









Dual-View Learning Based on Images and Sequences for Molecular Property Prediction

Xiang Zhang , Hongxin Xiang , Xixi Yang , Jingxin Dong , Xiangzheng Fu ,
Xiangxiang Zeng , *Senior Member, IEEE*, Haowen Chen , and Keqin Li , *Fellow, IEEE*

Abstract—The prediction of molecular properties remains a challenging task in the field of drug design and development. Recently, there has been a growing interest in the analysis of biological images. Molecular images, as a novel representation, have proven to be competitive, yet they lack explicit information and detailed semantic richness. Conversely, semantic information in SMILES sequences is explicit but lacks spatial structural details. Therefore, in this study, we focus on and explore the relationship between these two types of representations, proposing a novel multimodal architecture named ISMol. ISMol relies on a cross-attention mechanism to extract information representations of molecules from both images and SMILES strings, thereby predicting molecular properties. Evaluation results on 14 small molecule ADMET datasets indicate that ISMol outperforms machine learning (ML) and deep learning (DL) models based on single-modal representations. In addition, we analyze our method through a large number of experiments to test the superiority, interpretability and generalizability of the method. In summary, ISMol offers a powerful deep learning toolbox for drug discovery in a variety of molecular properties.

Index Terms—Drug design and development, images and SMILES strings, predict molecular properties, deep learning toolbox.

I. INTRODUCTION

DRUG discovery is time-consuming, expensive, and high-risk endeavor, with the average time exceeding 10 years

Manuscript received 28 June 2023; revised 4 December 2023; accepted 21 December 2023. Date of publication 28 December 2023; date of current version 7 March 2024. This work was supported in part by the Natural Science Foundation of Hunan Province under Grant 2023JJ30161, in part by the Natural Science Foundation of Changsha City under Grant kq2202137, in part by the Key Project of Scientific Research of Hunan Provincial Education Department under Grant 22A0022, in part by the Changsha City Takes the Lead in Major Science and Technology Projects under Grant KQ2102002, and in part by the Postgraduate Scientific Research Innovation Project of Hunan Province under Grant CX20220380. (*Corresponding authors: Hongxin Xiang; Haowen Chen.*)

Xiang Zhang, Hongxin Xiang, Xixi Yang, Jingxin Dong, Xiangzheng Fu, Xiangxiang Zeng, and Haowen Chen are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: zhxiang@hnu.edu.cn; xianghx@hnu.edu.cn; yangxixi@hnu.edu.cn; jxdong@hnu.edu.cn; excelsior511@126.com; xzeng@foxmail.com; hwchen@hnu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York, New York, NY 12561 USA (e-mail: lik@newpaltz.edu).

The code and supplemental materials are available at: <https://github.com/Mrzhang1999/ISMol>.

Digital Object Identifier 10.1109/JBHI.2023.3347794

and the average cost exceeding \$1-2 billion to bring a new drug approved for clinical use [1], [2]. To reduce the reliance on labor-intensive experiments and improve the efficiency of drug development [3], [4], significant efforts have been made in developing efficient computational tools and bioinformatics methods [5], [6], [7]. Molecular property prediction (MPP) is a fundamental task during drug discovery, which includes bioactivity prediction [8], [9], toxicity prediction [10], [11], drug-likeness prediction [12], [13], and so on. Quantitative structure-activity (property) relationship (QSAR/QSPR) models have increasingly become dominant methods in the selection of promising drug candidates. ML-based QSAR/QSPR models are data-driven and heavily dependent on appropriate molecular representations [14], [15]. Currently, the representations of molecules include molecular descriptors, graph, Simplified Molecular Input Line Entry System (SMILES), and image.

Molecular descriptors are mathematical representations that are algorithmically generated [16], which quantitatively describe the topological and physicochemical structure of molecules, such as fingerprint-based descriptors [17], [18]. Although descriptor-based methods have shown promising results, they frequently necessitate extensive feature engineering in the initial phase [19], [20], [21]. This poses a significant challenge for many researchers who may not possess the expertise or resources to generate high-quality features. Unlike molecular descriptors, structural information about atoms and bonds is displayed clearly in molecular graphs, enabling graph-based methods to be naturally suited for extracting molecular features [22], [23], [24], [25], [26]. However, the application of graph neural networks in MPP is currently constrained since they are prone to overfitting and over-smoothing problems [27], [28], [29]. Given that SMILES strings are linguistically defined graphical structures for representing chemical information, natural language processing (NLP) methods have been extensively adopted [30]. SMILES-based methods have achieved remarkable performance in molecular property prediction [31], [32], [33], [34]. Nonetheless, accurate predictions of molecular properties are still a challenge due to the limited spatial information of the molecules this representation contains [35]. Molecular images represent the detailed structural features of molecules through pixels, which is one of the most intuitive representations for humans. In recent times, there has been a growing emphasis on the field of bioimage analysis [36]. Specially, the feasibility of images as a novel representation was proven by an impressive method, ImageMol [37]. However, the molecular images contain weak

chemical semantics, making it difficult for models to directly extract chemical-related information from them, which requires more knowledge of chemistry to further improve performance.

While methods based on a single representation have achieved significant performance, they all rely on unimodal information. In contrast, multimodal models integrate two or more representations, which provide multiple views of molecules to enable more robust completion of MPP tasks. Many researchers have attempted various combinations of SMILES with different representations [38], [39], [40], [41], but due to sparse chemical semantics of molecular images, they have not yet succeeded in integrating images. However, molecular images contain abundant molecular structural information, as they display the topological features of molecules at high resolution, revealing the relative positions of atoms, bond lengths, angles, and other geometric parameters. These are precisely the details that the SMILES representation lacks. This leads us to a hypothesis of whether sequence information in SMILES strings and structural information in molecular images are compensated for each other. If this hypothesis holds, the structural details in images can be beneficial to improve the prediction performance of SMILES-based models and the chemical knowledge encoded in SMILES can also help the learning of molecular images.

In this study, we built a pre-trained DL model based on a dual-stream architecture called ISMol for MPP. We evaluated the performance of ISMol on 14 benchmark datasets and observed its state-of-the-art performance when compared to baselines. Ablation experiments on different modality inputs were conducted, and the results demonstrated competitive performance even when there was no corresponding modal input during the fine-tuning phase. In addition, we conducted ablation experiments on the components of ISMol, and the results showed that simply merging pixels features with SMILES could lead to the model easily learning inductive biases, resulting in a decrease in its performance. However, extensive experiments indicated a noteworthy enhancement in performance through the alignment and integration of both modalities via multiple pretraining tasks. Furthermore, we analysed the biological interpretation of ISMol and found that the chemical space generated by ISMol was more distinguishable than molecular descriptors (MACCS fingerprints). Subsequently, we conducted evaluations of the scaffold generalizability and temporal generalizability of ISMol and discovered that ISMol either equals or surpasses other existing methods. Finally, we performed visual analyses to intuitively explain how ISMol worked. We discovered that ISMol successfully extracted molecular structures from images and integrated them with the information contained in SMILES strings, enabling the model to read molecules from different perspectives.

Our contribution can be summarized into the following:

- We propose a pre-training model called ISMol to predict molecular properties, which is based on SMILES strings and molecular images.
- Extensive analyses indicate that ISMol exhibits strong performance superiority. Additionally, it possesses robust chemical interpretability, feature extraction capabilities, and generalization abilities.

- To the best of our knowledge, this is the first exploration of the relationship between SMILES strings and molecular images. It is the alignment and integration of image features with SMILES strings that truly enhances the model performance, whereas the simple mere act of concatenating them yields the opposite effect.

