Supplementary Information

Title: **Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework**

Zeng et al., *Nature Machine Intelligence* 2022

*To whom correspondence should be addressed:

Feixiong Cheng, Ph.D.

Lerner Research Institute, Cleveland Clinic, Ohio, USA

Tel: +1-216-444-7654; Fax: +1-216-636-0009

Email:chengf@ccf.org

Table of Contents

Supplemental Materials and Methods

A. Experiment setup

A.1 Downstream task details

We used four different types of datasets for molecular property prediction, drug metabolism prediction, compound-protein binding prediction and anti-viral activity prediction tasks.

Datasets of Molecular Property Prediction

Datasets. MoleculeNet[1] is a popular benchmark for molecular property prediction. Here, we used 8 classification datasets (BBBP, Tox21, ClinTox, HIV, BACE, SIDER, MUV and ToxCast) and 5 regression datasets (FreeSolv, ESOL, Lipophilicity, QM7 and QM9) from MoleculeNet to evaluate our ImageMol. All details for datasets are provided in **Table S1**. In these eight classification datasets, Tox21, ClinTox, SIDER, MUV and ToxCast are complex multiple binary classification tasks, which have 12, 2, 27, 17, 617 tasks and 7,831, 1,478, 1,427, 93,087 and 8,575 samples, respectively. The three remaining classification datasets (BBBP, HIV, and BACE) are single binary classification tasks with 2,039, 41,127, 1,513 samples respectively. In these five regression datasets, QM9 is a multiple binary classification dataset, which has 3 binary classification tasks with 133,885 samples. The four remaining datasets (FreeSolv, ESOL, Lipophilicity and QM7) are single

binary classification tasks with 642, 1,128, 4,200 and 21,786 samples, respectively. The details of 8 molecular classification datasets are described as follows:

- **BBBP** (Blood-Brain Barrier Penetration) dataset includes binary-classification records of barrier permeability properties between blood and brain of more than 2,000 compounds.
- **Tox21** (Toxicology in the 21st Century) is a dataset of compound toxicity, including qualitative toxicity measurements for 8k compounds on 12 different targets.
- **HIV** (Human Immunodeficiency Virus) dataset contains more than 40,000 records of whether the compound inhibits HIV replication for binary classification between active and inactive.
- **ClinTox** (Clinical trial Toxicity) dataset includes 1,491 drug compounds with known chemical structures for the binary classification between clinical trial toxicity (or absence of toxicity) and FDA approval status.
- **BACE** (BetA-seCretasE) dataset contains compounds that can be inhibitors of human β-secretase 1 (BACE-1).
- **SIDER** (Side Effect Resource) is a database of marketed drugs and adverse drug reactions (ADR). The version of the SIDER dataset in DeepChem classifies drug side effects into 27 system organ classes according to MedDRA's classification of 1,427 approved drugs.
- **MUV** (Maximum Unbiased Validation) group is a benchmark dataset

selected from PubChem BioAssay by applying a refined nearest neighbor analysis. The MUV dataset contains 17 challenging tasks of approximately 90,000 compounds, designed specifically to validate virtual screening techniques.

- **ToxCast** (Toxicity foreCaster) is an extended data collection of the same program as Tox21, providing toxicology data for large compound libraries based on in vitro high-throughput screening. The processed collection includes qualitative results of over 600 experiments on 8k compounds.

The details of 5 molecular regression datasets are described as follows:

- **FreeSolv** (Free Solvation) dataset is a collection of experimental and calculated hydration free energies and their experimental values for small molecules in water.
- **ESOL** (Estimated SOLubility) dataset is a regression dataset containing structures and water solubility data of compounds.
- **Lipophilicity** dataset collected from the ChEMBL database provides experimental results for 4200 compounds with respect to the octanol/water distribution coefficient (logD at pH 7.4), which is an important feature of drug molecules affecting membrane permeability and solubility.
- **QM7** (Quantum Machine 7) is a subset of GDB-13 (a database of nearly 1 billion stable and synthesizable organic molecules) that

records the calculated atomization energies of stable and synthesizable organic molecules, such as HOMO/LUMO, atomization energies, etc. It contains various molecular structures (such as triple bonds, cycles, amides and epoxy resins) and up to 7 heavy atoms C, N, O, and S.

- **QM9** (Quantum Machine 9) is a comprehensive dataset providing geometric, energetic, electronic, and thermodynamic properties for a subset of the GDB-17 database, including 134,000 stable organic molecules and up to 9 heavy atoms.

