bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MaskMol: Knowledge-guided Molecular Image Pre-Training Framework for Activity Cliffs with Pixel Masking

Zhixiang Cheng, Hongxin Xiang, Pengsen Ma, Li Zeng, Xin Jin, Xixi Yang, Jianxin Lin, Yang Deng, Bosheng Song, Xinxin Feng, Changhui Deng, and Xiangxiang Zeng*, *Senior Member, IEEE,*

***Abstract*—Activity cliffs, which refer to pairs of molecules that are structurally similar but show significant differences in their potency, can lead to model representation collapse and make the model challenging to distinguish them. Our research indicates that as molecular similarity increases, graph-based methods struggle to capture these nuances, whereas image-based approaches effectively retain the distinctions. Thus, we developed MaskMol, a knowledge-guided molecular image self-supervised learning framework. MaskMol accurately learns the representation of molecular images by considering multiple levels of molecular knowledge, such as atoms, bonds, and substructures. By utilizing pixel masking tasks, MaskMol extracts fine-grained information from molecular images, overcoming the limitations of existing deep learning models in identifying subtle structural changes. Experimental results demonstrate MaskMol's high accuracy and transferability in activity cliff estimation and compound potency prediction across 20 different macromolecular targets, outperforming 25 state-of-the-art deep learning and machine learning approaches. Visualization analyses reveal MaskMol's high biological interpretability in identifying activity cliff-relevant molecular substructures. Notably, through MaskMol, we identified candidate EP4 inhibitors that could be used to treat tumors. This study not only raises awareness about activity cliffs but also introduces a novel method for molecular image representation learning and virtual screening, advancing drug discovery and providing new insights into structure-activity relationships (SAR). Code is available at https://github.com/ZhixiangCheng/MaskMol.**

***Index Terms*—drug discovery, deep learning, activity cliff estimation, knowledge-guided pre-training, explainable artificial intelligence**

I. Introduction

Drug discovery has always posed a significant challenge in life sciences, and its outcome could tremendously impact medical research. Recently, the advancements in machine learning and artificial intelligence are now opening up new possibilities and leading to breakthroughs in the field of drug discovery [1]–[3]. Over the past few years, machine learning has made remarkable advancements in various aspects of early drug discovery, such as molecular generation [4]–[6], molecular optimization [7]–[9], and molecular property prediction [10]–[16]. These technologies are offering more efficient and accurate methods for developing new drugs.

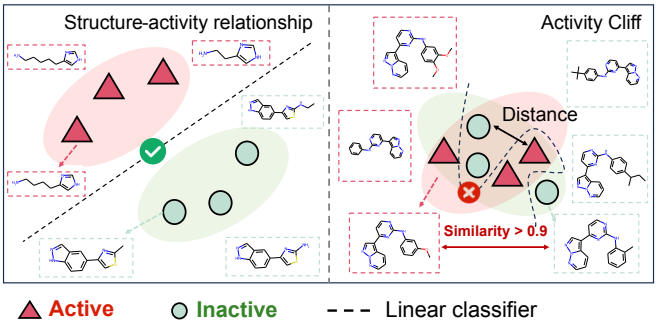


Fig. 1. Examples of SAR (*Left*) and activity cliff (*Right*) in feature space. Highly active molecules are laid in red boxes and low-active molecules in green boxes.

Molecular property prediction plays a vital role in the drug discovery and design process, as it directly impacts the safety, effectiveness, and efficiency of drug development [17]. The fundamental concept behind molecular property prediction is that molecules with similar structures tend to have similar properties [18]. As shown in Fig. 1 (Left), molecules with distinct scaffolds exhibit different activities, and they can be well separated. However, there are cases called activity cliffs [19], where two molecules with similar structures have significantly different biological activities (Fig. 1 Right). Predicting activity cliffs holds substantial importance in rational drug design and the efficient discovery of new therapeutic agents [20]. Anticipating cliffs provides crucial insights into SAR and optimizes lead compounds more effectively, leading to more reliable biological activity prediction.

Activity cliff task is a very important yet understudied task in the field of drug discovery. Previous studies [21]–[23] have observed that graph-based models have poor performance on activity cliffs. We conjecture that the graph-based representation learning methods cannot separate two similar molecules in the feature space, which is called representation collapse,

Zhixiang Cheng, Hongxin Xiang, Pengsen Ma, Xixi Yang, Jianxin Lin, Bosheng Song, and Xiangxiang Zeng are with the College of Information Science and Engineering, Hunan University, China. (E-mail: zhixiangcheng, xiangh, pengsenma, yangxixi, linjianxin, boshengsong, xzeng@hnu.edu.cn)

Li Zeng is with the Department of AIDD, Shanghai Yuyao Biotechnology Co. Ltd, China. (E-mail: lzeng@yuyaobiotech.com)

Xin Jin is with the Ningbo Institute of Digital Twin, Eastern Institute of Technology, China. (E-mail: jinxin@eias.ac.cn)

Yang Deng and Changhui Deng are with the Hunan Provincial Key Laboratory of Anti-Resistance Microbial Drugs, The Third Hospital of Changsha (the Affiliated Changsha Hospital of Hunan University), China. (E-mail: yangdeng, changhui-deng@cssdsyy.com)

Xinxin Feng is with the State Key Laboratory of Chem-/Bio-Sensing and Chemometrics, Hunan Provincial Key Laboratory of Biomacromolecular Chemical Biology, and School of Chemistry and Chemical Engineering, Hunan University, China. (E-mail: xinxin_feng@hnu.edu.cn)

**Corresponding author: Xiangxiang Zeng.*

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

resulting in poor performance on the activity cliffs. As shown in Fig. 2, we evaluate the performance of various GNN architectures on the activity cliffs, such as GCN [24], GAT [25], and MPNN [26]. Fig. 2 clearly shows that as the similarity between pairs of molecules increases, the distance in the feature space of graph-based methods decreases faster, proving our conjecture. We defined this phenomenon as representation collapse. Therefore, we turn to discover other representations of molecules and find that various graph-based representation learning methods are inferior to image-based representation learning methods in identifying differences between similar molecules. Although molecular graphs and images describe the same molecular information, they are essentially different due to modal differences (graph versus image) and feature extraction differences (GNN versus CNN). In the activity cliff task, pairs of molecules have very similar structures and significant differences in activity. For example, a difference of just one atom can lead to a completely different activity. Increasing the discrimination between two similar molecules is the key to the success of predicting activity cliffs for deep learning models. For the GNN model, the small structural difference will be over-smoothed [27] out during information aggregation, resulting in little difference in the extracted features. This is also why GNN methods perform poorly on activity cliff tasks. For images, the convolution operation in CNN has the characteristics of local connectivity and parameter sharing, which makes the model pay more attention to local features to preserve these differences [28]. These observations indicate that image-based methods can amplify the differences between two similar molecules and motivate us to develop an image-based method for more accurate activity cliff prediction.

Besides, obtaining labels for activity cliffs requires expensive and time-consuming wet experiments. The inadequacy of labeled data significantly impacts model performance. Thus, we turn our attention to the pretrain-finetune paradigm [29]–[34], because the pre-training process doesn't need labels, and few labels can be used in the fine-tuning phase to enhance performance. However, unlike natural images, molecular images are not as information-dense and have many blank areas. If we simply apply the pre-trained framework in computer vision such as MAE [30] directly to molecular images, it would be challenging for the model to utilize meaningful molecular knowledge to identify subtle changes in cliff molecules. Therefore, it is necessary to use molecular domain knowledge to guide the model to learn molecule structures.

Moreover, activity cliffs often arise due to subtle changes at various molecular levels [36], [37], such as specific atom substitutions, bond modifications, or functional group replacements. At the atomic level, substituting a hydrogen atom on a benzene ring with a chlorine atom can lead to significant changes in the molecule's binding interactions with receptors, thereby affecting its biological activity. Changing a single bond in a molecule to a double bond may alter the molecule's shape and electronic distribution, thereby affecting its interactions with targets and its biological activity. Replacing a hydroxyl group on a benzene ring with a methyl group. While the structural difference is insignificant, the hydroxyl group can form hydrogen bonds, significantly affecting the molecule's solubility and interactions with biological targets. As a result, our objective is to incorporate prior chemical knowledge into the model and utilize this activity cliff-related knowledge to instruct the model in learning molecules. Here, we present a novel self-supervised pre-training framework called MaskMol, which focuses on learning fine-grained representations from molecular images with knowledge-guided pixel masking. We design three pixel masking-based pre-training tasks with three different levels of knowledge, involving atomic knowledge, bond knowledge, and motif knowledge. These tasks enable MaskMol to comprehensively learn the local regions of molecules by pixel-level knowledge prompts. In summary, our main contributions are:

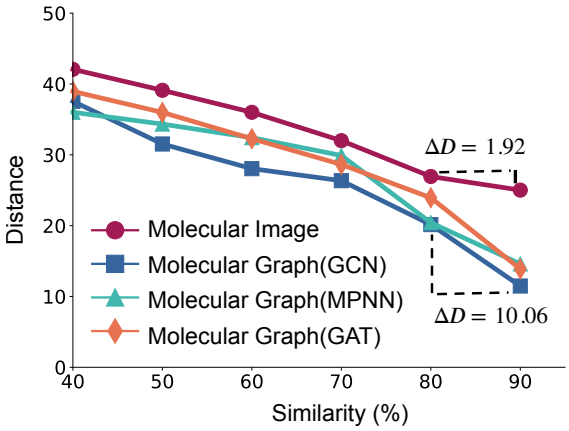


Fig. 2. Comparison of graph and image in feature space. The similarity is calculated using the Tanimoto coefficient on molecular ECFP fingerprint pairs [35]. The distance is defined as the average Euclidean distance in the 2D space of the 1000 molecule pairs after the encoder extracts features. Among them, the molecular image uses ResNet18 [28], and the molecular Graph adopts GCN, MPNN, and GAT as the encoder.