II. MATERIALS AND METHODS

A. Datasets

The pre-training initial database is a molecular dataset collected from two large-scale drug databases (ChEMBL [42] and ZINC [43]), which are publicly available [44]. Moreover, we subjected the initial database to screening, applying criteria for logP values within the range of $(-5, 7)$, molecular weights between $(12, 600)$, and heavy atom counts between $(3, 50)$. Furthermore, based on the molecular scaffold frequency statistics shown in Supplementary Fig. S1, we found that a large number of scaffolds appeared less than 10 times. Therefore, we required that the frequency of molecular scaffold occurrences be within the range of $[10, 600]$. If there are more than 600, we only take the first 600 molecules. In the end, we obtained 3.5 million molecules. RDKit was used to generate standardized RGB images. We formed a pair of the SMILES string and the corresponding image, called image-SMILES pair, and split all image-SMILES pairs into an 80/20 training/test set based on scaffold split. 14 small pharmaceutical datasets from ADMET-lab 2.0 [45] were used as our benchmark datasets to evaluate the performance of ISMol, which can be categorized as absorption, metabolism, excretion, and toxicity. Each of the datasets was split into the training set, validation set, and test set, with a ratio of 8:1:1. The detailed descriptions of datasets are provided in Supplementary Table SI.

B. ISMol Network Architecture

As demonstrated in Fig. 1(a), ISMol mainly included a visual encoder for extracting unimodal image features, a SMILES encoder to acquire unimodal SMILES features and a multimodal fusion module to align and fuse both. In addition, the architecture also was equipped with three task-specific heads (including Image-SMILES Matching head, Mask SMILES Modeling head and Fingerprint Class Predicting head) to accomplish different pre-training objectives.

Visual Encoder: Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, we resized it to a fixed size $x \in \mathbb{R}^{H_{rs} \times W_{rs} \times 3}$. After the data augmentations [37], the image was segmented into patches $\{x_1^v, x_2^v, \dots, x_{N_v}^v\}$, where (H, W) was the original image size, (H_{rs}, W_{rs}) was the resized image size and $x_i^v \in \mathbb{R}^{P^2 \times 3}$ was the i -th patch with the resolution $P \times P$. Then patches were flattened and linearly projected into 1-D sequences of patch embeddings by a linear transformation $E^v \in \mathbb{R}^{P^2 \times D}$. A special learnable token $x_{cls}^v \in \mathbb{R}^D$ embedding was inserted in front of the first patch to aggregate global visual information. Then, a

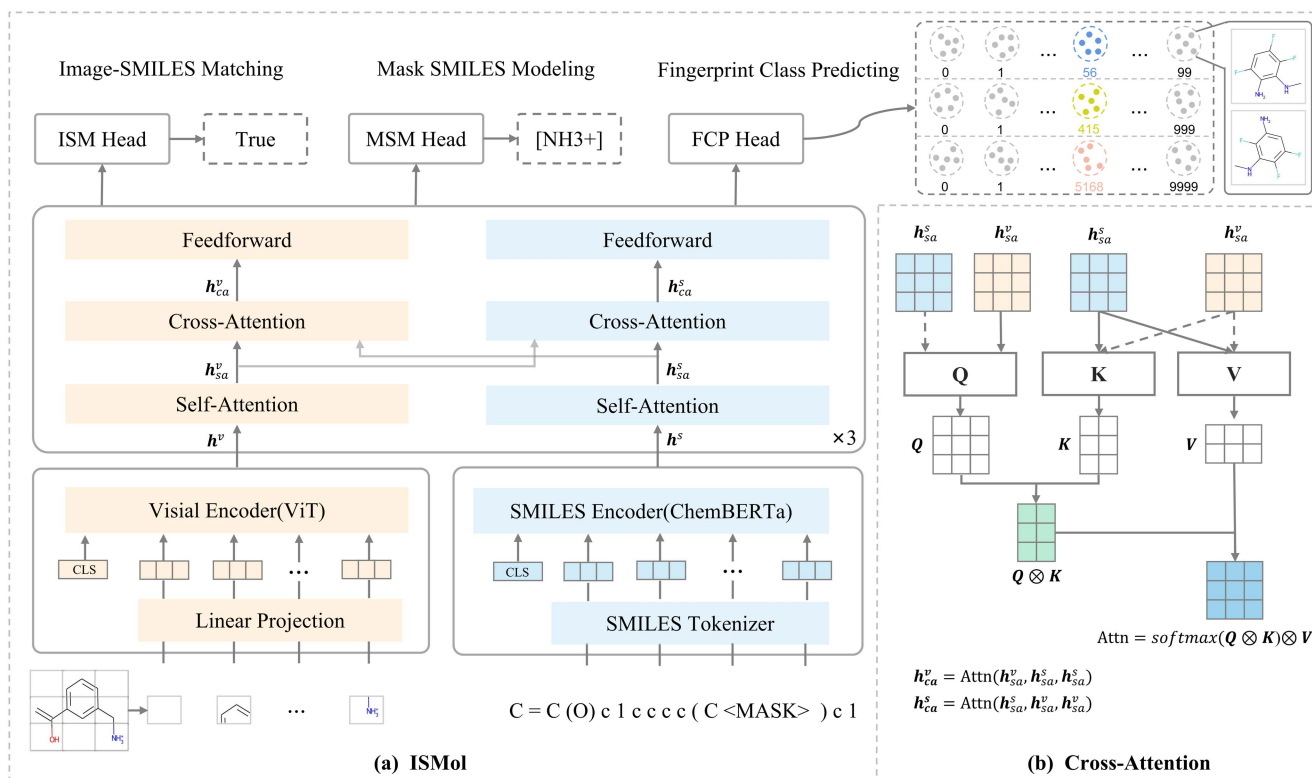


Fig. 1. Overview of the proposed architecture. (a) Dual-stream architecture of ISMol, (b) brief description of the cross-attention mechanism.

learnable embedding $E_{pos}^v \in \mathbb{R}^{(N+1) \times D}$ as positional embedding was added to all embeddings globally:

$$z^v = (x_{cls}, x_1^v E^v, x_2^v E^v, \dots, x_{N_v}^v E^v) + E_{pos}^v. \quad (1)$$

Finally, z^v was fed into transformer blocks and the final output was $h_v = [h_{cls}^v, h_1^v, \dots, h_i^v, \dots, h_{N_v}^v]$, where $h_i^v \in \mathbb{R}^D$ was the i -th hidden state extracted from visual patches. In our actual network, (H_{rs}, W_{rs}) was (224, 224), P was equal to 16, and the Vision Transformer (ViT) [46] was employed as the visual encoder.

SMILES Encoder: Given an input SMILES string, we first tokenized it to an index array in the corresponding vocabulary dict using a special regex pattern [47], then mapped the array to a list of learnable atomic embeddings. After that, a couple of transformer blocks were applied to encode the list as hidden state vectors $h_s = [h_{cls}^s, h_1^s, h_2^s, \dots, h_{N_s}^s]$, where h_{cls}^s was additional embedding to aggregate global SMILES information. ChemBERTa-77M-MLM was used as the initialization of our SMILES encoder, which we could call directly from DeepChem [48]. This encoder was a transformer that adopted masked language modeling (MLM) pretraining on unlabeled 77 million SMILES strings. And, it already had some knowledge of the SMILES grammar, which would save a vast amount of computational resources and time for our training.

