretraining tasks.

2.4 Training detail

We pretrained two million small molecules from the ZINC database for 100 epochs, and about 134 hundurds fragments were obtained using the BRICS algorithm. Instead of increasing the pretraining data size to achieve the best benchmark result, we kept the data size at two million molecules and conducted more rounds of exploration on adding strategies. We set the random seed to zero and the batch size to 256. The Adam optimizer was updated with a learning rate of 0.001; no weight decay or learning rate schedule was used to keep the system at a minimum. We included a ratio of 0.1 in our global training objective when combining the local and global loss.

The pretrained model was fine-tuned on eight classification datasets from MoleculeNet, and the batch size was reduced to 32. MoleculeNet classification datasets are the most accepted for small molecule property prediction, including three biophysics and five physiology datasets[21]. We added a dropout rate of 0.5 and reduced the batch size to 32 for small-size downstream tasks such as SIDER. Further, a linear layer was used to predict the final binary label and average the accuracy across all tasks for each dataset.

2.5 Virtual screening pipeline

We took Estrogen Receptor Alpha (ERα) binding data from the Nuclear Receptor Activity (NURA) dataset and divided it into reference and search data. The search data were then combined with two million molecules to form the final virtual screening dataset. NURA dataset contains information on small molecules that act as nuclear receptor modulators[22]. We obtained 1287 ERα binding active and 4861 inactive molecules from the 11 nuclear receptors of NURA. We sampled 20% of ERα data as reference data, which were used as a template for similarity search and fine-tuning. The other 80% ERα data were merged with two million small molecules from the ZINC database for screening. All weak active ERα binders were eliminated for simplicity.

We adjusted FragAdd on ERα reference data for the purpose of generating molecular representations. We set the batch size to 32, which is suitable for a small dataset, and fine-tuned the pretrained model for 30 epochs. To make the fine-tuning process easier, we excluded weak active data and only took into account absolute active or inactive data. During training, a linear layer was used to classify binding activity, and meaning pooling created molecular representations for similarity search.

We employed the Python library FAISS to carry out a molecular similarity search, utilizing embeddings from the GIN model and the Tanimoto coefficient to search for fingerprints. FAISS is a Python library for similarity searching and clustering of large-scale vectors[23]. The distance between molecular representations was calculated with the minimum Euclidean (L2) distance (the maximum inner product search could also be used). In this study, we chose the RDKit fingerprint and set the fingerprint size to 300, the same as the pretrained embedding. Additionally, the k-nearest fingerprint was defined by the Tanimoto coefficient, which is the ratio of the intersection of two vectors to the union of the two vectors.

We used AutoDock Vina (version 1.2.3), a widely used docking software for protein-ligand interaction[24], to investigate the interaction between unknown screening retrievals and ERα protein. To begin our analysis, we first created three-dimensional molecular structures with Open Babel[25]. We then carefully determined the center of the grid box, using the mean value of atoms coordinates within the binding pocket of ERα. This approach helped us to accurately define our docking search space, which was set to a dimension of 30 angstroms. Apart from these specific settings, we adhered to the default parameters provided by AutoDock Vina. Finally, we visualized the docking pose with Pymol and Discovery Studio[26, 27].

3 Result

3.1 Molecular property prediction benchmark

We compared FragAdd to other molecular machine learning frameworks by evaluating them on eight molecular property classification tasks[21], including Beta-secretase 1 Inhibition (BACE), Blood-Brain Barrier Permeability (BBBP), drug toxicity assessment (ClinTox, Tox21, and ToxCast), HIV inhibition (HIV), challenging virtual screening assays (MUV), and drug side effect profiling (SIDER). To ensure a fair comparison, we pretrained all the frameworks with the same model and training procedure. We used Area Under the Receiver Operating Characteristics (AUROC) as a metric and reported the mean and standard deviation of five fine-tune random seeds. For the setup of FragAdd, we set the probability of adding a fragment to the input molecule to 90%. Once a decision is made to add a fragment to the molecule, two subsequent modifications are independently performed on the fragment: each atom in the fragment has a 15% chance of undergoing mutation, and there is a separate 50% probability of breaking a ring within the fragment. These thresholds were set based on corresponding experiments. Several possibilities for mutation and ring breaking were tested. Excessively high rates of mutation and ring breaking could result in chemically implausible fragments, while excessively low rates might not generate sufficient diversity.

FragAdd achieves the best mean accuracy compared with Mask-like baselines, as shown in Table 1. From the distribution of bold best accuracy, each pretraining method has its areas of expertise. For example, Contextpred[12] performs the best in two datasets related to toxicity; MolCLR[13] and Grover[11] also excel in two pairs of tasks. Considering this accuracy pattern, evaluating the average performance across all eight datasets is crucial. FragAdd achieves the highest mean AUROC accuracy for the eight tested datasets, showing that the proposed adding strategy is comparable to previous best-performed baselines. For individual tasks, FragAdd surpasses all other baselines in two datasets that are highly related to drug discovery: BACE and ClinTox. BACE is a binary binding classification dataset for inhibitors of human β-secretase 1, and ClinTox includes FDA-approved drugs and drugs that have not passed clinical trials for toxicity reasons[29, 30]. We can infer that FragAdd has the potential to contribute to downstream applications in drug discovery.