- We first pinpoint the bottleneck in the molecular activity cliff task that the cliff molecules give rise to deep learning model representation collapse. Image-based model is superior to graph-based model due to alleviating representation collapse.
- We design a novel and multi-level knowledge-guided molecular image self-supervised learning framework (called MaskMol) using a pixel masking strategy. After pre-training on a large-scale dataset consisting of approximately two million molecules, MaskMol demonstrated a significant performance enhancement on activity cliff estimation datasets and compound potency prediction datasets.
- Explainable case study and visualization demonstrate that MaskMol strongly enables cliff awareness for bioactivity estimation and extracting meaningful SAR information for intuitive interpretation.
- Through MaskMol, we identified candidate EP4 inhibitors that could be used to treat tumors, demonstrating that MaskMol can be used as a promising method under activity cliff virtual screening scenario.

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

II. RELATED WORK

A. Activity Cliff Estimation

In the field of drug discovery, the prediction of activity cliffs has been a challenging task aimed at identifying pairs of molecules that are structurally similar yet exhibit significantly different biological activities. A variety of traditional machine learning methods have been proposed to address this issue in recent years, including [38]–[40]. MoleculeACE [21] is a benchmarking platform specifically focused on identifying activity cliffs and quantifying the corresponding model performance. So far, many pretraining methods have achieved remarkable success in molecular property prediction tasks, including [17], [41], [42]. However, activity cliff estimation based on pre-training methods has been disproportionately underinvestigated. The pioneering work of Instructbio [43] accurately estimates cliff molecules by combining semi-supervised learning and GNN. Different from Instructbio, we combine molecular images with multi-level knowledge in a pixel masking manner for efficient activity cliff estimation.

B. Masked Image Modeling

Masked image modeling (MIM), inspired by the recent success of masked language modeling in NLP [29], [44], has attracted growing interest in visual pre-training. The pioneering work of iGPT [45] conducts the autoregressive task of predicting the next pixel, which allows iGPT to achieve advanced image comprehension. SimMIM [46] directly predicts the raw RGBs of the masked patches through a linear layer. MAE [30] introduced an asymmetric encoder-decoder structure, with the encoder operating on the part of the observed signal and a lightweight decoder reconstructing the complete signal from the latent representation and the masked token. Different from the above works, we aim to combine multi-level chemical knowledge with MIM to accurately predict activity cliffs.

III. METHOD

A. Overview of MaskMol

This section gives an overview of our MaskMol, highlighted in Fig. 3. To accurately estimate molecular activity cliffs, we developed a knowledge-guided molecular image pre-training framework by fine-grained pixel masking, MaskMol. It consists of two parts: (1) three knowledge-guided pixel masking strategies, and (2) three knowledge-guided masked pixel prediction tasks for pre-training. See the Experimental Section for more descriptions on MaskMol.

B. Knowledge-guided Pixel Masking

Firstly, the conversion from molecular SMILES to molecular images is performed using RDKit. To eliminate any extraneous color effects, we proceed by removing all non-essential hues from the molecular images. Next, again leveraging RDKit, we apply green hues to atoms, bonds, and motifs separately. In the following, HSV detection isolates regions with green pixels within the highlighted image. To introduce an element of randomness, we select a subset of atom/bond masking images by randomly choosing a fraction (determined by the masking ratio γ) from the available masking atom/bond image sets. Consequently, we generate a set of masking images, totaling $\gamma \cdot N_{atom}$ and $\gamma \cdot N_{bond}$ in number. It is important to note that to ensure that motifs do not cross each other, we only randomly select one masking image from the set of masking motif images. Moving forward, the masked image is combined with the original molecular image. Precisely, we adjust the white region of the masking image to correspondingly modify the region of the molecular image. In this synthesis process, we end up with three masked molecular images: the masked atom/bond/motif image. The three images are input through ViT to obtain latent features and classified through different fully connected layers. The pre-trained molecular encoder is fine-tuned on downstream tasks to further improve model performance.

C. Knowledge-guided Masked Pixel Prediction

Definition. A molecule's 2D information can usually be represented as a graph $G = (\mathcal{V}, \mathcal{E})$ with atoms $\mathcal{V}$ as nodes and the edges $\mathcal{E}$ given by covalent bonds. But in our experiments, the molecule is expressed as the image $x \in \mathbb{R}^{H \times W \times C}$, where (H, W) is the resolution of the molecular image, C is the number of channels.

1) Atom-level Masked Pixel Prediction: We counted the atom types of molecules in the pre-training data and selected the ten most frequent atom types ($e.g.$, C, N, O, Cl). Correspondingly, the ten atom types serve as pseudo-labels for the atom-level masked pixel prediction (AMPP). Formally, the molecular image set and the pseudo-labels are $\left\{x_i \in \mathbb{R}^{224 \times 224 \times 3}\right\}_{i=1}^{N}$ and $y^{atom} \in \{0, 1, \cdots, 9\}^{10}$ respectively. For each x_i, we will get the mask atom image sets $M = \{M_j\}_{j=1}^{N_{atom}}$ by Masking. Random sampling M with a masking ratio γ to get the subset of M as $M^* = \{M_j\}_{j=1}^{m}$, where $m = \gamma \cdot N_{atom}$ denotes the masking image number of subset. Then, we can obtain the masking atom image, denoted as $\tilde{x}_i = x_i \Theta M^*$, where Θ indicates modifying the pixel value in x_i corresponding to the white pixel area in M to white. Following ViT [31], we divide a masking atom image $\tilde{x}_i$ into regular non-overlapping patches. To save calculation time and make our model pay more attention to the masked patches, we only calculate the loss of the masked patches $\Omega(\tilde{x}_i)$. Finally, the cost function of the AMPP task is as follows:

$$\mathcal{L}_{\text{AMPP}} = \arg\min_{\theta, W} \frac{1}{N} \sum_{i=1}^{N} \ell\left(\omega\left(f_\theta\left(\Omega\left(\tilde{x}_i\right)\right)\right), y^{atom}\right) \quad (1)$$

where f_θ and θ refer to the mapping function and corresponding parameters of the molecular encoder, ω represents the parameters of the fully connected classification layers, ℓ is the cross-entropy (CE) loss function.

2) Bond-level Masked Pixel Prediction: The workflow of the bond-level masked pixel prediction (BMPP) is similar to that of AMPP, and the difference is that there are only four bond types, $i.e.$, single, double, triple, and aromatic, and

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

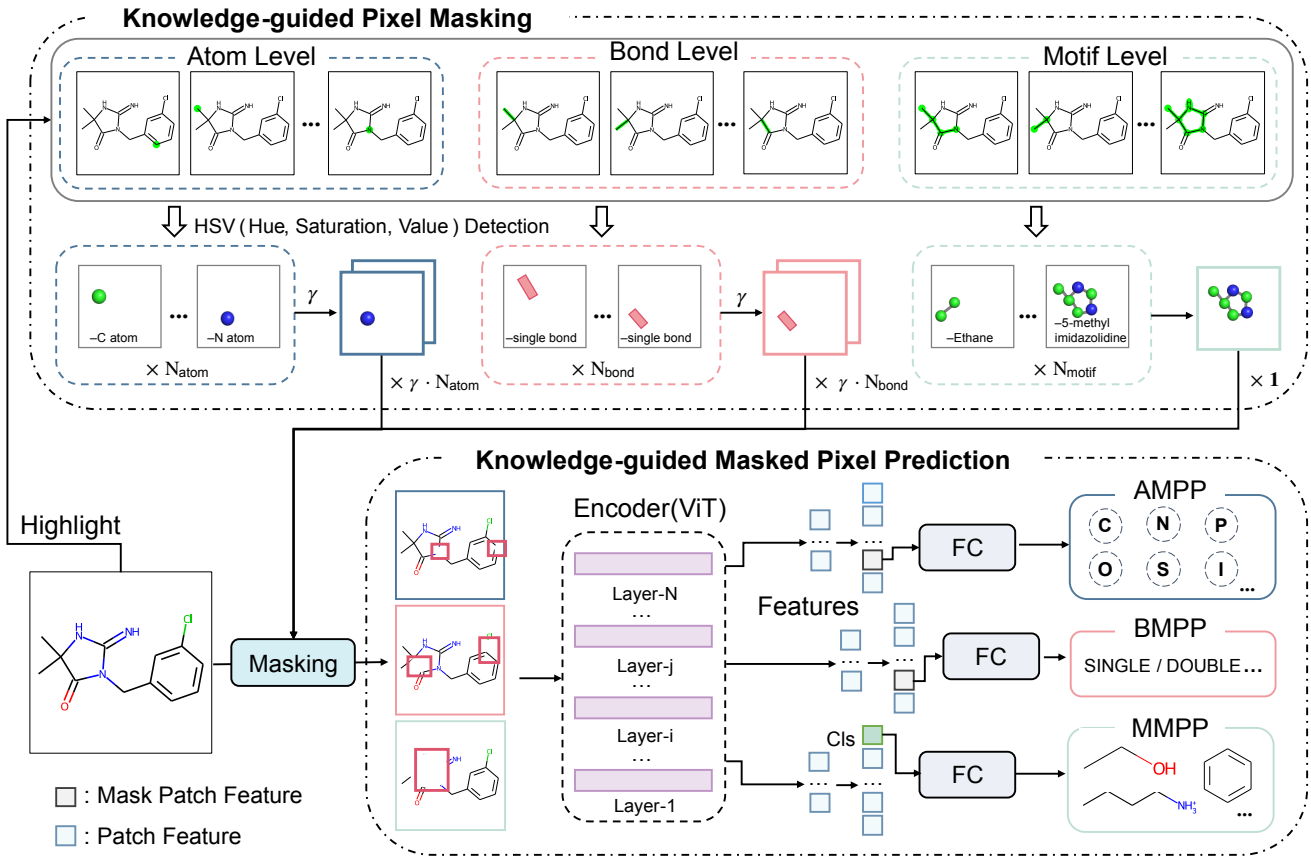


Fig. 3. The framework of our MaskMol pipelines. The framework comprises two major components, knowledge-guided pixel masking and knowledge-guided masked pixel prediction. RDKit generates molecular highlight images and HSV color detection is used to obtain molecular knowledge-guided masked images. In the pre-training stage, self-supervised learning tasks of masking are introduced to capture knowledge and structure information hidden inside the molecular image.