Multimodal Fusion Module: Specifically, this component comprised three cross-layers, each of which consisted of three essential sub-layers: a self-attention sub-layer, a cross-attention sub-layer, and a feed-forward sub-layer. Obviously, the attention

mechanism is:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\text{scale}}\right) \cdot V. \quad (2)$$

The Query (Q), Key (K) and Value (V) matrices in the self-attention sub-layer were obtained from the same modality mapping:

$$h_{sa}^v = \text{Attn}(h_v, h_v, h_v), h_{sa}^s = \text{Attn}(h_s, h_s, h_s), \quad (3)$$

where h_{sa}^v and h_{sa}^s denoted the self-attention results of the molecular image and SMILES string, respectively. The cross-attention mechanism, similar to self-attention, differs in the source of the Q, K, and V matrices. The Q matrix is sourced from the current modality, while the K and V matrices are sourced from another modality. As shown in Fig. 1(b), the solid line inputs represent the current visual modality injecting SMILES information, while the dashed line inputs represent the opposite. Thus, in the cross-attention sub-layer, information from another modality could be integrated into their respective representation:

$$h_{ca}^v = \text{Attn}(h_{sa}^v, h_{sa}^s, h_{sa}^s), h_{ca}^s = \text{Attn}(h_{sa}^s, h_{sa}^v, h_{sa}^v), \quad (4)$$

where h_{ca}^v and h_{ca}^s were the cross-attention results of the image and SMILES string respectively. Finally, the corresponding features were obtained through inputting into the feed-forward sub-layer respectively:

$$Z^v = [z_{cls}^v, z_1^v, \dots, z_N^v], Z^s = [z_{cls}^s, z_1^s, \dots, z_N^s], \quad (5)$$

where Z^v and Z^s were the output of the molecular image and SMILES string after one layer of cross-layer, respectively.

C. Pre-Training Objectives

To be detailed, we jointly optimized three objectives during the pre-training phase, of which two tasks (Masked SMILES Modeling and Image-SMILES Matching) were widely used in vision-language pre-training (VLP) and the third task (Fingerprint Class Predicting) depended on the features of molecular data and referred to the implementation in ImageMol.

Masked SMILES Modeling: The MLM task was first introduced in NLP, while it has been shown to be greatly beneficial to improve much performance in VLP [49], [50], [51]. We adopted a masking strategy analogy to ChemBERTa-77M-MLM for SMILES sequences. In our case, given a molecular image-SMILES pair (v, s) , we utilized the “<MASK>” token to randomly mask each atomic tokens in SMILES string s with a probability of 15%, and trained ISMol to reconstruct the masked tokens s_m via the unmasked tokens $s_{\setminus m}$ and its corresponding image v . Thus, the training objective was to minimize the loss of the reconstructed SMILES sequence:

$$\mathcal{L}_{msm}(\theta) = E_{(v,s) \sim D} f(s_m | v, s_{\setminus m}). \quad (6)$$

A linear layer with an activation function with default parameters was used as the MSM decoder header. Since the frequency of atoms in the dataset had varies greatly [52], f was the focal loss function to mitigate the sample imbalance:

$$f = -\alpha_\theta (1 - p_\theta)^\gamma \log(p_\theta), \quad (7)$$

where θ was the trainable parameters of ISMol, p_θ was a predicted probability value, and α and γ were hyperparameters.

Image-SMILES Matching: In this task, ISMol needed to identify whether the given image-SMILES pairs matched or not. We provided the model with matched or mismatched image-SMILES pair (v, s) with equal probability. After the inference of the backbone of ISMol, we concatenated the visual vector z_{cls}^v and the SMILES vector z_{cls}^s as a fused representation of the both modalities, and then it was fed to a fully connected layer with a sigmoid function to predict a score between 0 and 1, where 0 meant that the image and SMILES string did not match and 1 meant that they did. Consequently, the task could be regarded as a binary classification problem with the objective loss of minimizing the negative log-likelihood:

$$\mathcal{L}_{ism}(\theta) = -E_{(v,s) \sim D} [\log P_\theta(y | v, s)], \quad (8)$$

where $y \in \{0, 1\}$ indicated whether the image-SMILES pair matched or not.

Fingerprint Class Predicting: Each molecule has a unique MACCS fingerprint, which is an abstract representation that transforms (encodes) the molecule into a series of bit characters. We clustered their fingerprints using K-Means algorithm into 100, 1000, and 10000 categories, with similar molecular structure information between the same categories. The results of clustering were used as pseudo-labels. Both the information contained in images and extracted from SMILES strings should have the same labels. For the visual features, the training objective was to minimize the loss:

$$\mathcal{L}_{fcp-v}(\theta) = -E_{(v,s) \sim D} [\log(P_\theta(y_{100}|v) \cdot P_\theta(y_{1000}|v) \cdot P_\theta(y_{10000}|v))], \quad (9)$$

where $y_{100} \in \{0, 1, \dots, 99\}$, $y_{1000} \in \{0, 1, \dots, 999\}$ and $y_{10000} \in \{0, 1, \dots, 9999\}$ were the different clustering pseudo-labels. Similarly, for the SMILES features, it should be consistent with another ones:

$$\mathcal{L}_{fcp-s}(\theta) = -E_{(v,s) \sim D} [\log(P_\theta(y_{100}|s) \cdot P_\theta(y_{1000}|s) \cdot P_\theta(y_{10000}|s))]. \quad (10)$$

Six different linear layers with default parameters were utilized as the FCP head for three categories within both modality. The final loss was the sum of \mathcal{L}_{fcp-v} and \mathcal{L}_{fcp-s} :

$$\mathcal{L}_{fcp} = \mathcal{L}_{fcp-v} + \mathcal{L}_{fcp-s}. \quad (11)$$

Joint loss: The overall pre-training loss could be calculated as the sum of the three aforementioned losses. Specifically, the total objective loss was:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{msm} + \lambda_2 \mathcal{L}_{ism} + \lambda_3 \mathcal{L}_{fcp}, \quad (12)$$

where λ_1 , λ_2 and λ_3 were trade-off parameters, so as to keep the loss of each task in the same order of magnitude. There were set to 1, 8, and 1, respectively.

D. Experimental Details

During the pre-training stage, each of image-SMILES pairs was fed to two encoders separately for extracting different features, and both features were aligned and fused in the fusion module. We trained and tested the proposed model with a batch size of 4096. A total of 25,000 training steps were conducted, accumulating gradients 64 times per step. We used AdamW optimizer with a learning rate linearly warming up to over 2500 steps and decaying to 0 with the polynomial schedule.

After pre-training, we retained the backbone of ISMol, but replaced the pre-training heads with a downstream head consisting of fully connected neural network layers. The receiver operating characteristic curve (ROC-AUC) was employed as the evaluation metric for the classification tasks. The model that exhibited the highest ROC-AUC values on the validation set was employed to assess the test set. The outcomes obtained from the test set substantiated the performance of ISMol. It should be noted that hyper-parameters had a significant impact on the downstream tasks. Therefore, we applied a grid search strategy for hyper-parameters in an actual network to ensure that ISMol achieved optimal results. Four hyper-parameters were chosen: the dropout rate of fine-tuning head, the batch size, the learning rate and the max steps of gradient update. Supplementary Table SII illustrates the search method and search ranges of hyper-parameters. To eliminate the influence of random factors, we selected three different random seeds to run experiments for each task and reported the average values as the final result. All experiments were executed using PyTorch-Lightning framework on an NVIDIA RTX A6000 GPU.

III. RESULTS AND DISCUSSION

A. Performance Comparison With Baseline Models

To verify the performance of ISMol, we compared it with seven competitive methods across the benchmark datasets.