Comparison method. For a comprehensive comparison, we selected several different types of popular methods, which are the fingerprint-based method (AttentiveFP[2]), the sequence-based methods (TF_Robust[3] and X-MOL[4]), the graph-based methods (GraphConv[5], Weave[6], SchNet[7], MPNN[8], DMPNN[9], MGCN[10], Hu et al.[11], N-GRAM[12], MolCLR[13] , GCC[14], GPT-GNN[15], Grover[16], MGSSL[17], 3D InfoMax[18], G-Motif[16], GraphLoG[19], GraphCL[20], GraphMVP[21] and MPG[22]) and the molecular image-based method (Chemception[23]). These recently proposed methods show competitive results and superior performance on molecular property prediction task. Therefore, we selected these representative methods for comparison. In fingerprint-based methods, AttentiveFP uses an attention mechanism to extract molecular fingerprints for interpretable property prediction. In the sequence-based methods, TF_Robust is a deep neural network-based multitasking model; X-MOL is a transformer-

based model, which is pre-trained on 1.1 billion molecules. In the graph-based methods, Hu et al., N-GRAM[12], MolCLR, GCC, GPT-GNN, GROVER, MGSSL, 3D InfoMax, G-Motif, GraphLoG, GraphCL, GraphMVP and MPG are graph representation learning methods based on self-supervised learning. Within molecular image-based method, Chemception[23] has a well-designed CNN architecture focused on molecular property prediction. To quantitatively compare the advantages and disadvantages of ImageMol and these methods, ROC-AUC score is calculated as the evaluation metric.

Experimental setting. Due to the differences in data split between different methods, for fair comparison, we used multiple different data split ways to comprehensively evaluate our ImageMol. Currently, the scaffold split [17, 18, 21] and random scaffold split[12, 16, 22] are the mainstream and popular splitting methods, which is a challenging and realistic evaluation setting because molecular substructures do not overlap between training and test sets. Therefore, we evaluated the performance of ImageMol on both split methods, which split dataset to 8/10 training set, 1/10 validation set and 1/10 test set. To filter the effect of pretraining data differences on the results, we also re-pretrained MPG (called MPG-10M), GROVER (called GROVER-10M) on the 10 million molecules used by ImageMol and fine-tuned on the MPP task. In addition, in order to compare with Chemception[23], we use exactly the same experimental configuration as Chemception, which uses stratified split to divide 4/6 training set, 1/6 validation set and 1/6 test set. In order to

compare fairly with X-MOL[4], we reproduced the results of X-MOL under scaffold split, which has the same experimental setup as ImageMol. The final AUC performance was reported by calculating the mean and standard deviation of the experimental results from 3 independent runs with different random seeds. The details of hyperparameter optimization for training MPG-10M, GROVER-10M and X-MOL can be found in Table **S31(a)-(c)**.

Datasets of Drug Metabolism Prediction

Datasets. In drug discovery, Cytochrome P450 inhibitors and noninhibitors classification is important for predicting the tendency of molecules to cause significant drug interactions by inhibiting CYP and to determine which subtypes are affected. In this task, we use PubChem Data Set I (Training Set) and PubChem Data Set II (Validation Set) from[24] to evaluate the performance of the proposed ImageMol on human cytochrome P450 (CYP) inhibition. PubChem Data Sets I and II are two-category datasets and both of them include CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 isoforms (Table **S2**). In addition, we also combine the five separate tasks (1A2, 2C9, 2C19, 2D6, and 3A4) of PubChem Data Set I into a multi-labeled classification problem to evaluate the performance of ImageMol in multi-labeled scenarios.

Comparison method. We compared the proposed ImageMol with three latest molecular image-based methods (Chemception[23], ADMET-CNN[25] and