the pseudo-labels are $y^{bond} \in \{0,1,2,3\}^4$. The BMPP loss function is defined as follows:

$$\mathcal{L}_{\mathrm{BMPP}} = \arg\min_{\theta,W} \frac{1}{N} \sum_{i=1}^{N} \ell\left(\omega\left(f_\theta\left(\Omega\left(\tilde{x}_i\right)\right)\right), y^{bond}\right) \quad (2)$$

3) Motif-level Masked Pixel Prediction: Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS [47]) based on chemical reaction templates was utilized to partition functional groups. However, the functional group vocabulary obtained through the BRICS division is somewhat redundant. To address this issue, two rules defined in MGSSL [12] were applied to eliminate redundant functional groups. As a result, we obtained a motif vocabulary consisting of 9854 motifs. We opted for the top 200 motifs with the highest occurrence and eliminated molecules lacking these particular motifs to reduce time and space burdens on the MMPP task. The motif-level masked pixel prediction (MMPP) process is also consistent with AMPP. The difference is that the pseudo-labels are $y^{motif} \in \{0,1,\cdots,199\}^{200}$ and we only randomly sample a motif in M as M^* so that there is no intersection between motifs and the model can extract accurate motif information. It is worth noting that when calculating loss, we use the classification token feature $\tilde{x}_i^{cls}$ to classify and perform loss calculation. The MMPP loss function is defined as follows:

$$\mathcal{L}_{\mathrm{MMPP}} = \arg\min_{\theta,W} \frac{1}{N} \sum_{i=1}^{N} \ell\left(\omega\left(f_\theta\left(\tilde{x}_i^{cls}\right)\right), y^{motif}\right) \quad (3)$$

D. Pre-training and Fine-tuning

Here, we used ViT as our molecular encoder. After using data augmentations and masking to obtain masking molecular images $\tilde{x}_i$, we forward these images $\tilde{x}_i$ to the ViT model to extract latent features $f_\theta\left(\tilde{x}_i\right)$. Then, these latent features are used by three pretext tasks to calculate the total cost function $\mathcal{L}$, which is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{AMPP}} + \mathcal{L}_{\mathrm{BMPP}} + \mathcal{L}_{\mathrm{MMPP}} \quad (4)$$

In order to pretrain our MaskMol, we first gathered 2 million unlabeled molecules with drug-like properties from the PubChem database [48]. We divided the 2M pre-training data into a training set (95%) and a validation set (5%), and judged the pre-training performance through the accuracy of each task. Finally, the AMPP, BMPP, and MMPP accuracy can reach 99.3%, 98.0%, and 89.6%, respectively. After the initial pre-training phase, we proceed to fine-tune the pre-trained encoder for the specific downstream tasks. In particular, we incorporate an extra fully connected layer after the encoder. The output

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dimension of this layer is set to match the number of categories associated with the downstream tasks.

IV. EXPERIMENTS

To evaluate the effectiveness of the image-based representations learned by MaskMol, we choose wide-ranging popular or state-of-the-art baselines for comparison on activity cliff estimation benchmark (ACE), including 12 pre-training baselines and 11 traditional machine learning methods. We refer to the original paper [21] and follow its strategy to split the dataset. To assess the generalization of MaskMol, we employed a widely used splitting strategy known as scaffold split on ACE task and compound potency prediction (CPP) task. This is a more challenging but practical setting since the test molecules can be structurally different from the training set.

A. Experimental Setups

1) Datasets: (1) ***ACE:*** We adopt the MoleculeACE [21] framework, which focuses on identifying activity cliffs and quantifying model efficacy. A difference in bioactivity of at least one order of magnitude (i.e., 10× fold disparity) in reported K_i values is adopted as the threshold for defining activity cliff pairs. Similar to prior work, molecules were clustered based on substructure similarity using extended connectivity fingerprints (ECFPs). Each cluster was then split into a training set (80%) and a testing set (20%) using stratified random sampling based on the activity cliff label. Besides, we apply the scaffold splitting to construct the training (80%), validation (10%), and test sets (10%) to evaluate the generalization of MaskMol. (2) ***CPP:*** Following Janela et al. [49], activity classes are extracted from the ChEMBL database (version 30) [50]. The selection criteria involved choosing bioactive compounds that have the highest level of confidence in direct interactions with a human target protein (confidence score 9) and a specified potency value (IC_{50}). The potency values were transformed into a negative logarithmic scale (pIC_{50}). Here, we apply the scaffold splitting to construct the training (80%), validation (10%), and test (10%) sets.

2) Evaluation Metrics: The overall performance of MaskMol was quantified via the mean absolute error (MAE) or root-mean-square error (RMSE) computed on the bioactivity values (*i.e.*, pK_i or pIC_{50}): $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$, where $\hat{y}_i$ is the predicted bioactivity of the i-th molecule, y_i is the corresponding experimental value, and n represents the total number of molecules. On activity cliffs, the performance of MaskMol was quantified by computing the root-mean-square error ($\mathrm{RMSE}_{\mathrm{cliff}}$) on compounds that belonged to at least one activity cliff pair: $\mathrm{RMSE}_{\mathrm{cliff}} = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(\hat{y}_i - y_i)^2}$, where $\hat{y}_i$ is the predicted bioactivity of the i-th compound, y_i is the corresponding experimental value, and n_c represents the total number of compounds on activity cliffs.

B. Baselines

The performance regarding methods (MLP [51], GBM [52], RF [53], SVM [54], KNN [55], AFP [56], MPNN [26], GAT [25], GCN [24], CNN [57], LSTM [58]) is derived from MoleculeACE [21]. The performance regarding methods (MolCLR [59], GROVER [60], GEM [61], InstructBio [43]) is derived from InstructBio. We additionally execute experiments on the activity cliff estimation datasets following the same experimental setting used in Mole-BERT [10], EdgePred [42], GraphMVP [62], 3DInfomax [63], and ImageMol [17].

C. Implementation Details of MaskMol

MaskMol is pre-trained by SGD optimizer with a learning rate of 0.01, weight decay 10-5, momentum 0.9, and batch size 128 for approximately 2 days with 4 NVIDIA A100 GPU (40GB). In downstream tasks, the pre-trained model is fine-tuned using SGD optimizer with batch size {8, 16, 32, 64}, learning rate {5e-5, 5e-4, 5e-3, 0.01}, weight decay 10-5, momentum 0.9 on Ubuntu 18.04.1 with 15 vCPU Intel(R) Xeon(R) Platinum 8375C CPU @ 2.90GHz and NVIDIA 4090 (20GB).

D. Activity Cliff Estimation

As shown in Fig. 4, we compared the performance of MaskMol with three types of state-of-the-art self-supervised molecular representation models using MoleculeACE experimental set-up. Compared with the second-best model (InstructBio), the elevated RMSE of MaskMol ranges from 2.3% to 22.4% with an overall relative improvement of 11.4% across 10 ACE datasets, in particular for HRH3 dataset (19.4% RMSE improvement) and ABL11 dataset (22.4% RMSE improvement). In addition, MaskMol achieved lower RMSE values (Fig. 5) compared with traditional ECFP-based methods across multiple machine learning algorithms, including support vector machine [54], random forest [53], k-nearest neighbors [55], multilayer perception [51], and gradient boosting machine [52]. In summary, Our method MaskMol, surpasses other state-of-the-art methods, achieving the lowest RMSE in these comparisons. To further substantiate MaskMol's efficacy in identifying activity cliff pairs, we showcase results using $\mathrm{RMSE}_{\mathrm{cliff}}$ as an additional performance metric. On the DAT and OX2R datasets, MaskMol achieves a 6.7% improvement in $\mathrm{RMSE}_{\mathrm{cliff}}$ compared to the second-best method ($\mathrm{SVM}_{\mathrm{ECFP}}$). Taking into account the two metrics of RMSE and $\mathrm{RMSE}_{\mathrm{cliff}}$, MaskMol also has a lower value than any other state-of-the-art molecular representation models (Fig. 6). Furthermore, to evaluate the disparity between the prediction and label, we employ Kullback-Leibler Divergence (KLD [64]) for measuring distribution differences (Fig. 7). The KLD values of all ACE datasets are significantly lower and the distributions of label and prediction values are close, except CLK4. We hypothesize that the relatively pronounced discrepancies observed in the CLK4 dataset could be attributed to its limited molecule number (731), which may have resulted in an under-fitted model.

To test the generalization of MaskMol, we split the datasets using a scaffold split (Fig. 8). We found that MaskMol significantly outperforms $\mathrm{SVM}_{\mathrm{ECFP}}$ models across all 10 ACE datasets. For instance, the RMSE values of MaskMol (RMSE = 0.69) compared with $\mathrm{SVM}_{\mathrm{ECFP}}$ model (RMSE

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

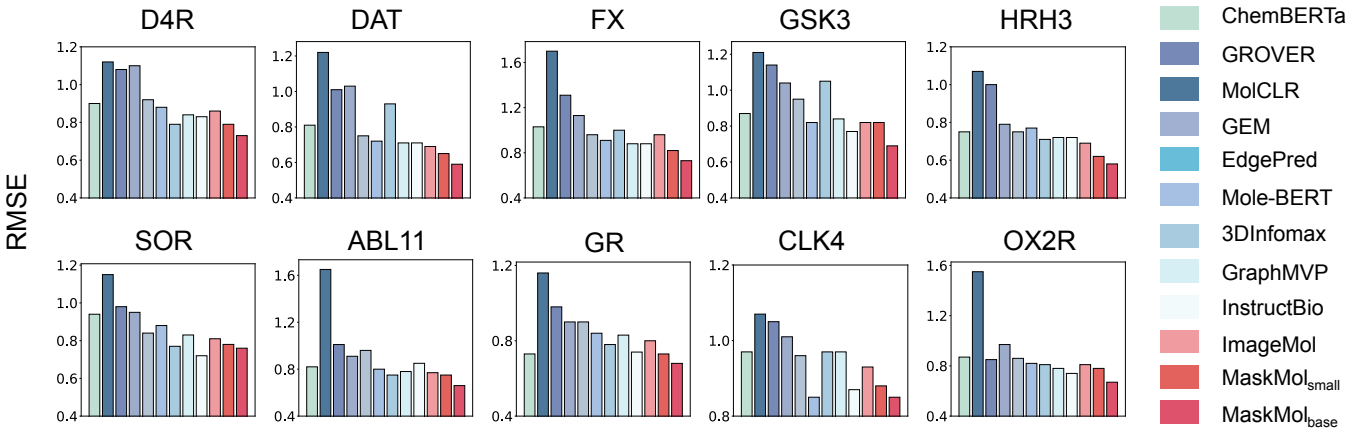


Fig. 4. Comparison of the error on all compounds (denoted as RMSE) on 10 ACE datasets from MoleculeACE with different representation pre-training methods.