TABLE I
PERFORMANCE OF ISMOL ON DRUG-DISCOVERY-RELATED DATASETS COMPARED TO OTHER METHODS

Dataset	Descriptor-based		Graph-based		SMILES-based		Image-based	Ours
	XGBoost-MACCS	XGBoost-ECFP4	HRGCN+	CD-MVGNN	ChemBERTa-77M-MTR	Knowledge-based BERT	ImageMol	ISMol
Pgb-sub	0.907±0.014	0.885±0.015	0.906±0.008	0.927±0.015	0.894±0.008	0.926±0.015	0.921±0.013	0.930±0.004
HIA	0.871±0.070	0.893±0.063	0.868±0.069	0.891±0.016	0.897±0.052	0.892±0.055	0.923±0.025	0.936±0.015
F(20%)	0.760±0.055	0.699±0.081	0.748±0.049	0.772±0.008	0.762±0.025	0.780±0.036	0.726±0.017	0.777±0.016
F(30%)	0.738±0.038	0.714±0.050	0.734±0.063	0.730±0.004	0.725±0.041	0.765±0.033	0.750±0.019	0.770±0.016
FDAMDD	0.838±0.039	0.824±0.033	0.832±0.050	0.829±0.009	0.815±0.046	0.844±0.037	0.829±0.008	0.826±0.017
CYP1A2-sub	0.786±0.070	0.766±0.072	0.769±0.075	0.807±0.066	0.799±0.073	0.774±0.095	0.870±0.088	0.892±0.025
CYP2C19-sub	0.672±0.135	0.680±0.146	0.740±0.104	0.879±0.059	0.762±0.022	0.732±0.127	0.823±0.003	0.893±0.032
CYP2C9-sub	0.723±0.039	0.744±0.042	0.744±0.058	0.736±0.033	0.735±0.015	0.732±0.047	0.746±0.024	0.811±0.036
CYP2D6-sub	0.798±0.039	0.772±0.034	0.800±0.040	0.811±0.047	0.797±0.052	0.808±0.03	0.819±0.010	0.826±0.023
CYP3A4-sub	0.792±0.040	0.776±0.044	0.775±0.041	0.792±0.025	0.796±0.008	0.776±0.047	0.801±0.003	0.855±0.005
T12	0.738±0.045	0.714±0.056	0.762±0.034	0.769±0.069	0.760±0.005	0.759±0.028	0.741±0.027	0.762±0.007
DILI	0.881±0.042	0.874±0.050	0.869±0.036	0.888±0.026	0.879±0.043	0.885±0.033	0.917±0.022	0.923±0.019
SkinSen	0.742±0.068	0.711±0.058	0.779±0.057	0.749±0.018	0.778±0.015	0.786±0.054	0.807±0.038	0.855±0.013
Respiratory	0.876±0.023	0.859±0.018	0.855±0.025	0.865±0.018	0.892±0.009	0.873±0.024	0.891±0.010	0.895±0.006
Average	0.794	0.779	0.798	0.818	0.807	0.809	0.826	0.854

Those methods could be classified into four categories: descriptor-based models, graph-based models, SMILES-based models, and image-based models. We provided a brief introduction of the methods, as below:

- XGBoost-MACCS is a traditional ML model based on molecular descriptor (MACCS fingerprints [17]).
- XGBoost-ECFP4 is a conventional ML model based on a different descriptor, namely ECFP4 fingerprints [18].
- HRGCN+ [53] combines molecular graph and descriptors, which can improve the predictive performance of the model.
- CD-MVGNN [22] modifies the interaction process between the atom-central view and bond-central view in the Multi-View Graph Neural Networks to improve the accuracy and robustness of predictions.
- ChemBERTa-77M-MTR [30] calculates 200 molecular properties for each compound by RDKit and uses these as labels to train a multi-task regression architecture to learn molecular knowledge.
- Knowledge-based BERT [29] proposes a multi-granularity pre-training model based on SMILES strings that aims to improve the ability to extract information from the SMILES sequences.
- ImageMol [37] pre-trains the ResNet network on a dataset of 10 million molecular images by setting multiple pretext tasks to enable it to extract chemical structures from molecular images.

The final results are shown in Table I, where the best results are highlighted in bold. Not surprisingly, methods based on molecular descriptors perform the worst. Typically, molecular fingerprints are used as auxiliary inputs to enhance the generalization capability, such as HRGCN+, rather than being used as the sole input for a model [28]. However, the combination of multiple representations does not necessarily lead to

a significant improvement in performance, and the efficiency of extracting and utilizing the information embedded in the representations is also crucial. Despite incorporating molecular descriptors into the graph representation, the performance of HRGCN+ (Average AUROC: 0.798) is not as effective as CD-MVGNN (Average AUROC: 0.818), which serves as evidence for this point. While ChemBERTa-77M-MTR (Average AUROC: 0.807) can infer potentially 200 or more properties of molecules, it falls short of encompassing their complete set of properties. Similarly, knowledge-based BERT (average AUROC: 0.809) analyzes the grammar rules of SMILES at different granularities, but its performance is limited due to its reliance on one-dimensional data and the absence of spatial features. In comparison, ISMol attains equal or better performance compared to the baseline models in 13 out of 14 datasets. Then, for statistical comparisons with competitive representations, we conducted T-tests (all obeying normal distributions), and the p-values are shown in Supplementary Table SIII. ISMol demonstrates statistical significance compared to graph-based (CD-MVGNN), and SMILES-based (ChemBERTa-77M-MTR) methods on five datasets. It also exhibits significant advantages compared to the image-based method (ImageMol) on three datasets. In particular, ISMol (Average AUROC: 0.854) achieves different degrees of improvement in the average AUROC compared to the image-based method (ImageMol, average AUROC: 0.826) and SMILES-based methods (ChemBERTa-77M-MTR, average AUROC: 0.807; Knowledge-based BERT, average AUROC: 0.809). This suggests that our previous hypothesis holds true, which is that the structural information in the image is complementary to the SMILES sequence information, thereby improving performance. ISMol exhibits outstanding performance on drug discovery-related datasets, making it as one of the competitive DL methods for drug discovery.

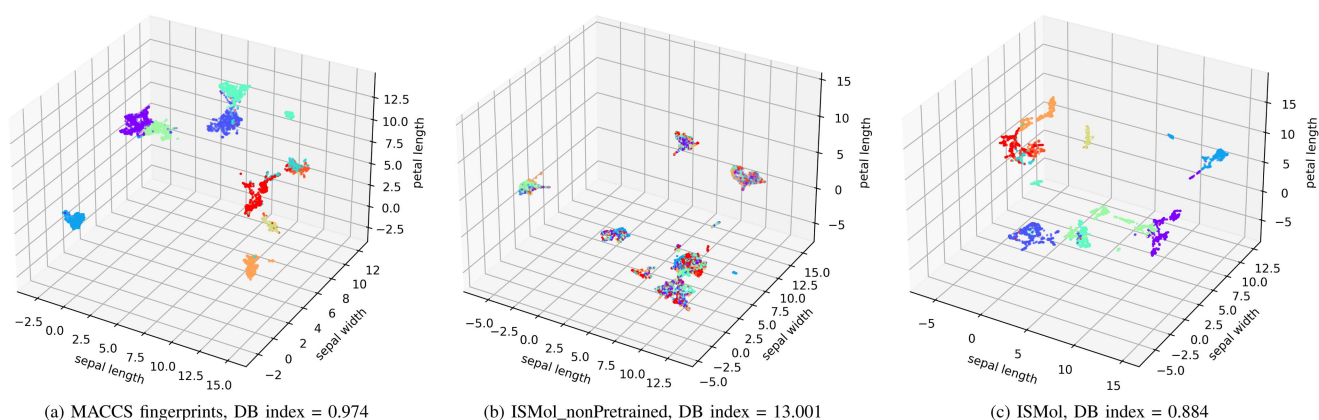


Fig. 2. Visualization of dimensionality reduction is performed on different molecular representations, with each color representing a different pseudo-label category.