3.2 Adding strategy exploration

In contrast to Mask and its variants, which only specify what to delete, the augmentation strategy of FragAdd presents more challenges and lacks approaches to address them. Mask deletes one feature for each chosen atom, while 1-hop Mask extends this range to the chosen atom and its 1-hop neighbors[11, 12]. This augmentation is straightforward and does not require additional thought. On the other hand, when it comes to adding strategy, questions such as what form should be added, how to connect the additional piece to the original molecule, and what training objective to use must be answered. These issues may be the reason why the default augmentation strategy for pretraining masks or hides something. Nevertheless, as discussed, for small molecules, adding strategy could be more

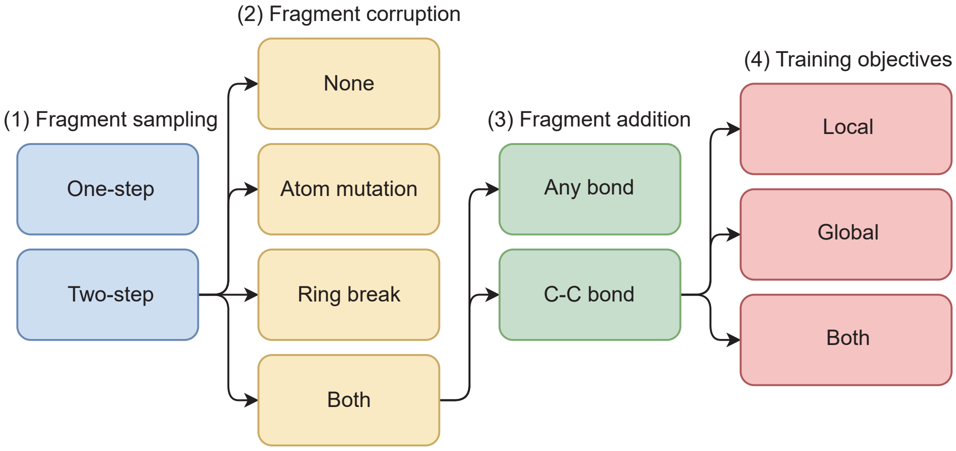
Table 1 AUROCs on eight molecular property classification datasets.

Method	AUROC accuracy (10^{-2})								
	BACE	BBBP	ClinTox	HIV	MUV	SIDER	Tox21	ToxCast	Average
Edgepred[28]	79.1 ± 2.2	71.1 ± 1.1	65.6 ± 1.4	76.9 ± 0.7	77.0 ± 1.8	61.8 ± 1.4	74.6 ± 0.7	62.6 ± 0.4	71.1
Infomax[14]	77.6 ± 1.0	70.1 ± 1.0	72.4 ± 1.2	78.4 ± 0.5	80.0 ± 0.9	59.4 ± 0.8	76.8 ± 0.5	63.8 ± 0.3	72.3
Contextpred[12]	81.4 ± 0.6	73.1 ± 0.7	70.1 ± 1.6	78.8 ± 0.2	78.8 ± 0.9	62.6 ± 0.4	**77.1 ± 0.2**	**64.7 ± 0.3**	73.3
Mask[12]	79.8 ± 0.8	71.4 ± 0.6	84.0 ± 1.1	78.9 ± 0.5	79.1 ± 1.4	60.2 ± 0.3	76.5 ± 0.3	64.1 ± 0.4	74.3
MolCLR[13]	80.5 ± 0.8	**74.2 ± 0.7**	79.4 ± 1.5	**79.6 ± 0.5**	79.1 ± 0.7	61.2 ± 0.7	75.8 ± 0.2	63.8 ± 0.2	74.2
Grover[11]	79.7 ± 0.4	69.6 ± 0.6	84.4 ± 2.0	79.2 ± 0.5	**80.4 ± 0.4**	**62.7 ± 0.2**	75.5 ± 0.3	64.4 ± 0.1	74.5
FragAdd	**84.8 ± 1.3**	72.3 ± 0.9	**85.1 ± 2.3**	77.8 ± 0.7	78.2 ± 1.2	62.4 ± 0.6	75.9 ± 0.6	63.7 ± 0.3	**75.0**

suitable for the requirements of molecular data and is worth exploring.