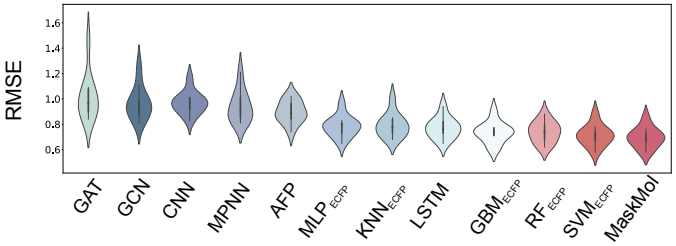


Fig. 5. The violin plot displays RMSE comparison with traditional machine learning methods.

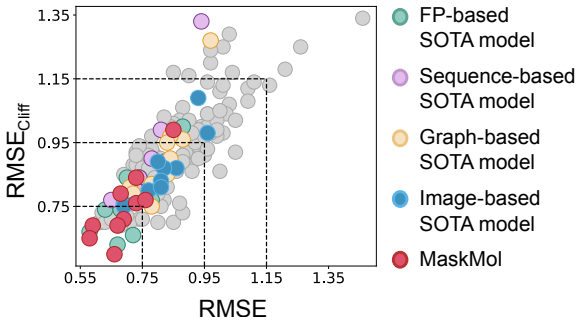


Fig. 6. Comparison between the error on activity cliffs compounds (denoted as $\mathrm{RMSE_{cliff}}$) and RMSE for all methods. Green, purple, yellow, and blue represent the SOTA methods based on fingerprint ($\mathrm{SVM_{ECFP}}$), sequence (LSTM), graph (GraphMVP), and image (ImageMol), respectively.

= 0.97) in the prediction of ABL11 are elevated by over 28.9%. We further evaluated the $\mathrm{RMSE_{cliff}}$, compared with $\mathrm{SVM_{ECFP}}$ models, MaskMol achieves better performance with a performance advantage of 6.4% on average, in particular for SOR (20.9% $\mathrm{RMSE_{cliff}}$ improvement). Compared with the molecule image pre-training model (ImageMol), the elevated RMSE of MaskMol ranges from 6% to 28.8% with a performance advantage of 17% on average, the elevated $\mathrm{RMSE_{cliff}}$ ranges from 9.4% to 40% with a performance advantage of 19.4% on average.

These results validate MaskMol's ability to precisely predict molecules exhibiting activity cliffs. Notably, ECFP-based methods demonstrate robust performance, whereas graph-based methods tend to underperform in activity cliff estimation. Graph-based models are vulnerable to representation collapse when faced with activity cliffs, and they face challenges in learning from non-smooth objective functions [23]. Furthermore, we found that image-based methods such as ImageMol have lower RMSE and $\mathrm{RMSE_{cliff}}$ than graph-based algorithms (EdgePred, GraphMVP, 3DInfomax, Mole-BERT). This further demonstrates that the CNN-based model can use local inductive biases to identify subtle cliff changes. Although InstructBio attempts to mitigate representation collapse by leveraging a substantial amount of unlabeled data as pseudo-labels, it still does not match the performance of ECFP-based methods. The addition of pseudo-labels helps to clarify class boundaries [65], suggesting that semi-supervised learning could emerge as a novel solution for addressing activity cliffs.

E. Compound Potency Prediction

Although MaskMol is primarily designed for solving fine-grained tasks such as ACE, it also performs well on the coarse-grained task of CPP. Compound potency prediction is crucial to the drug discovery and design process [66], [67]. Researchers aim to forecast the biological activity of chemical compounds, explicitly measuring their potency in terms of the amount needed to produce a desired effect. As shown in Fig. 9, MaskMol has a better performance compared with sequence-based (ChemBERTa), graph-based (MolCLR, MGSSL, MPG [68], and GraphMVP), and image-based models (ImageMol) using a scaffold split. Notably, on the BACE1 dataset, MaskMol achieves a small MAE of 0.56, while the best-performing baseline model (ImageMol), achieves 0.63. It is worth mentioning that MaskMol achieves this performance using only 2M pre-training data, compared to the 10M pre-training data used by ChemBERTa and ImageMol. This demonstrates that MaskMol can achieve superior performance with significantly less pre-training data.

F. Ablation Studies on MaskMol

We perform comprehensive experiments to investigate the impact of each component in MaskMol on the activity cliff estimation. As illustrated in Fig. 10, seven out of ten datasets have a pre-training gain of more than 30% and the gain reaches its peak at 45.87% on the DAT dataset. Furthermore, the average gain across all ACE datasets surpasses 34.43%, underscoring the substantial enhancement in MaskMol's performance at-

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

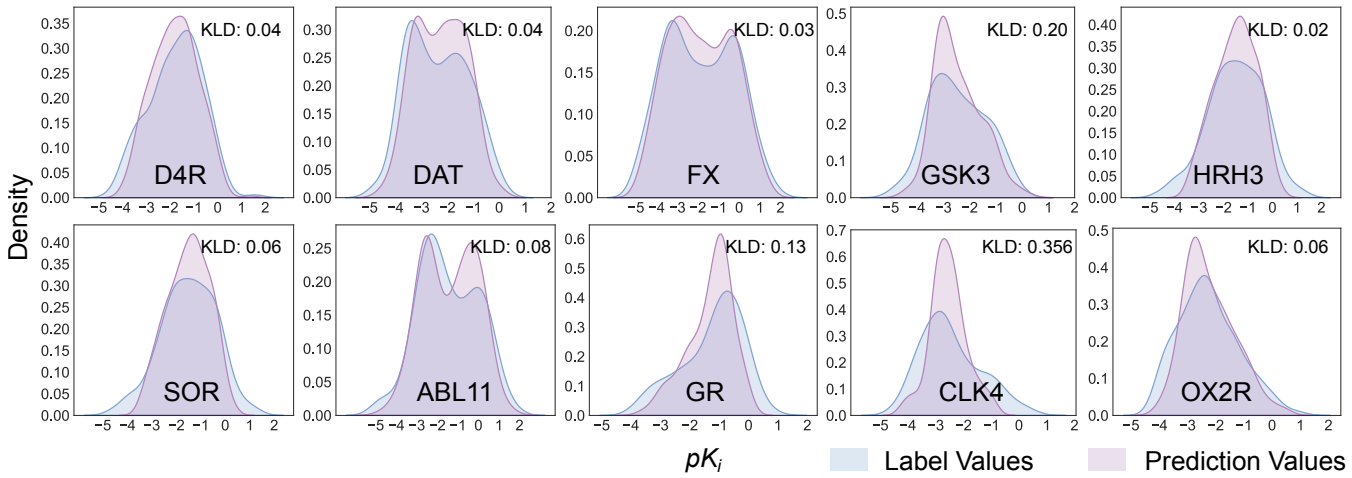


Fig. 7. Evaluation of the gap between label values and prediction values on ACE, in which the data distribution is determined by kernel density estimation. The kullback-leibler divergence (KLD) is to measure the difference between two probability distributions.

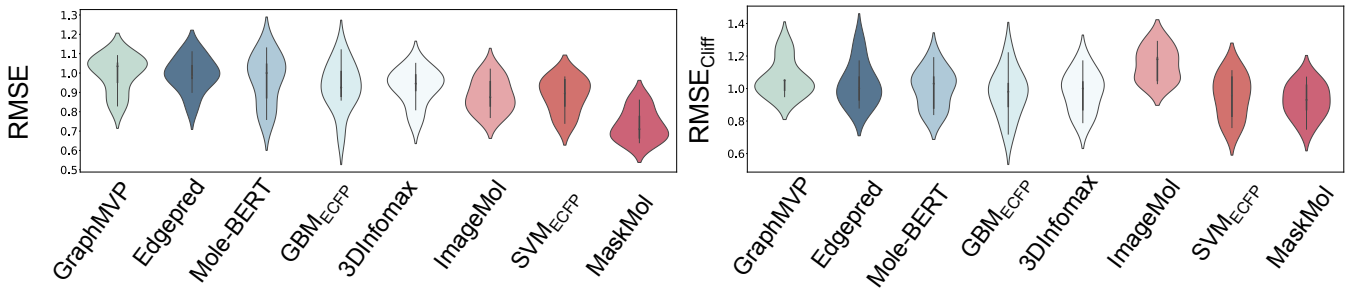


Fig. 8. The performance of MaskMol and baseline methods on an ACE task under scaffold splitting was measured in terms of RMSE.