QSAR-CNN[26]) with ROC-AUC metric to confirm the superiority of our method on molecular images and other molecular fingerprinting-based methods (MACCS-based and FP4-based methods[24]) with accuracy and ROC-AUC metrics to validate that our method can learn more information from molecular images than molecular fingerprints. We also compare ImageMol with sequence-based methods (RNN_LR, TRFM_LR, RNN_MLP, TRFM_MLP, RNN_RF, TRFM_RF[27] and CHEM-BERT[28]) and graph-based methods ($MolCLR_{GIN}$, $MolCLR_{GCN}$[13] and GROVER[16]) with more evaluation metrics (Accuracy, ROC-AUC, AUPR, F1, precision, recall, kappa) to verify the advantages of ImageMol. In molecular image-based methods, ADMET-CNN successfully established a molecular 2-D image-based CNN model and achieved good prediction performances on predicting the ADMET properties (including CYP1A2 inhibitory potency, P-gp inhibitory activity, etc.); QSAR-CNN applied transfer learning and data augmentation to train molecular image-based DenseNet121[29] model for developing quantitative structure-activity relationships (QSARs) to predict compound rate constants toward OH radicals. In molecular fingerprinting-based methods, two types of methods are used in the comparison, which includes traditional machine learning methods (SVM, C4.5 DT, k-NN and NB) and ensemble learning methods (CC-I, CC-II, etc.) respectively. In sequence-based methods, SMILES transformer[27] used RNN (Recurrent Neural Network)[30] and TRFM (TRansForMer)[31] to extract molecular representations and use LR (Logistic Regression)[32], MLP (Multi-

Layer Perception)[33] and RF (Random Forest)[34] as classifiers for downstream tasks. CHEM-BERT applied a pre-training task of BERT[35] on 9 million unlabeled molecules SMILES selected from ZINC[36] database. In graph-based methods, GROVER is a self-supervised message passing transformer, which is pre-trained on 10 million unlabelled molecules with node-level, edge-level and graph-level tasks. MolCLR developed GIN (Graph Isomorphism Network)[37] or GCN (Graph Convolutional Network)[5] encoders to learn differentiable representations on large unlabeled data (~10 million unique molecules) with three molecule graph augmentations (atom masking, bond deletion, and subgraph removal).

Experimental setting. For fairness, we keep the experimental settings consistent with these methods. When compared with fingerprinting-based and image-based methods, we first use 5-fold cross-validation on PubChem Data Set I to evaluate our performance, and then used the model trained in PubChem Data Set I to evaluate our performance on the external validation set PubChem Data Set II. When compared with sequence-based and graph-based methods, we used PubChem Data Set I to evaluate the performance of ImageMol with balanced scaffold split[16], which split the dataset to 80% training set, 10% validation set and 10% test set. Compared to scaffold split and random scaffold split, balanced scaffold split is a more scientific way to split data, which considers balancing sizes of scaffolds in train set, validation set and test set, rather than just putting the smallest in test set. The multi-

labeled learning is a more challenging setting, so we considered combining these five independent tasks (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) in PubChem Data Set I into one multi-labeled data for evaluation with balanced scaffold split. For reproduced methods, details of hyperparameter optimization can be find in Table **S31**.

Datasets of Compound-Protein Binding Prediction

Datasets. The top 10 G protein coupled receptors (GPCRs) datasets with the largest number of reported ligands (Table **S3**) from ChEMBL database (https://www.ebi.ac.uk/chembl/) and 10 KinomeScan datasets (Table **S4**) are used to predict drug-protein binding affinity (both regression task and classification task). In drug-kinase binding activity task, we used 10 common biochemical kinase profiling assays from KinomeScan data (https://lincs.hms.harvard.edu/kinomescan/). KinomeScan reports the "percent of control" of molecules binding to each kinase, where the control is DMSO and a 100% result means no inhibition of kinase binding to ligand in the presence of the compound, and a low percentage result means strong inhibition. Therefore, we use control as the criterion of activity, control=100% is inactive (non-inhibitor) and control<100% is active (inhibitor).

Comparison method and experimental setting. As in the setting of drug metabolism prediction under balanced scaffold split, we maintained the same setting in drug-protein binding.

Datasets of Anti-Viral Activity Prediction

Datasets. Anti-viral activities prediction is vital for the development of new drugs to treat COVID-19. We used anti-SARS-CoV-2 activities prediction as our task to prioritize compounds when screening in vitro. The experimental datasets are obtained from the COVID-19 portal[38] in the National Center for Advancing Translational Sciences (NCATS), which include 13 assays such as Spike-ACE2 protein-protein interaction (AlphaLISA), Spike-ACE2 protein-protein interaction (TruHit Counterscreen), ACE2 enzymatic activity, etc. These 13 assays represent five distinct categories: viral entry, viral replication, live virus infectivity, counterscreen and in vitro infectivity. Due to the extreme imbalance in these original datasets, the proportion of positive samples in the total samples ranges from 0.7% to 7.3%, so we filter out those samples without AC_{50} to generate our datasets and set AC_{50} greater than 10 and less than 10 as non-inhibitors and inhibitors, respectively. The overview of the processed datasets is summarized in Table **S18**. Each dataset contains binary-classification records of whether to inhibit SARS-CoV-2 activity. In addition, for a fair comparison with other method, we also used 11 existing SARS-CoV-2 datasets in REDIAL-2020[38] to train some models of anti-SARS-CoV-2 activities, and its statistical information is shown in Table **S27**.