To determine how to implement the adding strategy which described in Section 2.1, we explored four components that influence the diversity and difficulty of augmentation on small molecules (Fig. 2a). At the beginning of the FragAdd augmentation process, a fragment should be sampled from the fragment database. However, the generated fragment database is unbalanced for fragment size (number of atoms in fragment), resulting in a decrease in diversity when sample from the databse directly (one-step sampling). For example, fragments with size less than 3 or larger than 20 have nearly no chance of being selected. Therefore, a better sampling method that tackles the unbalancing problem of fragment size can contribute to the diversity of corruption. For fragment corruption, how the fragment can be damaged to adjust the difficulty to a reasonable level needs to be explored. Additionally, it is crucial to choose the connection bond in fragment addition step. If most connection bonds are obvious wrong, the model will only need to break the bond to separate the molecule into two parts, which makes it too easy for the model to learn valuable molecular information. Finally, training objectives directly affect the difficulty of the pretraining tasks locally or globally.

Based on the benchmark, we found the best solutions for the four chosen components, shown in Fig. 2b. Compared with one-step sampling, first choosing fragment size substantially improves the accuracy, showing the importance of maintaining the fragment



(a) Design options of each component

Component option	AUROC accuracy (10^{-2})								
	BACE	BBBP	ClinTox	HIV	MUV	SIDER	Tox21	ToxCast	Average
(1) One-step	85.2 ± 0.9	72.9 ± 0.7	77.6 ± 2.2	77.7 ± 0.7	78.5 ± 1.1	61.0 ± 0.7	76.3 ± 0.5	63.4 ± 0.5	74.1
(1) Two-step	84.8 ± 1.3	72.3 ± 0.9	85.1 ± 2.3	77.8 ± 0.7	78.2 ± 1.2	62.4 ± 0.6	75.9 ± 0.6	63.7 ± 0.3	75.0
(2) None	83.9 ± 0.8	72.4 ± 0.7	78.3 ± 2.4	77.9 ± 0.8	80.0 ± 1.2	62.1 ± 0.6	75.8 ± 0.4	63.6 ± 0.5	74.3
(b) Atom mutation	84.3 ± 0.8	72.1 ± 0.7	79.0 ± 2.4	78.1 ± 0.7	79.0 ± 0.7	61.8 ± 1.0	75.7 ± 0.4	63.0 ± 0.2	74.1
(2) Ring break	85.0 ± 0.8	73.0 ± 0.6	79.6 ± 2.4	77.2 ± 0.9	79.1 ± 1.2	62.2 ± 0.6	76.0 ± 0.4	63.5 ± 0.3	74.5
(2) Both	84.8 ± 1.3	72.3 ± 0.9	85.1 ± 2.3	77.8 ± 0.9	78.2 ± 1.2	62.4 ± 0.6	75.9 ± 0.6	63.7 ± 0.3	75.0
(3) Any bond	84.7 ± 1.5	72.8 ± 0.5	78.2 ± 2.1	77.5 ± 0.6	79.8 ± 1.4	62.0 ± 0.4	76.5 ± 0.5	63.5 ± 0.3	74.4
(3) C-C bond	84.8 ± 1.3	72.3 ± 0.9	85.1 ± 2.3	77.8 ± 0.7	78.2 ± 1.2	62.4 ± 0.6	75.9 ± 0.6	63.7 ± 0.3	75.0
(4) Local	81.2 ± 1.2	73.8 ± 0.5	78.7 ± 3.2	79.7 ± 0.8	79.6 ± 1.4	60.0 ± 1.3	76.4 ± 0.5	63.0 ± 0.5	74.1
(4) Global	86.1 ± 0.7	72.0 ± 0.8	81.2 ± 2.3	77.7 ± 0.8	77.9 ± 0.9	61.6 ± 1.1	76.0 ± 0.5	62.9 ± 0.4	74.4
(4) Both	84.8 ± 1.3	72.3 ± 0.9	85.1 ± 2.3	77.8 ± 0.7	78.2 ± 1.2	62.4 ± 0.6	75.9 ± 0.6	63.7 ± 0.3	75.0

(b) AUROCs of each component option

Fig. 2 Exploration of FragAdd design options. (a) Design options of each component. The process consists of four sections. The first section uses a sampling technique to randomly choose a fragment from the fragment database. In the next section, the chosen fragment is deliberately altered to make it chemically invalid. The third section involves adding a chemical bond to connect the fragment with the input molecule. Finally, training objectives are set up to detect the corrupted fragment. The local objective predicts the atoms that are part of the corrupted fragment, while the global objective determines the number of atoms added to the input molecule. (b) AUROC of each design option. The final choice for each component of FragAdd is determined based on the results obtained.

size distribution normalized. For fragment corruption, atom substitution and ring scaffold hoping contribute independently to the invalid chemical information. Additionally, the carbon-carbon (C-C) bond proved to be a more effective choice for connecting fragments than any random bond. This superiority can be attributed to the high prevalence of C-C bonds in our pretraining dataset, where they constitute approximately 59% of all bonds in small molecules. Furthermore, carbon atoms in these molecules are typically connected to more than 1.05 hydrogen atoms on average, a higher connectivity compared to other atoms (e.g., “O”: 0.06, “N”: 0.32). This statistical prevalence of C-C bonds and the connectivity pattern of carbon atoms make the C-C bond attachment more chemically reasonable and effective for maintaining molecular integritye. Results also show that local and global training objectives are essential to pretrain performance, as they learn rich local information while producing a high-quality graph representation.