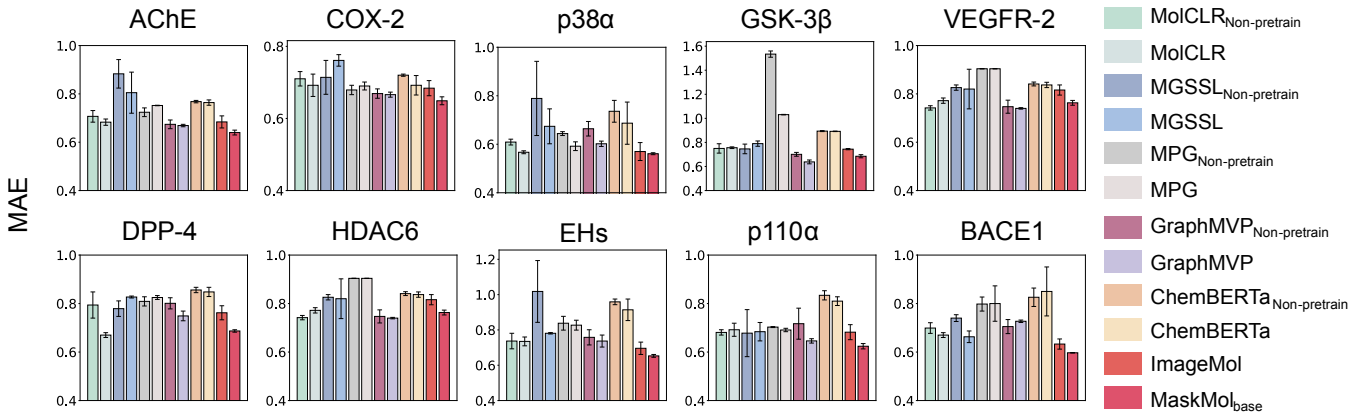


Fig. 9. The performance of MaskMol and baseline methods on the CPP task under scaffold splitting was measured in terms of MAE. All the prediction results were reported based on three independent runs on three random seeded(seed=0,1,2). Data are presented as mean ± standard deviation (SD).

tributable to knowledge-guided masked pixel prediction tasks. Unlike graphs, graph treats molecules as nodes and bonds, which encode a large amount of chemical information such as atom types and bond types. For an initialized image model, molecules are input into the model in the form of RGB pixels, which do not contain any chemical information. The model's understanding of molecular images is limited to the fact that the image is composed of some "line." Therefore, it is necessary to help the model understand the chemical information in the image, which allows the model to understand the specific meaning of the "lines." in the image. This is why we can see that MaskMol has greater improvement gains than graph-based GROVER before and after pre-training (34.43% versus 8.53%). The observed decline in performance for "w/o AMPP" (RMSE 4.5% decline), "w/o BMPP" (RMSE 16.4% decline), and "w/o MMPP" (RMSE 21% decline) indicates that the removal of any level knowledge-guide task adversely affects MaskMol's performance, with MMPP being the most influential. We also explored the impact of pre-training with different data scales. The size of pre-training dataset for $\mathrm{MaskMol}_{\mathrm{base}}$ and $\mathrm{MaskMol}_{\mathrm{small}}$ are 20K and 2M respectively. We found that the average RMSE performance increased from 0.76 to 0.70 as the pre-trained data scale increased. This suggests that MaskMol will be further improved as more molecules

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

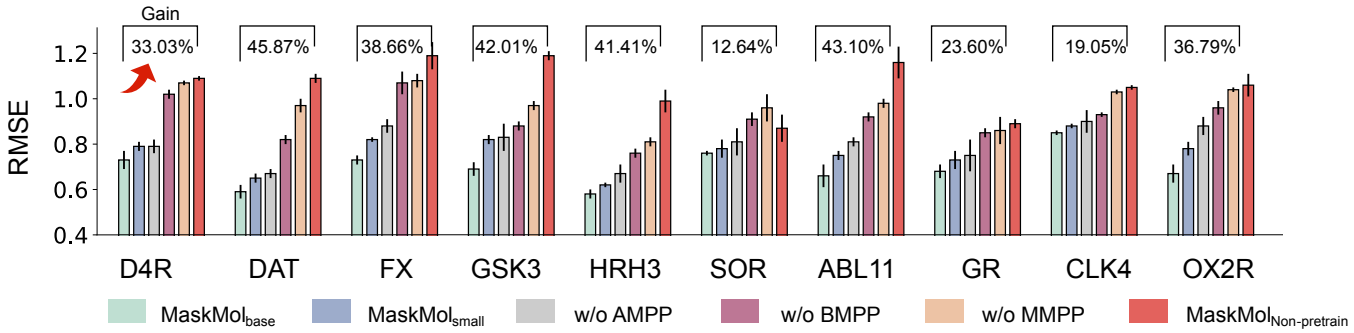


Fig. 10. The ablation study of the pretext task in the MoleculeACE setting. "w/o AMPP" denotes removing the AMPP pretext task component during pre-training, "w/o BMPP" and "w/o MMPP" in the same way. "Gain" indicates the improvement of the $\mathrm{MaskMol_{base}}$ compared to the $\mathrm{MaskMol_{Non\text{-}Pretrain}}$.

are added to the pre-training dataset.

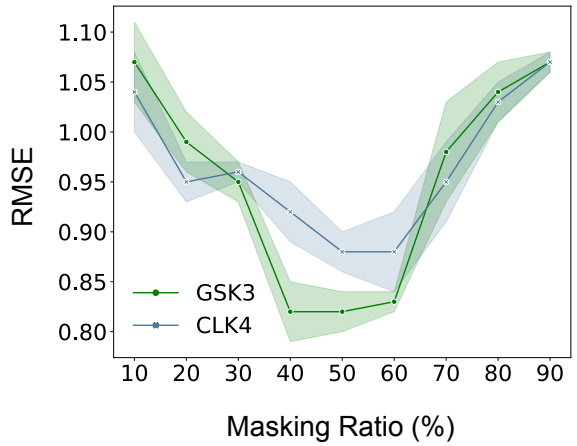


Fig. 11. Ablation study on the masking ratio used in the pre-training stage. The x-axis represents the masking ratio in the AMPP and BMPP tasks. The error bands stand for the standard deviations.

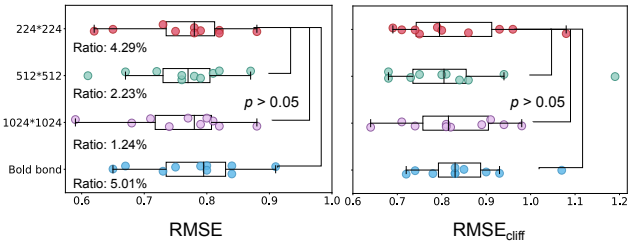


Fig. 12. Experiments on different image sizes and useful pixel ratios in ACE datasets. The ratio is the molecule pixels divided by the whole image pixels on 12,590 images. RDKit is used to bold the chemical bonds to generate an image with a higher ratio of useful pixels. Statistical analysis is performed with the Mann-Whitney U test.

Additionally, we delve into analyzing the implications of the masking ratio, examining how its value affects MaskMol's overall performance (Fig. 11). It is worth noting that the optimal masking ratio in our study significantly deviates from the typical ratios used in BERT and MAE. BERT typically employs a masking ratio of 15%, whereas MAE utilizes a masking ratio as high as 75%. However, we found that a 50% masking ratio yields optimal results in our experiments. Molecular images are rather sparse with most pixels being

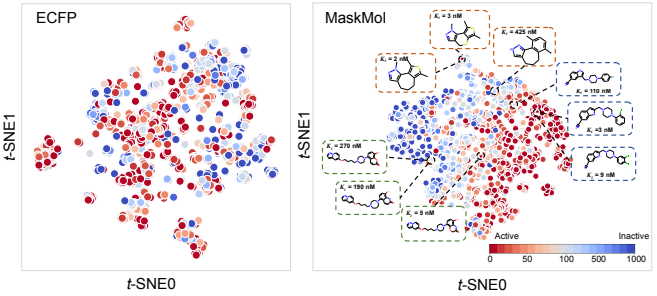


Fig. 13. Visualization of molecular ECFP fingerprint *(Left)* and molecular representations derived from MaskMol *(Right)* via t-SNE on D4R dataset from ACE task, respectively. Each point is colored according to its corresponding K_i value. The larger the K_i value, the cooler the color, and the smaller the corresponding K_i value, the warmer the color. Each pair of structurally similar but bioactivity-distinct molecules are laid in boxes with the same color.

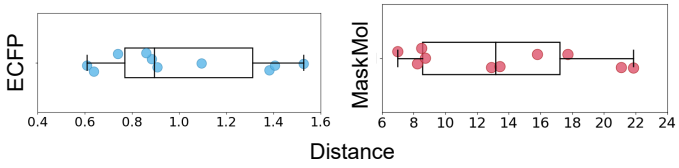


Fig. 14. Comparison of ECFP fingerprint and MaskMol in quantifying the overall relative distance of active cliff pairs in the latent features space.

empty and the resolution of the images is important in such settings. Thus, we research the impact of image size and the ratio of empty spaces to useful pixels on the learned representations (Fig. 12). The results show that the image size and useful pixel ratio achieved similar performance on ACE dataset (p>0.05, Mann-Whitney U test [69]).

G. Interpretation of MaskMol

1) Investigation of MaskMol Representation: We use t-SNE to compare the representation learned by MaskMol with the ECFP fingerprints feature (Fig. 13). The t-SNE algorithm maps similar molecular representations to adjacent points in two dimensions. We observe that ECFP can only be mapped based on structure, resulting in active and inactive molecules being mixed in the feature space. Through multi-level knowledge-guided masked pixel prediction tasks, MaskMol can be aware of changes in atom/bond/motif when any atom/bond/motif in the image changes. Thus, the representations learned by

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MaskMol can effectively distinguish between active and inactive molecules, with a clear boundary between them. Additionally, we have included some randomly selected pairs of activity cliffs in the figure to illustrate the similar and dissimilar molecules learned by MaskMol based on their biological activity. MaskMol can learn similar representations from molecules with similar structures and properties and map molecules with significant differences in structures and properties to distinct feature spaces. This demonstrates that MaskMol learns the topological structure information between molecules and uses properties to differentiate between molecules.

To measure the distance between active cliff pairs in feature space, we introduce a distance metric $d = \frac{1}{N}\sum_{i=1}^{N}\rho_i$, $\rho_i = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, where $M_1 = (x_1, y_1), M_2 = (x_2, y_2)$ are active cliff pair coordinates in the unified feature space. Fig. 14 illustrates that in all ACE datasets, the distance between active cliff pairs in the feature space generated by MaskMol is considerable, significantly greater than that of ECFP. This observation highlights the effectiveness of MaskMol in accurately estimating activity cliffs, as it can capture subtle structural variations and utilize them to describe and represent molecules.