Comparison method. We chose two representative methods for

experimental comparison, Jure'GNN[11] and REDIAL-2020[38]. Jure'GNN is a pre-training method based on graph and graph neural network (GNN), which used molecular graph as the input data of the GNN and introduced a series of pre-training strategies to train the GNN to obtain better molecular embedding. REDIAL-2020 is a suite of computational models based on manual features, which extracts a total of 22 features of three different types (19 fingerprints-based, 1 pharmacophore-based and 2 physicochemical descriptors-based) to train the machine learning model from scikit-learn package. In this task, we used a total of 6 evaluation metrics, namely accuracy, sensitivity, precision, ROC-AUC, AUPR and F1. In addition, like the drug metabolism prediction task, we also evaluated the performance of ImageMol under balanced scaffold split and compared ImageMol with more sequence-based (RNN_LR, TRFM_LR, RNN_MLP, TRFM_MLP, RNN_RF, TRFM_RF and CHEM-BERT) and graph-based methods ($MolCLR_{GIN}$, $MolCLR_{GCN}$ and GROVER).

Experimental setting. In order to compare our ImageMol with Jure's GNN, we reproduced Jure's GNN by using the public source code they provided to extract molecular features and added a fully connected layer for fine-tuning on downstream tasks. We uniformly split these datasets into 80% training set and 20% test set, and report the AUC and AUPR results on test set. We also compared our method with REDIAL-2020. To compare fairly with REDIAL-2020, we use the same experimental configuration as REDIAL-2020[38]. Note that REDIAL-2020 provides a new data preprocessing method and divides the

training set, validation set and test set, so we directly use these divided datasets to perform our evaluation process (Table **S27**). For the experimental results, we use the model that achieves the best performance on the validation set to evaluate the results of the test set. Finally, accuracy, F1, sensitivity, precision and ROC-AUC metrics are reported in the experiment. Under balanced scaffold split, the details of experimental setting are the same as for drug metabolism prediction task.

A.2 Selection of K in *K-means*

In the clustering pseudo-label classification task, we determined the K values to be 100, 1000, and 10000, respectively. In order to determine the value of K in K-Means method, we first use different K values, ranging from 1 to 14000, to cluster the dataset and to calculate the sum of squared distances. Then, we use the K value as the x-axis and sum of squared distances as the y-axis to draw a curve. Finally, a knee point detection algorithm [39] is used to find the knee point of this curve. As shown in Figure **S28**, the dotted line indicates the K value corresponding to the "elbow" point. Obviously, the larger the K value, the more difficult it is for ImageMol to perform the clustering pseudo-label classification task. Therefore, we select two K values (K =100 and 1000) on the left side of the "elbow" point and one K value (k=10,000) on the right side of the "elbow" point.

A.3 Hyperparameters of pre-training and finetuning

The hyperparameters of the pre-training and fine-tuning process are shown in Table **S30**. In the pre-training task, our model is pre-trained by SGD optimizer with learning rate 0.01, weight decay 10^{-5}, momentum 0.9 and batch size 256 for approximately 6 days on the Amazon server of the instance p3.16xlarge with 8 Tesla V100 GPU (32G). In downstream task, the pre-trained model is fine-tuned using SGD optimizer with batch size [8, 16, 32, 64, 128], learning rate from 5e-4 to 0.5, weight decay 10^{-5}, momentum 0.9 and epoch from 10 to 60 on Ubuntu 18.04.1 with Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz and Tesla T4 (16GB).