B. Biological Interpretation of ISMol

To gain an understanding of the ability of ISMol to effectively extract and distinguish molecular structures, we briefly visualize the chemical space generated by ISMol. UMAP is a manifold learning algorithm known for its ability to reduce dimensionality while preserving the global structure of data [54]. Using the UMAP algorithm, we performed dimensionality reduction on MACCS fingerprints of those molecules, as well as on the embeddings generated by ISMol before and after the pre-training. After that, the Davies-Bouldin (DB) index was used to assess the quality of clustering results, which measured the similarity between each cluster and its most closely related cluster, and then computed the average of these similarities. We randomly selected 10000 molecules from ChEMBL and clustered their MACCS fingerprints, with the resulting cluster labels assigned as the molecules' pseudo-labels. We compared each molecule based on: (a) MACCS fingerprints, (b) embeddings generated by non pre-trained ISMol and (c) embeddings generated by pre-trained ISMol. Fig. 2 shows the visualized results, where the molecular features generated by pre-trained ISMol have a more discrete distribution than MACCS fingerprints, with ISMol (DB index = 0.884) outperforming MACCS fingerprints (DB index = 0.974). In addition, the pre-training strategy greatly improved the ability of ISMol to characterize the molecules (DB index = 13.001), which implied that the strategy brought with it gains for the model.

Then, to assess ISMol's biological interpretation within downstream datasets, we selected representative datasets from each of the four major property classes, including HIA in the absorption property class, CYP3A4-sub in the metabolism property class, SkinSen in the excretion property class, and T12 in the toxicology property class. As illustrated in Fig. 3, the DB index with the pre-trained ISMol are consistently lower than those obtained using MACCS fingerprints on the selected datasets. The results indicate that the chemical space generated by ISMol was more discriminative than that of MACCS fingerprints. Furthermore, comparing the results of pre-trained ISMol with non-pretrained ISMol, the pre-training strategy significantly optimized the chemical space of the model.

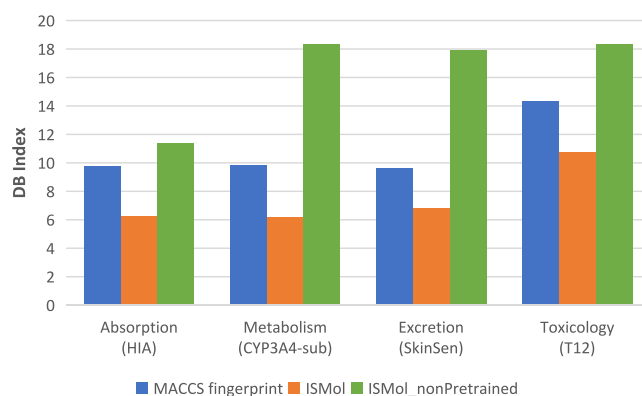


Fig. 3. Evaluation results of dimensionality reduction and clustering on the representations of four selected downstream task datasets. The lower the DB index value, the better the performance.

C. Ablation Experiments

We conducted ablation experiments on the input of ISMol. As ISMol required the use of image-SMILES pairs as inputs, its constraint was more stringent than those of the single-input methods. We were inspired to investigate whether ISMol could perform well with only one modality of information inputs. To tackle the problem of missing modality inputs, we utilized the placeholder approach for substitution. In particular, we used an empty string as a placeholder for SMILES when the correct image was fed (ISMol_onlyImage), and a blank image as a placeholder for the image when the correct SMILES string was entered (ISMol_onlySMILES). The results are shown in Table II. It is surprising that the performance of ISMol achieved comparable performance to the image-based method (ImageMol) or SMILES-based (Knowledge-based BERT) method when the input was only images (ISMol_onlyImages, average AUROC: 0.826) or SMILES strings (ISMol_onlySMILES, average AUROC: 0.820). This seems to indicate that ISMol was able to extract unimodal information more effectively after pre-training. Because it enhanced the ability to extract current modality information with another modality, even without corresponding

TABLE II
ABLATION EXPERIMENTS ON COMPONENTS AND INPUTS OF ISMOL

Dataset	ChemBERTa-77M-MLM	ViT	ISMol_nonPretrained	ISMol_onlySMILES	ISMol_onlyImages
Pgb-sub	0.874±0.029	0.835±0.010	0.879±0.026	0.916±0.013	0.898±0.01
HIA	0.877±0.013	0.892±0.023	0.891±0.027	0.905±0.007	0.927±0.027
F(20%)	0.755±0.059	0.633±0.026	0.642±0.018	0.733±0.018	0.744±0.01
F(30%)	0.703±0.056	0.605±0.031	0.678±0.012	0.722±0.025	0.715±0.007
FDAMDD	0.729±0.014	0.724±0.024	0.690±0.021	0.783±0.009	0.808±0.006
CYP1A2-sub	0.824±0.074	0.776±0.051	0.816±0.098	0.873±0.061	0.880±0.011
CYP2C19-sub	0.721±0.030	0.763±0.050	0.769±0.077	0.847±0.054	0.853±0.017
CYP2C9-sub	0.736±0.010	0.684±0.029	0.700±0.009	0.773±0.006	0.766±0.012
CYP2D6-sub	0.771±0.028	0.701±0.010	0.749±0.015	0.769±0.011	0.813±0.009
CYP3A4-sub	0.814±0.014	0.789±0.022	0.737±0.070	0.830±0.004	0.802±0.018
T12	0.687±0.005	0.721±0.016	0.722±0.008	0.741±0.011	0.734±0.009
DILI	0.848±0.021	0.833±0.045	0.875±0.022	0.902±0.018	0.910±0.041
SkinSen	0.771±0.048	0.773±0.087	0.793±0.088	0.817±0.037	0.825±0.099
Respiratory	0.785±0.027	0.782±0.047	0.813±0.008	0.866±0.015	0.893±0.010
Average	0.778	0.751	0.768	0.820	0.826

inputs during the fine-tuning stage. This once again confirms the validity of the hypothesis we previously mentioned, namely that the information from these two modalities could complement each other. Meanwhile, this implies that ISMol can perform well even under a single input condition.

We then conducted ablation experiments on the components of ISMol. We separately tested the performance of the visual encoder (ViT) and the SMILES encoder (ChemBERTa-77M-MLM), both of which were initialized with weights provided by the official release, as shown in Table II. The image-based method (ViT, average AUROC: 0.751), performed the worst. This could be attributed to the fact that simple pixel features contained weak chemical information, and the introduction of blank pixels further hindered the model to make correct inferences. The performance of ChemBERTa-77M-MLM (Average AUROC: 0.778) was also predictably unsatisfactory, as it struggled to capture the spatial structure of molecules well solely through the masking language modeling task. Intuitively, incorporating additional information would likely improve predictive performance. However, comparing the experiments of ISMol_nonPretrained (Average AUROC: 0.768), we found that simply concatenating the two modalities did not appear to improve the performance and may have even led to a decrease. The introduction of low-quality pixel features made the SMILES-based model more prone to learning induction bias, which could result in incorrect predictions. When the model had not been pre-trained, our previous assumption was wrong, which meant that hastily adding images not only did not improve predictive performance but instead reduced it.

D. Essential Pre-Training

The pre-training strategy yields remarkable results in NLP, CV, and other fields [29]. This strategy enables proficient

learning and performance on small-scale datasets. In most MPP tasks, there is only a limited amount of relevant training data available, which is insufficient for DL models to learn valuable patterns through normal training. Hence, pre-training a model on abundant unlabeled data to acquire molecular knowledge and then fine-tuning it on the actual tasks is a workable approach [30], [37]. By comparing the performance of pre-trained and non pre-trained ISMol on the downstream dataset, we observed an average performance improvement of 8.6% with the pre-training strategy. This proved the effectiveness of the approach in our case. The strategy enabled ISMol to extract critical features from molecular images and SMILES strings more efficiently and with greater robustness.

E. Generalizability Experiments

With the continual expansion of chemical databases, the generalizability of the model to new data becomes a crucial issue. We conducted a survey of the number and proportion of scaffolds that appeared in the pre-training dataset for the 14 ADMET datasets. The results are available in Supplementary Table SIV. In comparison to ImageMol, our dataset size was reduced by almost 2/3, yet the average scaffold coverage in the 14 downstream tasks decreased by only 14.7%. Additionally, our SMILES encoder was based on the ChemBERTa-77M-MLM model, which had been pre-trained on a dataset of 77 million molecules, contributing to its benefits for downstream tasks. The performance statistics in Table I also indicated that ISMol outperformed Image-based model. Therefore, the scaffold generalizability of ISMol surpasses that of ImageMol.