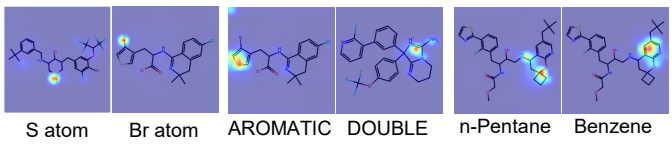


Fig. 15. Examples of MaskMol's heatmaps of three knowledge levels highlighted by Grad-CAM. A hotter color area indicates higher MaskMol attention. Given a masked image with different knowledge levels, MaskMol gives the masked knowledge type and focuses on the correct mask area.

2) Explaining MaskMol via Attention Visualization: We applied three levels of knowledge-guided pixel masking to the molecular images and used Grad-CAM [70] to visualize the areas of attention (Fig. 15). The results show that MaskMol accurately classifies the knowledge and focuses on the appropriate masked areas. This indicates that our three knowledge-guided masked pixel prediction tasks allow the model to identify different molecular chemical structures.

In Fig. 16, we provide a comparative analysis of key substructures associated with activity cliffs, as extracted by various deep learning (DL) methods. We select the top-3 most crucial edges detected by PGExplainer [71]. GNNs tend to allocate attention to insignificant regions of the cliff molecule and emphasize the identical structure. This observation supports our hypothesis that GNNs are susceptible to representation collapse when dealing with active cliffs, thereby hindering their ability to correctly identify cliff molecules. We can see that ImageMol focuses on large areas of the molecules, while MaskMol, without pre-training, only focuses on the entire molecule and ignores irrelevant blank areas. However, neither of them pays attention to the important substructure that affects the activity. MaskMol successfully identifies the most informative substructure and judges compound activity based on these substructures. These plots convincingly prove that MaskMol recognizes subtle differences in activity cliff pairs' substructures and can provide reliable and informative insights for medicinal chemists in identifying key substructures.

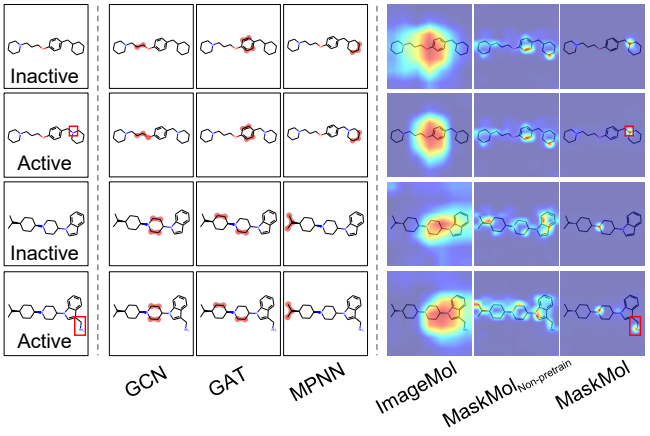


Fig. 16. Visualization of the explanatory structure found by different DL methods on activity cliffs. Image-based methods use Grad-CAM to visualize attention, while graph-based approaches employ PGExplainer to mine essential structures.

3) Chemistry-intuitive exposition of MaskMol: We use Substructure-Mask Interpretation (SME [72]) to further quantify the contribution of substructure to MaskMol predictions. We define the impact of the masking substructure on the overall prediction as the attribution. We make two predictions with MaskMol, one before and one after applying the substructure masking to the molecular image, and consider the difference between the predicted values as the attribution: $\text{Attribution}_{\text{sub}} = f(x) - f(x_{sub})$, among them, x represents the molecular image, x_{sub} represents the molecular image of masking substructure, and f represents MaskMol. By calculating the contribution of substructure to model predictions, we can gain insight into the impact of substructure on molecule activity. As depicted in Fig. 17, adding substructures such as benzene ring (Attribution = -1.93, K_i = 5,370 nM) and ethyl alcohol (Attribution = -0.95, K_i = 758 nM), the attributions are lower than zero, and the influence of the benzene ring is greater than that of ethyl alcohol, which is highly consistent with the molecular activity value. It can also be found that the position of the propyl group affects the activity, and the attribution value also makes the same judgment. Fig. 17 also shows the same conclusion in the DAT dataset. In addition to biological activity, we also present the chemically intuitive explanation of MaskMol on Mutagenicity. Fig. 18 and Fig. 19 display the analysis of different substructures based on their Mutagenicity. A positive attribution indicates that the substructure contributes to toxicity, while a negative attribution suggests that the substructure has a detoxifying effect. Fig. 19 reveals that nitro, amino, and quinone groups enhance the model's ability to predict toxicity, while carboxyl groups improve the model's prediction of non-toxicity. This observation aligns with previous studies, which have identified aromatic nitro, aromatic amino, and quinone groups as toxic and carboxyl groups as detoxifying [73]–[75].

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

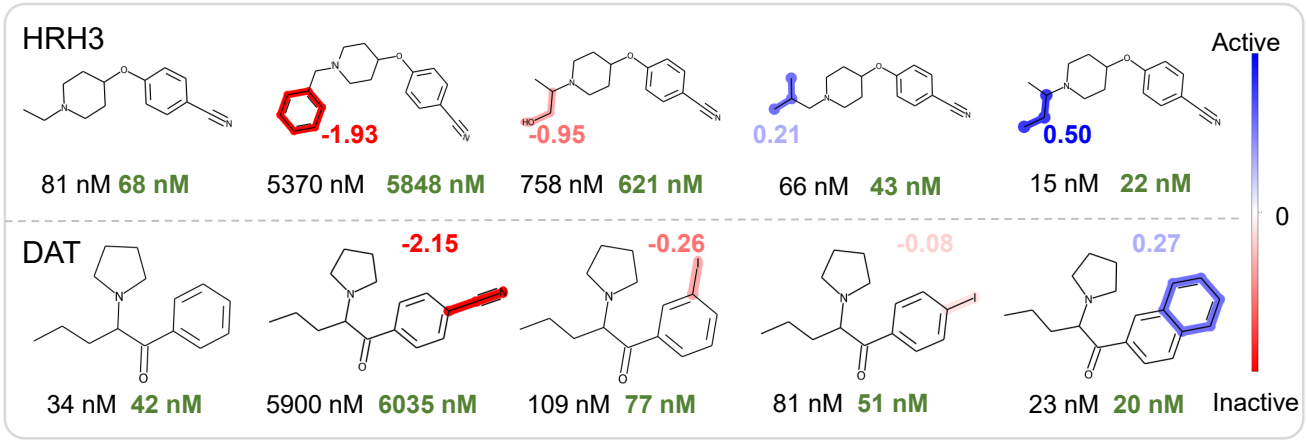


Fig. 17. The attribution visualization of four compounds of HRH3 and DAT receptors, respectively. Black represents the true value and green represents the prediction value.

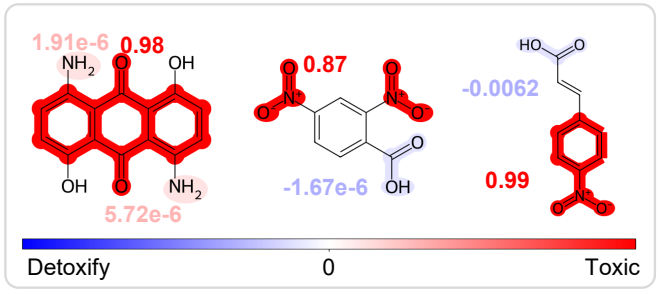


Fig. 18. The attribution visualization of three compounds of mutagenicity. The mutagenicity model achieved high ROC-AUC scores of 0.90.

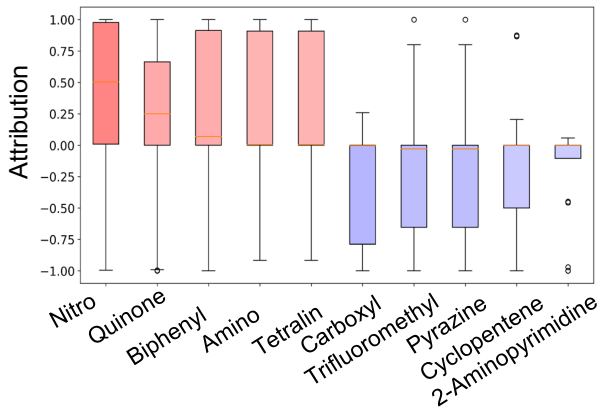


Fig. 19. The attribution of the different functional groups in the whole dataset. Only the functional groups that appear more than twenty times in the dataset are included. Blue hues indicate negative attributions below zero, signifying the substructure's inclination towards the property, while red hues represent positive attributions above zero, implying the substructure's unfavorable impact on the property.

In summary, this visualization provides evidence that MaskMol is subtle structure-aware and exploits structural differences to make accurate predictions. Thus, MaskMol can provide meaningful and fresh SAR insights to help medicinal chemists in structural optimization and de novo design.

H. Virtual screening using MaskMol

EP4 receptor has been widely investigated and recognized as a promising drug target for cancer immunotherapy [76]. We manually collected data from multiple sources, including the BindingDB [77], ChEMBL database, and patent libraries targeting EP4. Canonicalization of the molecules was achieved utilizing RDKit, and duplication of SMILES was deleted, resulting in a finalized dataset comprising 1633 molecules. We evaluated the performance of MaskMol on EP4 targets with a random split of 8:1:1. We found that MaskMol has a low RMSE on the test set (RMSE = 0.577), and the prediction values are linearly correlated with the label values in Fig. 5f left (R^2 = 0.789). The t-SNE visualization in the latent space showed a clear boundary between inhibitors and non-inhibitors (Fig. 20a grey dots). To test the generalization ability of MaskMol, we constructed an additional patent set (131 molecules) from the extended patents and literature as an external validation set (R^2 = 0.755). We found that inhibitors and non-inhibitors in the patent test were also perfectly separated. MaskMol identified 9 known EP4 inhibitors and visualized these 9 molecules to embedding space (Fig. 20b), suggesting structural identification ability of MaskMol to learn discriminative information. These nine molecules (Grapiprant [78], L001 [79], CJ-042794 [80], MK-2894 [81], CR6086 [82], ONO-4578 [83], E7046 [84], HL-43 [85], and AMX12006 [86]) have been validated (including cell assay, clinical trial, or other evidence) as potential EP4 inhibitors. These findings demonstrate the ability of MaskMol to provide robust and generalizable molecular representation and prediction of inhibitors of targets, making it an efficient and effective virtual screening method.