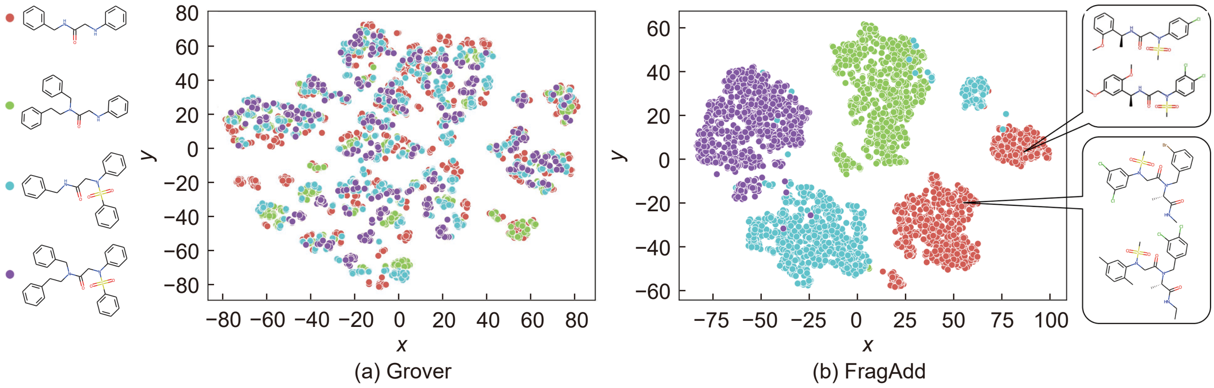
3.3 Visualization of molecular representation

We assessed whether molecular representations carry meaningful chemical information by visualizing embeddings from four structurally related scaffolds. To do this, we used t-SNE to plot the embeddings, expecting that molecules belonging to different scaffolds would be clustered[31]. Instead of selecting the most popular scaffolds in the dataset, we chose four structurally related scaffolds, as shown in Fig. 3a (popularity was determined by the number of times the scaffold appeared in the cross-pretraining dataset). The four chosen scaffolds only differ in the presence of one or two fragments, so this separation task requires more powerful extraction capabilities.

As opposed to Grover, FragAdd generates molecular representations that contain information capturing the slight difference between the scaffolds, as shown in Fig. 3. Grover fails to discriminate molecules by their scaffolds, and FragAdd separates four groups of points into clusters. The comparison shows that FragAdd learns the structure details about the existence of fragments in molecules. We also noticed that FragAdd generates subgroups under the same color, especially for scaffolds colored blue and red, which have subgroups far away in the t-SNE space. We further found that the subgroups significantly differ in side chains, showing that FragAdd can learn structural information deeper than the algorithm used to calculate the scaffold.

3.4 Application in virtual screening

We replaced the fingerprint method used in virtual screening with FragAdd and investigated whether it could help to retrieve more desired molecules from the screening database. Virtual screening is a common technique for the in silico development of new medicines[32–34], which searches for molecules with the highest probability of a particular property or activity in molecule libraries. To generate molecular representations with abundant chemical information, pretraining methods have been employed[35]. This approach is advantageous over the traditional fingerprint method, as it does not require the use of