In addition, to further measure the performance of the models in terms of generalizability on retrospective and prospective, we evaluated the performance of different models on Epidermal Growth Factor Receptor (EGFR) dataset with retrospect splitting

TABLE III
PROSPECTIVE AND RETROSPECTIVE EXPERIMENTS ON ISMOL

Split Way	XGBoost-MACCS	CD-MVGNN	ChemBERTa-77M-MTR	ImageMol	ISMol
prospect	0.584±0.004	0.753±0.005	0.725±0.006	0.714±0.009	0.749±0.011
retrospect	0.674±0.002	0.761±0.006	0.938±0.002	0.887±0.003	0.972±0.002

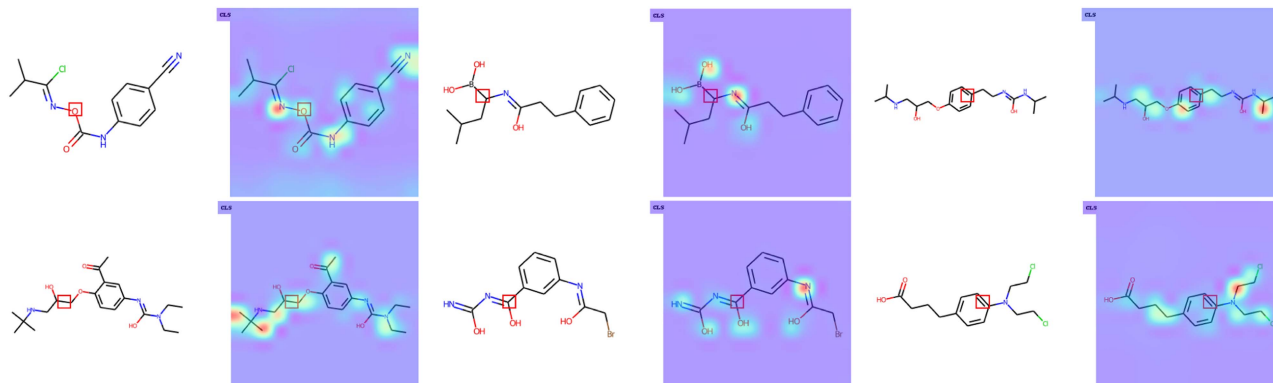


Fig. 4. Visualization results of attention map in visual feature extractor. On the left is the original image, and on the right is the visualization of the attention map with the CLS marker in its upper left corner and the red region represents a query.

and prospect splitting. In details, we collected 5,833 molecules targeting the EGFR from ChEMBL database, which is the most popular drug target and ranks first in innovative drug development globally [55]. We labeled a molecule as a positive sample if its IC₅₀ (the concentration at which the ratio of apoptotic cells to total cells is 50%) was less than 30 nM. Otherwise, the molecule was labeled as a negative sample. Subsequently, we searched for the ‘Molecule ChEMBL ID’ in the PubChem database and recorded its ‘create date’. Based on the timestamps, we split the data into training, validation, and test sets. Specifically, we used the earliest added molecules for the training set, the mid-time added molecules for the validation set, and the most recently added molecules for the test set. This type of splitting was referred to as the ‘prospect split’ and conversely, we referred to it as the ‘retrospect split’. The statistical results can be found in Supplementary Table SV. Subsequently, we fine-tuned each representative models based on different representations with retrospect splitting and prospect splitting to evaluate the performance retrospectively and prospectively. The results are shown in Table III. The best results are indicated in bold. We found that ISMol achieves or approaches the best performance under different time splitting settings, indicating its superior generalization to both unseen new data and historical data.

F. Explanatory Analysis

As we all know, in molecular images, the molecular structures are projected into Euclidean space, making it difficult to mine and fully leverage the information contained therein [28]. More than just performance improvement, we were particularly interested in how ISMol extracted and utilized structural information from molecular images. To this end, we visualized the attention

maps in our visual encoder. As shown in Fig. 4, warmer colors indicate greater attention allocated to the corresponding region within the attention mechanism, while cooler colors indicate lower attention. ISMol effectively extracted molecular structures by focusing on the spatial information of other atoms or functional groups around the queried area, rather than insignificant blank areas. Such a feature could potentially make the molecular image a valuable representation in fields such as chemistry and biology.

In general, the alignment and fusion of diverse data could help a model to enhance the comprehension of the data features, thereby reducing inductive bias and generalization errors [41]. For an intuitive interpretation of the alignment and fusion process of the two modalities in the fusion module, an exemplar molecule was selected randomly, with the SMILES string “O=[P+](O)c1ccc(Cl)cc1”. The molecule was converted into an image and input into ISMol along with its SMILES string. Then, we visualized the attention maps of the different cross-attention layers in the multimodal fusion module. For a brief description, we removed the attention scores of the nonsense patches, and the results were shown in Fig. 5. As the number of cross-attentive layers increased, the interaction between the two modalities gradually reinforced and became clearer. Multiple different vertical lines in the figures show that ISMol indeed paid more attention to the specific features of the other modality in the current modality. The complete attention maps can be seen in Supplementary Fig. S2.

Then, we further studied the results of aligning and fusing after the fusion module. We used specified SMILES sub-strings as query terms to calculate the similarity with all patches and selected those with high similarity for visualization. Table IV reveals the query results that it was essentially possible to

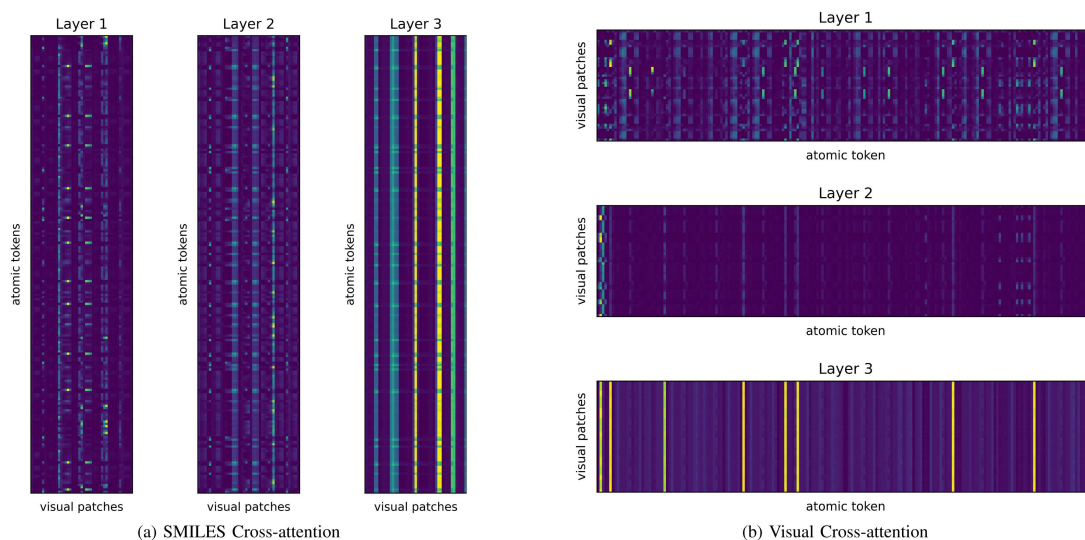


Fig. 5. Visualisation of cross-attention maps from the different layers in the multimodality fusion module. (a) is cross-attention visualization in the SMILES modality. (b) is the visualization of cross-attention in the visual modality.