V. CONCLUSION

In the field of early-stage drug discovery, machine learning is gaining prominence, yet the concept of activity cliffs remains underexplored. Activity cliffs, which refer to structurally similar molecules with significant differences in potency, are critical for virtual screening and developing models that understand complex structure-activity relationships. Traditional graph-based methods often struggle with representation

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

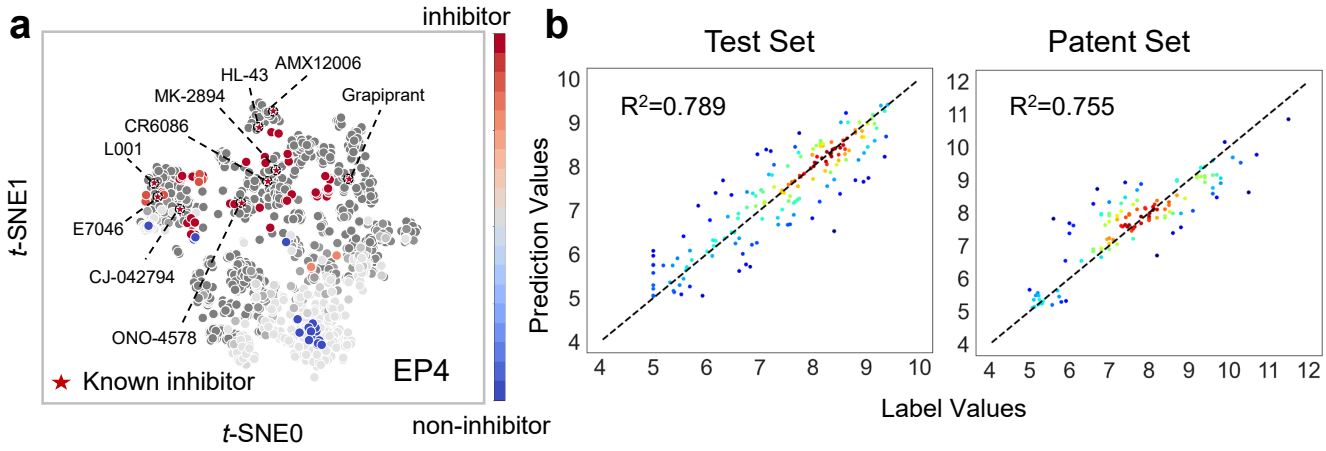


Fig. 20. **(a)** Visualization of molecular representations of molecules from the EP4 derived from MaskMol. Grey dots represent molecules used to train MaskMol, and dark grey dots represent inhibitors. Red and blue dots represent inhibitors and non-inhibitors in the patent set, respectively. The dashed circles indicate that the molecules had been previously identified as inhibitors of EP4 in previous studies. **(b)** The gap between prediction values and label values of test set (*Left*) and patent set (*Right*). The black dotted line represents y=x, and the closer to this line, the warmer the point color.

collapse due to high similarity between activity cliffs. To address this, we developed MaskMol, a knowledge-guided self-supervised learning framework utilizing molecular images. MaskMol employs three pre-training tasks with pixel masking, incorporating atomic, bond, and motif knowledge. This approach enables MaskMol to effectively learn local molecular regions and detect subtle changes in activity cliffs. Experimental results confirm MaskMol's superior accuracy in predicting activity cliffs and its performance compared to other state-of-the-art algorithms. Extensive experiments and ablation studies validate the effectiveness of each MaskMol component and determine the optimal ratio for knowledge-guided pixel masking. Furthermore, MaskMol identifies critical substructures responsible for activity cliffs through visualization, enhancing researchers' understanding of compounds and facilitating the drug discovery process. This study not only raises awareness about activity cliffs but also introduces a novel method for molecular image representation learning and virtual screening, advancing drug discovery and providing new insights into structure-activity relationships.

Future potential directions may improve MaskMol further: (1) Incorporating more chemical knowledge (such as fingerprints knowledge, 3D space structure knowledge, and chemical reaction knowledge) into image model is a promising future direction. Fingerprint-based methods have demonstrated excellent performance in predicting activity cliffs. Therefore, incorporating information from multiple fingerprints, such as MACCS [87], ECFP, PharmPrint [88], and USRCAT [89], into the image model can enhance its accuracy. Currently, the consideration of activity cliffs does not include chiral cliffs. Learning 3D spatial structure information may be beneficial for predicting chiral cliffs. Despite the structural similarity of cliff molecules, their reaction synthesis processes differ, providing a unique perspective that can be used to identify these molecules. (2) Fine-grained alignment of images and other representations (for example SMILES and graph). By applying our proposed knowledge-guided pixel masks, we can perform fine-grained masking on images. This approach allows us to align images and graphs (or SMILES) more precisely, guiding the model to learn from multiple views and improving its ability to capture subtle differences. (3) Multi-task learning of multiple activity cliff prediction datasets. This approach will not only predict activity cliffs but also address related tasks such as toxicity prediction and pharmacokinetic parameter prediction. By tackling multiple tasks simultaneously, we expect to improve the model's overall generalization and performance.

ACKNOWLEDGMENT

The work was supported by National Natural Science Foundation of China (Grant Nos. 62450002, 62202153, 62272151, 62372159, 62302156, 61972138, 62106073, 62122025, 62102140 and U22A2037), Hunan Provincial Natural Science Foundation of China (Grant Nos. 2024JJ4015, 2023JJ40180, 2022JJ20016 and 2021JJ10020), The Science and Technology Innovation Program of Hunan Province (Grant Nos. 2022RC1100, 2022RC1099), Postgraduate Scientific Research Innovation Project of Hunan Province (Grant No. CX20220380), the project of Hunan Provincial Key Laboratory of Anti-Resistance Micro-bial Drugs (No:2023TP1013).

REFERENCES

[1] N. Fleming, "How artificial intelligence is changing drug discovery," *Nature*, vol. 557, no. 7706, pp. S55–S55, 2018.
[2] X. Zeng, F. Wang, Y. Luo, S.-g. Kang, J. Tang, F. C. Lightstone, E. F. Fang, W. Cornell, R. Nussinov, and F. Cheng, "Deep generative molecular design reshapes drug discovery," *Cell Reports Medicine*, 2022.
[3] J.-P. Vert, "How will generative ai disrupt data science in drug discovery?" *Nature Biotechnology*, pp. 1–2, 2023.
[4] Y. Diao, D. Liu, H. Ge, R. Zhang, K. Jiang, R. Bao, X. Zhu, H. Bi, W. Liao, Z. Chen *et al.*, "Macrocyclization of linear molecules by deep learning to facilitate macrocyclic drug candidates discovery," *Nature Communications*, vol. 14, no. 1, p. 4552, 2023.
[5] D. Flam-Shepherd, K. Zhu, and A. Aspuru-Guzik, "Language models can learn complex molecular distributions," *Nature Communications*, vol. 13, no. 1, p. 3293, 2022.

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[6] O. Mahmood, E. Mansimov, R. Bonneau, and K. Cho, "Masked graph modeling for molecule generation," *Nature communications*, vol. 12, no. 1, p. 3156, 2021.

[7] X. Yang, L. Fu, Y. Deng, Y. Liu, D. Cao, and X. Zeng, "Gpmo: Gradient perturbation-based contrastive learning for molecule optimization," 2023.

[8] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," in *International conference on machine learning*. PMLR, 2018, pp. 2323–2332.

[9] ——, "Hierarchical generation of molecular graphs using structural motifs," in *International conference on machine learning*. PMLR, 2020, pp. 4839–4848.

[10] J. Xia, C. Zhao, B. Hu, Z. Gao, C. Tan, Y. Liu, S. Li, and S. Z. Li, "Molebert: Rethinking pre-training graph neural networks for molecules," in *The Eleventh International Conference on Learning Representations*, 2022.

[11] D. Xue, H. Zhang, D. Xiao, Y. Gong, G. Chuai, Y. Sun, H. Tian, H. Wu, Y. Li, and Q. Liu, "X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis," *bioRxiv*, pp. 2020–12, 2020.

[12] Z. Zhang, Q. Liu, H. Wang, C. Lu, and C.-K. Lee, "Motif-based graph self-supervised learning for molecular property prediction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 870–15 882, 2021.

[13] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in neural information processing systems*, vol. 33, pp. 5812–5823, 2020.

[14] S. Luo, T. Chen, Y. Xu, S. Zheng, T.-Y. Liu, L. Wang, and D. He, "One transformer can understand both 2d & 3d molecular data," *arXiv preprint arXiv:2210.01765*, 2022.

[15] Z. Guo, P. Sharma, A. Martinez, L. Du, and R. Abraham, "Multilingual molecular representation learning via contrastive pre-training," *arXiv preprint arXiv:2109.08830*, 2021.

[16] H. Li, R. Zhang, Y. Min, D. Ma, D. Zhao, and J. Zeng, "A knowledge-guided pre-training framework for improving molecular representation learning," *Nature Communications*, vol. 14, no. 1, p. 7568, 2023.

[17] X. Zeng, H. Xiang, L. Yu, J. Wang, K. Li, R. Nussinov, and F. Cheng, "Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework," *Nature Machine Intelligence*, pp. 1–13, 2022.

[18] J. B. Hendrickson, "Concepts and applications of molecular similarity," *Science*, vol. 252, no. 5009, pp. 1189–1190, 1991.

[19] D. Stumpfe and J. Bajorath, "Exploring activity cliffs in medicinal chemistry: miniperspective," *Journal of medicinal chemistry*, vol. 55, no. 7, pp. 2932–2942, 2012.

[20] A. J. Wedlake, M. Folia, S. Piechota, T. E. Allen, J. M. Goodman, S. Gutsell, and P. J. Russell, "Structural alerts and random forest models in a consensus approach for receptor binding molecular initiating events," *Chemical Research in Toxicology*, vol. 33, no. 2, pp. 388–401, 2019.