B. Supplementary Methods

B.1 Molecular image and fingerprint generation

In this study, we use image as molecular representation. Before molecular image generation, we first removed molecules that cannot be resolved by RDKit (https://github.com/rdkit/rdkit) and contains disconnected ions or fragments. We then removed salts, isotopes and stereochemical information from the SMILES sequences and the remaining molecules were further standardized by charge neuralization. Finally, we only kept drug-like molecules with logp (lipid-water partition coefficient) between -5 and 7, molecular weight between 12 and 600, and number of heavy atoms between

3 and 50 and transformed them to canonical SMILES. During molecular image generation, we used RDKIT's MolsToGridImage method to convert SMILES into standard and unique image. The MolsToGridImage method keeps the aromaticity of all bonds and atoms. 2D coordinates of each atom as orientation information are added, while none of the rotatable bonds are flipped. Unlike molecular graph, molecular image is composed of a pile of pixels rather than vertices and edges. In detail, we first filter out molecules without SMILES information in the original dataset. Second, we transform the SMILES sequences to molecular images using RDKit and set the image size to 224 × 224. Finally, these molecular images with the same size will be used as the initial dataset of our method. Considering that molecular fingerprints are easy to obtain and can express some priori knowledge of molecules, we chose *MACCS keys* to assist our pre-training process to make our model learn molecule-related priori knowledge. The *MACCS keys* are one of the commonly used structural molecular fingerprint [40], which contain 166 keys related to molecular structure. In our work, we used RDKit to generate a distinct 166-D molecular fingerprint for each molecule.

B.2 Pre-task details in pre-training

This section will describe the pre-training details of ImageMol with five pre-tasks. The overall data flow of the ImageMol framework during training is shown in Figure **S30**. In general, the original input images X is processed into

three different datasets. Augmented images X^{aug} is obtained by using data augmentation on X, including *RandomHorizontalFlip()*, *RandomGrayscale(p=0.2)* and *RandomRotation(degrees=360)* in torchvision. Shuffled images X^{Jig} is obtained by performing a jigsaw puzzle on X^{aug}. The puzzle rule uses "permutations 100" in [41]. Masked images X^{mask} is obtained by adding the mask matrix in X^{aug}, and the values in the matrix are filled with the mean value. The examples about masked images are shown is Figure **S29**. Then, randomly select a batch of data from these three datasets and input them into ResNet18 without classification layer to extract 512-D latent features $z^{aug}, z^{jig}, z^{mask}$. Finally, these latent features are input into the sub-network for each task for further processing.

Figures **S1-S5** show the architecture of each pre-training strategy. In multi-granularity chemical clusters classification (MG3C) task (Figure **S1**), chemical fingerprints are first extracted from SMILES and input into unsupervised KMEANS with different K values to produce clusters with different structure granularity. Then, these clusters are treated as pseudo-labels of molecular images. Finally, the molecular encoder and structural classifier are jointly used to predict the labels of molecular images and optimizing the loss between pseudo-labels and predicted labels in pre-training. The structural classifier is a multi-task learner that receives 512-dimensional features as input and then forward-propagates to 3 fully connected layers with different numbers of neurons (100, 1000 and 10000) for

classifying different clustering granularity.

In molecular rationality discrimination (MRD) task (Figure **S2**), we first disrupt the molecular structure to construct an irrational molecular image, which uses a 3x3 grid to decompose the molecular image into 9 patches and randomly shuffle them to form an irrational image. The original images are viewed as rational molecular images. Then, these rational and irrational molecules will be input to the molecular encoder to extract 512-D features. Finally, these features are forward propagated to rationality classifier for rationality judgment. The rationality classifier is a simple MLP structure that takes 512-dimensional feature as input and directly outputs 2-dimensional results (rational or irrational).

In jigsaw puzzle prediction (JPP) task (Figure **S3**), similar to MRD, we first decompose the molecular image into 9 patches and label the original permutations $(1, 2, 3, 4, 5, 6, 7, 8, 9)$. Then, we randomly shuffle the permutation and re-stitch into new images like $(7, 1, 6, 2, 0, 5, 4, 3, 8)$ or $(7, 8, 5, 6, 3, 2, 0, 1, 4)$. In particular, we randomly select from 100 defined permutations, which can be obtained from permutations_100.npy (https://github.com/fmcarlucci/JigenDG/blob/master/permutations_100.npy). Finally, the Molecular encoder is used to extract features of rearranged images and subsequently input into the jigsaw classifier for predicting the permutation (100 classification). The Jigsaw classifier is a simple MLP, which consists of a 512-dimensional input layer and a 100-dimensional output layer.

In MASK-based contrastive learning (MCL) task (Figure **S4**), we randomly mask a 16×16 region in the molecular image, which is filled using the mean of the image (some masked examples in Figure **S29**). Subsequently, image pairs (original image, masked image) are fed into the molecular encoder to extract features and maximize the similarity. Here, the Euclidean distance is used to constrain the similarity between two features, and we should minimize the Euclidean distance to ensure greater similarity.