(a) Grover

(b) FragAdd

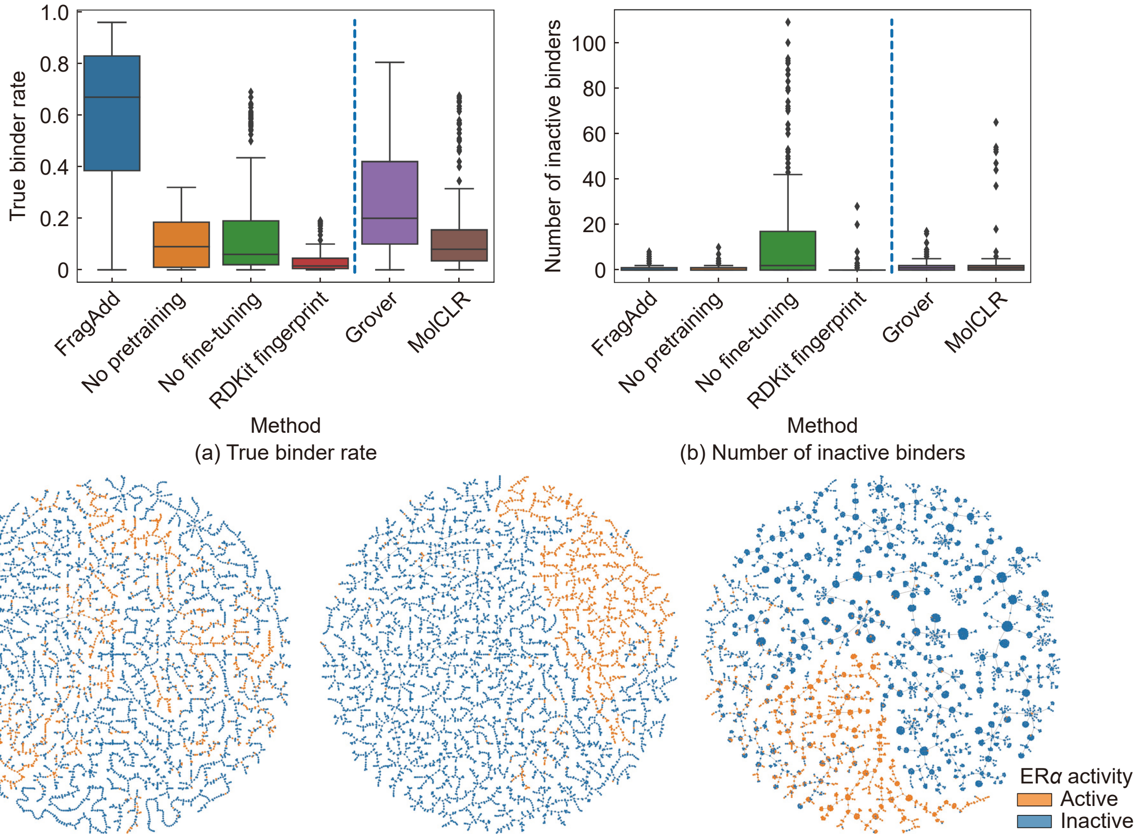
Fig. 3 Visualization (with t-SNE) of molecule representations pretrained by Grover and FragAdd. In this study, a total of four molecule scaffolds with similar characteristics were chosen and displayed on the left side. The main goal was to accurately segregate the molecules belonging to different scaffolds into distinct clusters. In the t-SNE space, Grover merged the four molecule categories, whereas FragAdd effectively differentiated the four scaffolds, forming clusters and even subgroups. The right panel demonstrates that even within the same scaffold, the subgroups exhibit noticeable variations in their side chains.

artificial rules to extract chemical information. Nevertheless, the application of pretraining in virtual screening has not been extensively studied.

We created a scenario to find molecules that bind to the estrogen receptor α (ERα) from the top-k output of a molecular similarity search. ERα is a crucial therapeutic target, especially considering that approximately 70% of breast cancer patients exhibit ERα positive status[36, 37]. Given this prevalence and the critical role of ERα in the disease's progression, our study focuses on this receptor to better understand its interactions and potential avenues for therapeutic intervention.

The dataset of ERα, comprising 6148 molecules, was split into reference and search subsets in a 1:4 ratio, and the search subset was combined with two million molecules to form the final search dataset. The reference subset was employed to fine-tune the model and served as the basis for reference molecules during the search process. We used a k-nearest neighbor search for each reference molecule, calculating the distance between molecular representations and setting k to 200. As most molecules in the search data do not have ERα binding activity labels, we used different methods to analyze known and unknown retrievals (known retrievals include molecules that have a binding label).

The analysis of known ERα ligands suggests that pretraining and fine-tuning are beneficial for virtual screening, as demonstrated in Fig. 4. FragAdd achieved the highest true binder rate for known binders and retrieved more than half of the true binders in the top 200 outputs for each reference molecule. The traditional fingerprint method was not successful in



(a) True binder rate

(b) Number of inactive binders

(c) Before fine-tuning

(d) After fine-tuning

(e) No pretraining

Fig. 4 Search engine (virtual screening) for estrogen receptor α binders. (a) True binder rate among the top 200 retrievals. (b) Number of known inactive binders among the top 200 retrievals. Box plots illustrate the qualities of search results by using various true binders as queries. (c) Visualization (with tmap) of ERα active and inactive binders with pretraining before fine-tuning. (d) Visualization with pretraining after fine-tuning. The promotion of the segregation of active and inactive binders can be observed through fine-tuning. (e) Visualization without pretraining. The results indicate that pretraining improves the quality of molecule representations at specific tmap coordinates, instead of grouping tmap coordinates. Therefore, by pretraining and fine-tuning FragAdd, a search engine for estrogen receptor α binders can be achieved.

retrieving enough true binders, which highlights the advantages of deep learning compared to the fingerprint method. We further explored the roles of pretraining and fine-tuning in virtual screening. Combining the true binder rate and inactive number results, we found that fine-tuning improves performance by decreasing the number of inactive binding molecules. To gain an intuitive understanding of the function of pretraining and fine-tuning, we visualized ERα data using tmap[38]. Comparing before and after fine-tuning reveals that fine-tuning helps classify active and inactive binding to reduce the inactive number. Without pretraining, many molecules mix with other ones instead of forming a tree structure, which indicates that pretraining assists in learning the chemical features of each molecule.