[21] D. van Tilborg, A. Alenicheva, and F. Grisoni, "Exposing the limitations of molecular machine learning with activity cliffs," *Journal of Chemical Information and Modeling*, vol. 62, no. 23, pp. 5938–5951, 2022.

[22] J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras, and F. Wang, "Taking a respite from representation learning for molecular property prediction," *arXiv preprint arXiv:2209.13492*, 2022.

[23] J. Xia, L. Zhang, X. Zhu, and S. Z. Li, "Why deep models often cannot beat non-deep counterparts on molecular property prediction?" *arXiv preprint arXiv:2306.17702*, 2023.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[26] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

[27] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[34] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI blog*, 2018.

[35] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[36] Y. Hu and J. Bajorath, "Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the chembl database," *Journal of chemical information and modeling*, vol. 52, no. 7, pp. 1806–1811, 2012.

[37] D. Stumpfe, H. Hu, and J. Bajorath, "Advances in exploring activity cliffs," *Journal of Computer-Aided Molecular Design*, vol. 34, pp. 929–942, 2020.

[38] J. Iqbal, M. Vogt, and J. Bajorath, "Prediction of activity cliffs on the basis of images using convolutional neural networks," *Journal of Computer-Aided Molecular Design*, pp. 1–8, 2021.

[39] J. Park, G. Sung, S. Lee, S. Kang, and C. Park, "Acgcn: graph convolutional networks for activity cliff prediction between matched molecular pairs," *Journal of Chemical Information and Modeling*, vol. 62, no. 10, pp. 2341–2351, 2022.

[40] M. Dablander, T. Hanser, R. Lambiotte, and G. M. Morris, "Exploring qsar models for activity-cliff prediction," *Journal of Cheminformatics*, vol. 15, no. 1, p. 47, 2023.

[41] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[42] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265*, 2019.

[43] F. Wu, H. Qin, W. Gao, S. Li, C. W. Coley, S. Z. Li, X. Zhan, and J. Xu, "Instructbio: A large-scale semi-supervised learning paradigm for biochemical problems," *arXiv preprint arXiv:2304.03906*, 2023.

[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[45] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[46] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

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

[48] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, "Pubchem 2019 update: improved access to chemical data," *Nucleic acids research*, vol. 47, no. D1, pp. D1102–D1109, 2019.

[49] T. Janela and J. Bajorath, "Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1246–1255, 2022.

[50] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey *et al.*, "The chembl bioactivity database: an update," *Nucleic acids research*, vol. 42, no. D1, pp. D1083–D1090, 2014.

[51] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[52] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[53] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

bioRxiv preprint doi: https://doi.org/10.1101/2024.09.04.611324; this version posted September 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[54] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[55] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[56] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] Y. Wang, J. Wang, Z. Cao, and A. Barati Farimani, "Molecular contrastive learning of representations via graph neural networks," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 279–287, 2022.

[60] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, "Self-supervised graph transformer on large-scale molecular data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 559–12 571, 2020.

[61] X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu, and H. Wang, "Geometry-enhanced molecular representation learning for property prediction," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 127–134, 2022.

[62] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang, "Pre-training molecular graph representation with 3d geometry," *arXiv preprint arXiv:2110.07728*, 2021.

[63] H. Stärk, D. Beaini, G. Corso, P. Tossou, C. Dallago, S. Günnemann, and P. Liò, "3d infomax improves gnns for molecular property prediction," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 479–20 502.

[64] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[65] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[66] W. Torng and R. B. Altman, "Graph convolutional neural networks for predicting drug-target interactions," *Journal of chemical information and modeling*, vol. 59, no. 10, pp. 4131–4149, 2019.

[67] M. Sakai, K. Nagayasu, N. Shibui, C. Andoh, K. Takayama, H. Shirakawa, and S. Kaneko, "Prediction of pharmacological activities from chemical structures with graph convolutional neural networks," *Scientific reports*, vol. 11, no. 1, p. 525, 2021.

[68] P. Li, J. Wang, Y. Qiao, H. Chen, Y. Yu, X. Yao, P. Gao, G. Xie, and S. Song, "An effective self-supervised framework for learning expressive molecular global representations to drug discovery," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab109, 2021.

[69] M. P. Fay and M. A. Proschan, "Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules," *Statistics surveys*, vol. 4, p. 1, 2010.

[70] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[71] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," *Advances in neural information processing systems*, vol. 33, pp. 19 620–19 631, 2020.

[72] Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh *et al.*, "Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking," *Nature Communications*, vol. 14, no. 1, p. 2585, 2023.

[73] Z. Wu, D. Jiang, J. Wang, C.-Y. Hsieh, D. Cao, and T. Hou, "Mining toxicity information from large amounts of toxicity data," *Journal of Medicinal Chemistry*, vol. 64, no. 10, pp. 6924–6936, 2021.

[74] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee, and Y. Tang, "In silico prediction of chemical ames mutagenicity," *Journal of chemical information and modeling*, vol. 52, no. 11, pp. 2840–2847, 2012.

[75] P. G. Polishchuk, V. E. Kuz'min, A. G. Artemenko, and E. N. Muratov, "Universal approach for structural interpretation of qsar/qspr models," *Molecular Informatics*, vol. 32, no. 9-10, pp. 843–853, 2013.

[76] S. Peng, P. Hu, Y.-T. Xiao, W. Lu, D. Guo, S. Hu, J. Xie, M. Wang, W. Yu, J. Yang *et al.*, "Single-cell analysis reveals ep4 as a target for restoring t-cell infiltration and sensitizing prostate cancer to immunotherapy," *Clinical Cancer Research*, vol. 28, no. 3, pp. 552–567, 2022.

[77] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D198–D201, 2007.

[78] K. Nakao, A. Murase, H. Ohshiro, T. Okumura, K. Taniguchi, Y. Murata, M. Masuda, T. Kato, Y. Okumura, and J. Takada, "Cj-023,423, a novel, potent and selective prostaglandin ep4 receptor antagonist with antihyperalgesic properties," *Journal of Pharmacology and Experimental Therapeutics*, vol. 322, no. 2, pp. 686–694, 2007.

[79] J. He, X. Lin, F. Meng, Y. Zhao, W. Wang, Y. Zhang, X. Chai, Y. Zhang, W. Yu, J. Yang *et al.*, "A novel small molecular prostaglandin receptor ep4 antagonist, l001, suppresses pancreatic cancer metastasis," *Molecules*, vol. 27, no. 4, p. 1209, 2022.

[80] A. Murase, T. Okumura, A. Sakakibara, H. Tonai-Kachi, K. Nakao, and J. Takada, "Effect of prostanoid ep4 receptor antagonist, cj-042,794, in rat models of pain and inflammation," *European journal of pharmacology*, vol. 580, no. 1-2, pp. 116–121, 2008.

[81] M. Blouin, Y. Han, J. Burch, J. Farand, C. Mellon, M. Gaudreault, M. Wrona, J.-F. Lévesque, D. Denis, M.-C. Mathieu *et al.*, "The discovery of 4-{1-[({2, 5-dimethyl-4-[4-(trifluoromethyl) benzyl]-3-thienyl} carbonyl) amino] cyclopropyl} benzoic acid (mk-2894), a potent and selective prostaglandin e2 subtype 4 receptor antagonist," *Journal of medicinal chemistry*, vol. 53, no. 5, pp. 2227–2238, 2010.

[82] G. Caselli, A. Bonazzi, M. Lanza, F. Ferrari, D. Maggioni, C. Ferioli, R. Giambelli, E. Comi, S. Zerbi, M. Perrella *et al.*, "Pharmacological characterisation of cr6086, a potent prostaglandin e 2 receptor 4 antagonist, as a new potential disease-modifying anti-rheumatic drug," *Arthritis Research & Therapy*, vol. 20, pp. 1–19, 2018.

[83] T. Kotani, H. Takano, T. Yoshida, R. Hamasaki, G. Kohanbash, K. Takeda, and H. Okada, "Inhibition of pge2/ep4 pathway by ono-4578/bms-986310, a novel ep4 antagonist, promotes t cell activation and myeloid cell differentiation to dendritic cells," *Cancer Research*, vol. 80, no. 16_Supplement, pp. 4443–4443, 2020.

[84] D. I. Albu, Z. Wang, K.-C. Huang, J. Wu, N. Twine, S. Leacu, C. Ingersoll, L. Parent, W. Lee, D. Liu *et al.*, "Ep4 antagonism by e7046 diminishes myeloid immunosuppression and synergizes with treg-reducing il-2-diphtheria toxin fusion protein in restoring anti-tumor immunity," *Oncoimmunology*, vol. 6, no. 8, p. e1338239, 2017.

[85] Y. Jin, Q. Liu, P. Chen, S. Zhao, W. Jiang, F. Wang, P. Li, Y. Zhang, W. Lu, T. P. Zhong *et al.*, "A novel prostaglandin e receptor 4 (ep4) small molecule antagonist induces articular cartilage regeneration," *Cell Discovery*, vol. 8, no. 1, p. 24, 2022.

[86] D. Das, D. Qiao, Z. Liu, L. Xie, Y. Li, J. Wang, J. Jia, Y. Cao, and J. Hong, "Discovery of novel, selective prostaglandin ep4 receptor antagonists with efficacy in cancer models," *ACS Medicinal Chemistry Letters*, vol. 14, no. 6, pp. 727–736, 2023.

[87] N. Salim, J. Holliday, and P. Willett, "Combination of fingerprint-based similarity coefficients using data fusion," *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 435–442, 2003.

[88] M. J. McGregor and S. M. Muskal, "Pharmacophore fingerprinting. 1. application to qsar and focused library design," *Journal of chemical information and computer sciences*, vol. 39, no. 3, pp. 569–574, 1999.

[89] A. M. Schreyer and T. Blundell, "Usrcat: real-time ultrafast shape recognition with pharmacophoric constraints," *Journal of cheminformatics*, vol. 4, pp. 1–12, 2012.