In molecular image reconstruction (MIR) task (Figure **S5**), we build our GAN model based on context encoders [42]. The detail of GAN model is described in Figure **S5**. In generator, firstly, the latent features z^{aug} are forward to a single hidden layer MLP model, which accepts 512-d input and obtains a 128-d output. Subsequently, four ConvTranspose2D layer with BatchNorm2D and ReLU are used. In ConvTranspose2D, the numbers represent input channels, output channels, kernel size and stride respectively. Finally, a ConvTranspose2D layer with Tanh activation function is used to generate 64×64 images. In discriminator, X^{aug} is first preprocessed to resize to 64×64. Then resized X^{aug} and X^{rec} are input to a Conv2d with LeakyReLU and three Conv2d with BatchNorm2D and LeakyReLU (negative slope is 0.2). In Conv2d, the numbers have the same meaning as ConvTranspose2D. Finally, a Conv2d is used to discriminate the real or fake of input images.

C. Supplementary Results and Discussion

C.1 Results on the pre-training

As shown in Figure **S31**, it shows the details of the loss change of ImageMol during pre-training. We did not show the training details of the Image reconstruction task because the loss is adversarial. In general, the loss of ImageMol in the remaining four pre-tasks is a decreasing trend and gradually converges, which shows that our ImageMol can learn different information about molecular images in these pre-tasks.

C.2 Performance comparison with existing approaches

Baselines. We compared ImageMol with a large number of baselines on multiple tasks (molecular property prediction, drug metabolism prediction and anti-SARS-CoV-2 activities prediction) and different experimental settings (stratified split, scaffold split, random scaffold split, balanced scaffold split). We compared ImageMol with four different types of models, which are fingerprint-based models (AttentiveFP [2], MACCS-based and FP4-based methods [including SVM, C4.5 DT, k-NN, NB, CC-I, CC-II, etc.] [24], REDIAL-2020), sequence-based models (TF_Robust [3], RNN_LR, TRFM_LR, RNN_MLP, TRFM_MLP, RNN_RF, TRFM_RF [27] and CHEM-BERT [28]), graph-based models (GraphConv [5], Weave [6], SchNet [7], MPNN [8], DMPNN [9], MGCN [10], Hu et al. [11], N-GRAM [12], MolCLR [13] , GCC [14], GPT-GNN [15], Grover [16], MGSSL [17], 3D InfoMax [18], G-Motif [16], GraphLoG [19], GraphCL [20], GraphMVP [21]

and MPG [22]) and image-based models (Chemception [23], ADMET-CNN [25] and QSAR-CNN [26]), respectively. Results of most baselines were obtained from their original papers, except for these cases, such as: (1) under scaffold split, we reproduced the results of MolCLR (using GIN with the best performance) because it uses a different experimental setup (without considering chirality); (2) under balanced scaffold split, we reproduced the results of RNN_LR, TRFM_LR, RNN_MLP, TRFM_MLP, RNN_RF, TRFM_RF, CHEM-BERT, $MolCLR_{GIN}$, $MolCLR_{GCN}$ and GROVER because they did not run on CYP450 and SARS-CoV-2 datasets; (3) We reproduced Chemception, ADMET-CNN and QSAR-CNN results as they differ from our experimental setup (including dataset and split).

Fingerprint-based models. We selected several state-of-the-art fingerprint-based methods (AttentiveFP [2], traditional models and their ensemble models based on MACCS and FP4 [24] and REDIAL-2020 [38]) on three finetuning tasks. ImageMol achieved better performance compared to AttentiveFP on all benchmark datasets (including classification and regression tasks) with an average improvement of 8.0% and a low average standard deviation of 0.6% in classification task (Table **S6**). The traditional machine learning models include SVM, C4.5 Decision Tree (DT), k-Nearest Neighbors (KNN) and Naive Bayes (NB) and the ensemble models includes five different combinations of SVM, C4.5 (DT), KNN, NB and three ensemble strategies (Mean, Maximum,

Multiply). We found that ImageMol can outperform these methods on almost all benchmarks (Figure **2.f** and Table **S9**) with an improvement from 0.5% to 3.7%, which shows the features extracted by ImageMol are richer than manual features. Compared with REDIAL-2020, ImageMol achieves state-of-the-art performance on almost all evaluation metrics with average improvements of 4.0% (ACC), 5.8% (F1), 9.3% (SEN), 0.9% (PREC) and 5.5% (AUC) (Table **S20**).