In contrast to known ERα ligands, the lack of binding activity labels for unknown retrievals makes it difficult to analyze them. To address this, we conducted a docking study to assess their binding to the ERα protein (Fig. 5). Docking is a computational technique used to predict protein-ligand interactions and binding affinity. We used the affinity gap to evaluate the binding of the unknown retrievals. FragAdd achieved the closest affinity gap to zero, indicating that it retrieves better unknown binders than the traditional fingerprint method. This confirms that both pretraining and fine-tuning are essential for unknown retrievals. To further understand the affinity gap result, we visualized the docking pose of a high-affinity unknown retrieval, ZINC1627292. The molecule interacts with the protein target through two hydrogen bonds on either side of the molecule and a T-shaped stacking between benzene rings. Of the three interactions, the hydrogen bond with His524 and the Pi-Pi interaction with Phe404 are conserved in the natural binders for ERα. For both known and unknown retrievals, FragAdd increases the number of potential binders in the top 200 outputs.

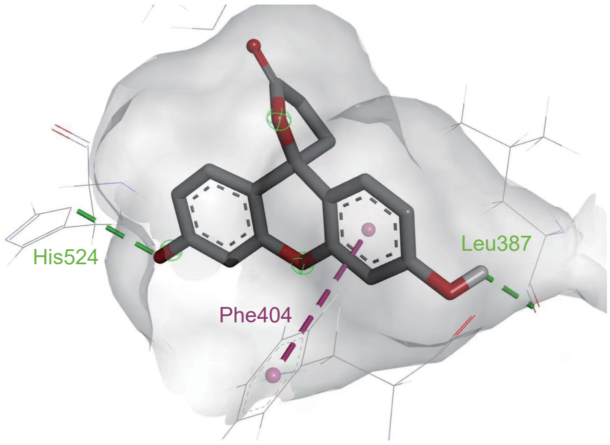
3.5 Combination of FragAdd with other methods

FragAdd preserves the original molecule component, thus allowing for the integration of other augmentation techniques. As an addition approach, FragAdd only adds a bond to one carbon atom in the original molecule; this means that FragAdd is compatible with Mask and its derivatives, raising the question of whether FragAdd can be combined with other methods. If it can, FragAdd will offer a new choice for other pretraining frameworks.

FragAdd improves the average performance added to other methods, indicating that the adding and deleting strategies could be used simultaneously. To implement this idea, we conducted Mask-like augmentation on the input molecule and then attached a fragment to the masked molecule and added the two loss items. We tested this operation for Infomax, Atom Mask, and Bond Mask. Bond Mask hides bond types for some

Method	MAE
FragAdd	0.022
No pretraining	0.137
No fine-tuning	0.148
RDKit FP	0.336

(a) Binding affinity MAE



(b) Docking case study

Fig. 5 Docking evaluation between ERα protein and novel binders discovered by FragAdd. (a) Mean Absolute Error (MAE) between binding affinities of the true binder and the novel binder. Hence, the novel binders uncovered by FragAdd exhibit comparable binding affinities to established binders. (b) Case study on ERα protein PDB:1X7E and novel ligand ZINC:1627292. Hydrogen bonds were represented by green lines, and T-shaped Pi-Pi stacking was represented by pink lines. Previous studies have shown that His524 and Phe404 act as inherent binding sites for ERα. The observation of hydrogen bonds and Pi-Pi interactions at these sites suggests that the newly discovered molecule by FragAdd could potentially bind to ERα.

bonds inside the molecular graph. For Infomax and Atom Mask, FragAdd improves more than 1% accuracy after being combined with FragAdd. Moreover, for Bond Mask, the accuracy stays the same, showing that it is better to adjust the ratio of loss items for the best combination performance.

4 Conclusion

We propose a pretraining framework, FragAdd, which uses fragments from decomposition as an additional part of an adding strategy, as an alternative to the Mask-based strategy in small molecule pretraining. Our results show that FragAdd outperforms previous baselines in molecular property prediction and virtual screening tasks. It achieved the best average accuracy in eight classification datasets, and excelled in two datasets related to drug discovery. This performance is attributed to the extraction of molecular representations that capture structure details. We also found that both pretraining and fine-tuning are essential for virtual screening, and that FragAdd can be used in conjunction with other self-supervised methods.

A pretrained model based molecule search engine has the potential to greatly accelerate the drug discovery process. However, we have noticed that FragAdd occasionally incorporates excessive structural variations, resulting in a bias during subsequent virtual screening. Additionally, the training of FragAdd has utilized the same model and dataset as previous studies, which might not be adequate for achieving optimal performance. Currently, we are focusing on developing a dependable molecule search engine that can cater to the specific requirements of biomedical research.