TABLE IV
VISUALIZATION OF EXAMPLE MOLECULAR IMAGES ALIGNED WITH SMILES SUB-STRINGS

Original image	SMILES (with focus)			
	<chem>O=[P+](O)c1ccc(Cl)cc1</chem>	<chem>O=[P+](O)c1ccc(Cl)cc1</chem>	<chem>O=[P+](O)c1ccc(Cl)cc1</chem>	<chem>O=[P+](O)c1ccc(Cl)cc1</chem>

We highlight the queried substrings for emphasis with the focus style.

locate the corresponding pixels in the image when using different SMILES sub-strings for querying. The results indicate that IS-Mol achieved effective alignment and fusion of information from both SMILES strings and molecular images. More examples can be found in Supplementary Table SVI.

Finally, we investigated whether the pixel patches were injected with biochemical knowledge after the interaction with the SMILES features. A straightforward approach we adopted was to first cluster (and color) the visual hidden states, and subsequently overlay the clustering results onto the original image. As the final results are shown in Fig. 6, ViT could distinguish structural features based on pixels, but failed to separate out the white background. This drawback would cause ViT to reason based on blocks of nonsensical pixels, thus severely impairing its performance [37]. Instead, ISMol was able to discriminate well white backgrounds and isolate the structures of the molecule in the image. Meanwhile, the image patches were endowed with certain chemical structural features rather than just simple pixels. These results again demonstrated that ISMol not only effectively integrated information from both sources, but also provided complementary insights for each other.

G. Discussion

In this work, we place a substantial emphasis on the hypothesis of information complementarity between molecular images and SMILES sequences, and we innovatively propose the utilization of visual knowledge to enhance drug discovery. Conventional wisdom tends to undervalue molecular images as a representation primarily due to their explicit information being projected into Euclidean space, rendering it challenging to harness effectively. However, it is worth noting that two-dimensional structural information, encompassing geometric spatial features such as atomic distances and bond angles, becomes explicit. We endeavor to confer explicit semantics on this information by aligning and fusing it with SMILES sequences. This approach parallels human cognition, where, upon encountering an image, the visual system captures its structural features and, through existing cognitive systems, imparts concrete cognitive entities to the image to achieve a comprehensive understanding of a subject. To equip our model with alignment and fusion capabilities, we undertake pretraining on a large-scale dataset. Subsequently, we fine-tune the model on ADMET datasets and compare its

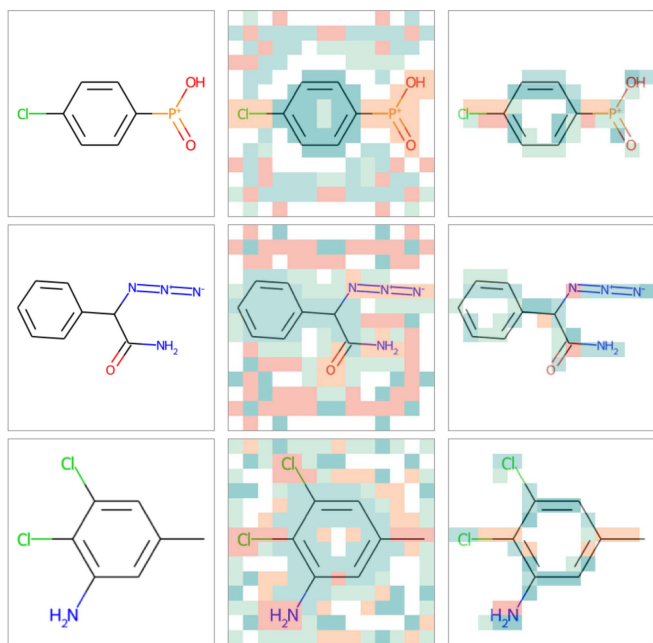


Fig. 6. Visualization of patch clusters for images as produced from ISMol (densely clustered patches). The first column shows the original images, the second column shows the clustering results of the patches generated by ViT, and the last column shows the clustering results of the patches generated by ISMol.

performance with the latest competitive models, showcasing superior results. Furthermore, we conduct an extensive array of experiments to substantiate the validity of our approach.

However, we posit that molecular images, as a novel representation, are not without their inherent challenges. The primary issue lies in the fact that simple pixel features lack explicit chemical information, as evidently illustrated in Fig. 4 on the left side with the original image. Notably, we employ the clustering results of the MACCS fingerprint to distill image features. This process is a clear attempt to imbue pixel features with chemical significance, prompting us to consider an additional question: if different fingerprints or explicit chemical knowledge were used to distill image features, could more explicit chemical information be conferred upon pixel features, thereby rendering structural features within the image explicit? This question stands as a subject for future investigation. Furthermore, molecular images contain nearly 95% blank space [37]. Without specialized training, models like ViT are susceptible to making misjudgments. However, the introduction of a substantial computational burden by blank areas remains a problem that is challenging to circumvent. Nevertheless, due to the rapid advancement in current computility, this problem is gradually becoming less critical.

IV. CONCLUSION

In this study, we placed emphasis on validating the hypothesis of complementarity between molecular images and SMILES strings, and proposed a pre-training model based on the both,

called as ISMol. ISMol performed favorably compared to the recently published seven methods on 14 public ADMET datasets. In response to the fast-growing chemical databases, we compared the generalizability of different models and found that ISMol outperformed others. This suggested that ISMol had great retrospective and prospective capabilities. We performed experiments with the ablation of different inputs and the components of input. Numerous experiments revealed that the alignment and fusion of these types of features after pre-training allowed them to complement each other, which brought a gain in performance improvement. In biological interpretation, we found the proposed pre-training strategies significantly optimized the discriminability of ISMol embedding vectors, which is better than MACCS fingerprints. It indicated that ISMol has robust biological interpretability. Importantly, we discovered that ISMol focused on neighboring atoms and bonds in images to extract molecular structures, and achieved alignment and fusion of the two types of features through the cross-attention mechanism. In general, ISMol explored the relationship between molecular images and SMILES strings and demonstrated promising performance in the field of drug development. In future work, we will explore the benefits of contrastive learning for molecular property prediction. Additionally, the consideration of other modalities of data to enhance the robustness of the model will be pursued.

REFERENCES

- [1] T. Gaudelot et al., "Utilizing graph machine learning within drug discovery and development," *Brief. Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab159.
- [2] I. V. Hinkson, B. Madej, and E. A. Stahlberg, "Accelerating therapeutics for opportunities in medicine: A paradigm shift in drug discovery," *Front. Pharmacol.*, vol. 11, 2020, Art. no. 770.
- [3] D. Sun, W. Gao, H. Hu, and S. Zhou, "Why 90% of clinical drug development fails and how to improve it?," *Acta Pharm. Sinica B*, vol. 12, no. 7, pp. 3049–3062, 2022.
- [4] T. Takebe, R. Imai, and S. Ono, "The current status of drug discovery and development as originated in United States academia: The influence of industrial and academic collaboration on drug discovery and development," *Clin. Transl. Sci.*, vol. 11, no. 6, pp. 597–606, 2018.
- [5] D. Giordano, C. Biancanello, M. A. Argenio, and A. Facchiano, "Drug design by pharmacophore and virtual screening approach," *Pharmaceuticals*, vol. 15, no. 5, 2022, Art. no. 646.
- [6] F. S. Willard and D. P. Siderovski, "Editorial [Hot Topic: GPCR high throughput screening (Part 1)](Guest Editors: David P. Siderovski and Francis S. Willard)," *Combinatorial Chem. High Throughput Screening*, vol. 11, no. 5, p. 336, 2008.
- [7] X. Li, Y. Lin, X. Meng, Y. Qiu, and B. Hu, "An l_0 regularization method for imaging genetics and whole genome association analysis on Alzheimer's disease," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3677–3684, Sep. 2021.
- [8] X. Li and D. Fourches, "Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MoLPMoFiT," *J. Cheminformatics*, vol. 12, no. 1, pp. 1–15, 2020.
- [9] Q. Ye et al., "Identification of active molecules against mycobacterium tuberculosis through machine learning," *Brief. Bioinf.*, vol. 22, no. 5, 2021, Art. no. bbab068.
- [10] L. Wei, X. Ye, T. Sakurai, Z. Mu, and L. Wei, "ToxiBTL: Prediction of peptide toxicity based on information bottleneck and transfer learning," *Bioinformatics*, vol. 38, no. 6, pp. 1514–1524, 2022.
- [11] L. Wei, X. Ye, Y. Xue, T. Sakurai, and L. Wei, "ATSE: A peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism," *Brief. Bioinf.*, vol. 22, no. 5, 2021, Art. no. bbab041.
- [12] W. Beker, A. Wołos, S. Szymkuć, and B. A. Grzybowski, "Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks," *Nature Mach. Intell.*, vol. 2, no. 8, pp. 457–465, 2020.