Acknowledgment

This work was supported by the National Key R&D Program of China (Nos. 2019YFA0905700 and 2021YFC2101500), and the National Natural Science Foundation of China (No. 62072283).

References

[1] H. F. Lynch and C. T. Robertson, Challenges in confirming drug effectiveness after early approval, *Science*, vol. 374, no. 6572, pp. 1205–1207, 2021.

[2] M. Schlander, K. Hernandez-Villafuerte, C. Y. Cheng, J. Mestre-Ferrandiz, and M. Baumann, How much does it cost to research and develop a new drug? A systematic review and assessment, *PharmacoEconomics*, vol. 39, no. 11, pp. 1243–1269, 2021.

[3] S. Simoens and I. Huys, R&D costs of new medicines: A landscape analysis, *Front. Med.*, vol. 8, p. 760762, 2021.

[4] Q. Jiao, Z. Qiu, Y. Wang, C. Chen, Z. Yang, and X. Cui, Edge-gated graph neural network for predicting protein-ligand binding affinities, in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine* (*BIBM*), Houston, TX, USA, 2021, pp. 334–339.

[5] H. Beck, M. Härter, B. Haß, C. Schmeck, and L. Baerfacker, Small molecules and their impact in drug discovery: A perspective on the occasion of the 125th anniversary of the Bayer Chemical Research Laboratory, *Drug Discov. Today*, vol. 27, no. 6, pp. 1560–1574, 2022.

[6] Y. Ye, Unleashing the power of big data to guide precision medicine in China, *Nature*, vol. 606, no. 7916, pp. 49–51, 2022.

[7] Y. Wang, Z. Qiu, Q. Jiao, C. Chen, Z. Meng, and X. Cui, Structure-based protein-drug affinity prediction with spatial attention mechanisms, in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine* (*BIBM*), Houston, TX, USA, 2021, pp. 92–97.

[8] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, Self-Supervised Representation Learning: Introduction, advances, and challenges, *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 42–62, 2022.

[9] Y. LeCun and I. Misra, Self-supervised learning: The dark matter of intelligence, https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/, 2021.

[10] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, and J. Pei, Transfer learning for drug discovery, *J. Med. Chem.*, vol. 63, no. 16, pp. 8683–8694, 2020.

[11] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, Self-supervised graph transformer on large-scale molecular data, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Virtual Event, 2020, pp. 12559–12571.

[12] W. H. Hu, B. W. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, Strategies for pre-training graph neural networks, presented at Int. Conf. Learning Representations (ICLR), Virtual Event, 2020.

[13] Y. Wang, J. Wang, Z. Cao, and A. Barati Farimani, Molecular contrastive learning of representations via graph neural networks, *Nat. Mach. Intell.*, vol. 4, no. 3, pp. 279–287, 2022.

[14] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, Deep graph Infomax, presented at Int. Conf. Learning Representations (ICLR), Vancouver, Canada, 2018.

[15] J. Milton and J. Treffers-Daller, Vocabulary size revisited: The link between vocabulary size and academic achievement, *Appl. Linguist. Rev.*, vol. 4, no. 1, pp. 151–172, 2013.

[16] X. Zhang, C. Chen, Z. Meng, Z. Yang, H. Jiang, and X. Cui, CoAtGIN: Marrying convolution and attention for graph-based molecule property prediction, in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine* (*BIBM*), Las Vegas, NV, USA, 2022, pp. 374–379.

[17] G. Landrum, RDKit: Open-source cheminformatics, https://www.rdkit.org, 2023.

[18] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey, On the art of compiling and using 'drug-like' chemical fragment spaces, *ChemMedChem*, vol. 3, no. 10, pp. 1503–1507, 2008.

[19] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, Gated graph sequence neural networks, presented at Int. Conf. Learning Representations (ICLR), San Juan, Puerto Rico, 2016.

[20] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, How powerful are graph neural networks? presented at Int. Conf. Learning Representations (ICLR), Vancouver, Canada, 2018.

[21] Z. Wu, B. Ramsundar, E. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, MoleculeNet: A benchmark for molecular machine learning, *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018.

[22] C. Valsecchi, F. Grisoni, S. Motta, L. Bonati, and D. Ballabio, NURA: A curated dataset of nuclear receptor modulators, *Toxicol. Appl. Pharmacol.*, vol. 407, p. 115244, 2020.

[23] J. Johnson, M. Douze, and H. Jégou, Billion-scale similarity search with GPUs, *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[24] O. Trott and A. J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function. efficient optimization, and multithreading, *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.

[25] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminf.*, vol. 3, no. 1, p. 33, 2011.

[26] W. L. DeLano, PyMOL: An open-source molecular graphics tool, *CCP4 Newsletter On Protein Crystallography*, vol. 40, no. 1, pp. 82–92, 2002.

[27] Dassault Systèmes, BIOVIA discovery studio visualizer, https://www.3ds.com, 2023.

[28] W. Hamilton, Z. T. Ying, and J. Leskovec, Inductive representation learning on large graphs, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 1025–1035.

[29] G. Subramanian, B. Ramsundar, V. Pande, and R. A. Denny, Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches, *J. Chem. Inf. Model.*, vol. 56, no. 10, pp. 1936–1949, 2016.

[30] K. M. Gayvert, N. S. Madhukar, and O. Elemento, A data-driven approach to predicting successes and failures of clinical trials, *Cell Chem. Biol.*, vol. 23, no. 10, pp. 1294–1301, 2016.

[31] G. Hinton and S. Roweis, Stochastic neighbor embedding, in *Proc. 15th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2002, pp. 857–864.

[32] A. A. Sadybekov, A. V. Sadybekov, Y. Liu, C. Iliopoulos-Tsoutsouvas, X. P. Huang, J. Pickett, B. Houser, N. Patel, N. K. Tran, F. Tong, et al., Synthon-based ligand discovery in virtual libraries of over 11 billion compounds, *Nature*, vol. 601, no. 7893, pp. 452–459, 2022.

[33] F. Gentile, J. C. Yaacoub, J. Gleave, M. Fernandez, A. T. Ton, F. Ban, A. Stern, and A. Cherkasov, Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking, *Nat. Protoc.*, vol. 17, no. 3, pp. 672–697, 2022.

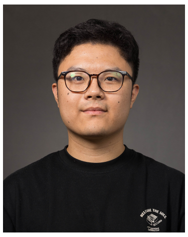
[34] J. Wang, Z. Qiu, X. Zhang, Z. Yang, W. Zhao, and X. Cui, Boosting deep learning-based docking with cross-attention and centrality embedding, in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine* (*BIBM*), Las Vegas, NV, USA, 2022, pp. 360–365.

[35] K. Atz, F. Grisoni, and G. Schneider, Geometric deep learning on molecular representations, *Nat. Mach. Intell.*, vol. 3, no. 12, pp. 1023–1032, 2021.

[36] D. Bafna, F. Ban, P. S. Rennie, K. Singh, and A. Cherkasov, Computer-aided ligand discovery for estrogen receptor alpha, *Int. J. Mol. Sci.*, vol. 21, no. 12, p. 4193, 2020.

[37] M. Kriegel, H. J. Wiederanders, S. Alkhashrom, J. Eichler, and Y. A. Muller, A PROSS-designed extensively mutated estrogen receptor *α* variant displays enhanced thermal stability while retaining native allosteric regulation and structure, *Sci. Rep.*, vol. 11, no. 1, p. 10509, 2021.

[38] D. Probst and J. L. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees, *J. Cheminf.*, vol. 12, no. 1, p. 12, 2020.



Zhaoxu Meng received the bachelor degree from Shandong University, China in 2023. He was at the School of Life Sciences, Shandong University, China, and worked at the School of Computer Science and Technology, Shandong University, China. He is currently pursuing the PhD degree at UT Southwestern Medical Center, Dallas, USA, studying under the program computational biology in biomedical engineering. His research interests include bioinformatics, deep learning, neuroscience, and immunology.

Cheng Chen received the bachelor and master degrees from Qingdao University of Science and Technology, China in 2017 and 2020, respectively. He is currently pursuing the PhD degree at Shandong University, China. His research interests include bioinformatics, data mining, and deep learning.

Xuan Zhang received the bachelor degree from Shandong University, China in 2021. He is currently pursuing the PhD degree at the School of Computer Science and Technology, Shandong University, China. His research interests include bioinformatics, deep learning, and drug discovery.

Xuefeng Cui received the bachelor, master, and PhD degrees from University of Waterloo, Canada in 2004, 2006, and 2014, respectively. In 2016, he joined the Institute for Interdisciplinary Information Sciences (IIIS) at Tsinghua University, China, as a tenure-track assistant professor. In 2019, he was promoted to the position of full professor at Shandong University, China. He is currently a professor at the School of Computer Science and Technology, Shandong University, China. His research focuses primarily on the advancement of deep learning and parallel algorithms to tackle biological problems that have practical applications in daily life.

Wei Zhao received the PhD degree in food nutrition and safety from Huazhong Agricultural University, Wuhan, China in 2021. He is an associate professor at the State Key Laboratory of Microbiology Technology, Shandong University, Qingdao, China. His research focuses on AI-assisted design and druggability evaluation of small molecule antitumor drugs, the development of specific tumor-targeted allosteric inhibitors, and the structural biology of complexes. He also commits to break through the druggability bottleneck problems of the traditional leading compounds with cytotoxicity.