- [13] J. Sun et al., "Prediction of drug-likeness using graph convolutional attention network," *Bioinformatics*, vol. 38, no. 23, pp. 5262–5269, 2022.
- [14] K. V. Chuang, L. M. Gunsalus, and M. J. Keiser, "Learning molecular representations for medicinal chemistry: Miniperspective," *J. Med. Chem.*, vol. 63, no. 16, pp. 8705–8722, 2020.
- [15] M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, "Choosing feature selection and learning algorithms in QSAR," *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 837–843, 2014.
- [16] B. Chandrasekaran, S. N. Abed, O. Al-Attraqchi, K. Kuche, and R. K. Tekade, "Computer-aided prediction of pharmacokinetic (ADMET) properties," in *Dosage Form Design Parameters*. New York, NY, USA: Elsevier, 2018, pp. 731–755.
- [17] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," *Methods*, vol. 71, pp. 58–63, 2015.
- [18] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.
- [19] T. Chen et al., "Xgboost: Extreme gradient boosting," *R Package Version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.
- [20] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev, "Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 2048–2056, 2003.
- [21] B. Gellin, J. F. Modlin, and R. F. Breiman, "Vaccines as tools for advancing more than public health: Perspectives of a former director of the national vaccine program office," *Clin. Infect. Dis.*, vol. 32, no. 2, pp. 283–288, 2001.
- [22] H. Ma et al., "Cross-dependent graph neural networks for molecular property prediction," *Bioinformatics*, vol. 38, no. 7, pp. 2003–2009, 2022.
- [23] J. Wang et al., "DeepAtomicCharge: A new graph convolutional network-based architecture for accurate prediction of atomic charges," *Brief. Bioinf.*, vol. 22, no. 3, 2021, Art. no. bbab183.
- [24] X. Pan, H. Wang, C. Li, J. Z. Zhang, and C. Ji, "MolGpka: A web server for small molecule pKa prediction using a graph-convolutional neural network," *J. Chem. Inf. Model.*, vol. 61, no. 7, pp. 3159–3165, 2021.
- [25] M. Withnall, E. Lindelöf, O. Engkvist, and H. Chen, "Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction," *J. Cheminformatics*, vol. 12, no. 1, pp. 1–18, 2020.
- [26] H. Ma et al., "Multi-view graph neural networks for molecular property prediction," 2020, *arXiv:2005.13607*.
- [27] G. Li et al., "DeepGCNs: Making GCNs go as deep as CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6923–6939, Jun. 2023.
- [28] Z. Li, M. Jiang, S. Wang, and S. Zhang, "Deep learning methods for molecular representation and property prediction," *Drug Discov. Today*, vol. 27, 2022, Art. no. 103373.
- [29] Z. Wu et al., "Knowledge-based BERT: A method to extract molecular features like computational chemists," *Brief. Bioinf.*, vol. 23, no. 3, 2022, Art. no. bbac131.
- [30] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa-2: Towards chemical foundation models," 2022, *arXiv:2209.01712*.
- [31] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: A pre-trained transformer for computational chemistry," *Mach. Learn.: Sci. Technol.*, vol. 3, no. 1, 2022, Art. no. 015022.
- [32] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "Smiles-bert: Large scale unsupervised pre-training for molecular property prediction," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform.*, 2019, pp. 429–436.
- [33] B. Winter, C. Winter, J. Schilling, and A. Bardow, "A smile is all you need: Predicting limiting activity coefficients from SMILES with natural language processing," *Digit. Discov.*, vol. 1, no. 6, pp. 859–869, 2022.
- [34] D. Xue et al., "X-MOL: Large-scale pre-training for molecular understanding and diverse molecular analysis," *Sci. Bull.*, vol. 67, no. 9, pp. 899–902, 2022.
- [35] J. Jiang et al., "MultiGran-SMILES: Multi-granularity SMILES learning for molecular property prediction," *Bioinformatics*, vol. 38, no. 19, pp. 4573–4580, 2022.
- [36] D. D. Nogare, M. Hartley, J. Deschamps, J. Ellenberg, and F. Jug, "Using AI in bioimage analysis to elevate the rate of scientific discovery as a community," *Nature Methods*, vol. 20, no. 7, pp. 973–975, 2023.
- [37] X. Zeng et al., "Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework," *Nature Mach. Intell.*, vol. 4, no. 11, pp. 1004–1016, 2022.
- [38] Z. Zeng, Y. Yao, Z. Liu, and M. Sun, "A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 862.
- [39] X.-C. Zhang et al., "MG-BERT: Leveraging unsupervised atomic representation learning for molecular property prediction," *Brief. Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab152.
- [40] Z. Guo, W. Yu, C. Zhang, M. Jiang, and N. V. Chawla, "GraSeq: Graph and sequence fusion learning for molecular property prediction," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 435–443.
- [41] Z. Guo, P. Sharma, A. Martinez, L. Du, and R. Abraham, "Multilingual molecular representation learning via contrastive pre-training," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2022, pp. 3441–3453.
- [42] A. Gaulton et al., "ChEMBL: A large-scale bioactivity database for drug discovery," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1100–D1107, 2012.
- [43] J. J. Irwin et al., "ZINC20—a free ultralarge-scale chemical database for ligand discovery," *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 6065–6073, 2020.
- [44] W. Hu et al., "Strategies for pre-training graph neural networks," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [45] G. Xiong et al., "ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W5–W14, 2021.
- [46] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [47] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, and T. Laino, "Found in translation: Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models," *Chem. Sci.*, vol. 9, no. 28, pp. 6091–6098, 2018.
- [48] B. Ramsundar, "Molecular machine learning with DeepChem," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2018.
- [49] F.-L. Chen et al., "VLP: A survey on vision-language pre-training," *Mach. Intell. Res.*, vol. 20, no. 1, pp. 38–56, 2023.
- [50] X. Gao, Y. Wang, W. Hou, Z. Liu, and X. Ma, "Multi-view clustering for integration of gene expression and methylation data with tensor decomposition and self-representation learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 3, pp. 2050–2063, May/Jun. 2023.
- [51] X. Gao et al., "Multi-view clustering with self-representation and structural constraint," *IEEE Trans. Big Data*, vol. 8, no. 4, pp. 882–893, Aug. 2022.
- [52] Z. Xu, J. Li, Z. Yang, S. Li, and H. Li, "SwinOCSR: End-to-end optical chemical structure recognition using a swin transformer," *J. Cheminformatics*, vol. 14, no. 1, pp. 1–13, 2022.
- [53] Z. Wu et al., "Hyperbolic relational graph convolution networks plus: A simple but highly efficient QSAR-modeling method," *Brief. Bioinf.*, vol. 22, no. 5, 2021, Art. no. bbab112.
- [54] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, 2018.
- [55] S. Hu, "Current status of anti-EGFR agents," in *Novel Sensitizing Agents for Therapeutic Anti-EGFR Antibodies*, S. Hu, Ed. Cambridge, MA, USA: Academic Press, 2023, ch. 1, pp. 1–12.