Sequence-based models. Due to the simplicity and efficiency of the Simplified Molecular-Input Line-entry System (SMILES) sequence, it has become one of the most popular molecular representation [4, 43]. We compared our ImageMol with several popular pre-training models (TF_Robust [3], RNN_LR, TRFM_LR, RNN_MLP, TRFM_MLP, RNN_RF, TRFM_RF [27], CHEM-BERT [28] and X-MOL [4]) in benchmark datasets. The performance of our ImageMol can outperform the state-of-the-art sequence-based pre-training models on molecular property prediction task with an absolute improvement in ROC-AUC ranging from 2.2% to 21.0% (Figure **2.d** or Table **S5** and Figure **2.e** or Table **S6**), drug metabolism prediction task with an absolute improvement in ROC-AUC ranging from 1.1% to 4.7% on single task setting (Table **S10**) and 3.5% to 16.4% on multi-labeled setting (Table **S11**), drug-protein binding prediction task with an absolute improvement in RMSE ranging from 0.003 to 4.600 on efficiency regression prediction (Table **S12**)

and in ROC-AUC ranging from 0.9% to 56.4% on activity classification prediction (Table **S13**), and anti-SARS-CoV-2 activities prediction task with an average ROC-AUC improvement of 3.7% ranging from 1.4% to 11.2% (Table **S21**). Especially, ImageMol outperforms X-MOL with statistical significance below 0.05 on the SIDER and ToxCast datasets (Table **S14(a)**). These results show that molecular image-based representation has obvious advantages compared with sequence-based molecular representation because these models can only learn 1D sequence information but lacks 2D structural information.

Graph-based models. Considering that molecules can be naturally represented as graphic structures, some graph-based methods [11, 16] have recently emerged to learn the 2D topological structure information of molecules. We compared our ImageMol with multiple graph-based pre-training models, including GraphConv [5], Weave [6], SchNet [7], MPNN [8], DMPNN [9], MGCN [10], Hu et al. (Jure's GNN) [11], N-GRAM [12], MolCLR [13] , GCC [14], GPT-GNN [15], Grover [16], MGSSL [17], 3D InfoMax [18], G-Motif [16], GraphLoG [19], GraphCL [20], GraphMVP [21] and MPG [22]. The performance of ImageMol comprehensively exceeds the state-of-the-art graph-based pre-training models on molecular property prediction task (Figure **2.d** or Table **S5** and Figure **2.e** or Table **S6**), drug metabolism prediction task with an absolute improvement in ROC-AUC ranging from 0.4% to 2.1% on single task setting

(Table **S10**) and 0.5% to 2.3% on multi-labeled setting (Table **S11**), drug-protein binding prediction task with an absolute improvement in RMSE ranging from 0.004 to 0.179 on efficiency regression prediction (Table **S12**) and in ROC-AUC ranging from 6.0% to 38.9% on activity classification prediction (Table **S13**), and anti-SARS-CoV-2 activities prediction task with an average ROC-AUC improvement of 2.0% ranging from 0.1% to 12.8% (Table **S21**). It is worth noting that ImageMol achieves slightly worse performance than MPG and GROVER on the Lipophilicity dataset in Table **S6**. According to thermodynamic theory, Lipophilicity is associated with hydrogen bonds [44], whereas we do not explicitly encode hydrogen atoms and hydrogen bonds in ImageMol. Therefore, the performance of ImageMol in Lipophilicity will be improved by explicitly encoding hydrogen atoms and hydrogen bonds in molecular images. Considering the impact of pre-training data differences on the results, we show the performance of models (MolCLR, MPG-10M, GROVER-10M) with the same pretrained dataset as ImageMol. We found that ImageMol still achieves better performance compared to these methods (Table **S5** and Table **S6**). Furthermore, the performance advantage of ImageMol is statistically significant on the BBBP, ClinTox, HIV, SIDER, Tox21 and ToxCast datasets (Table **S14**). These results show the advantage of using molecular images as a representation. Although both molecular graph and image are based on 2D representation, they are significantly different in representation type. The molecular graph focuses on topological information

at the atomic level, while the molecular image focuses on the spatial structure information at the pixel level. In spatial structure, more rich information is included, such as the shape of molecules, the angle of chemical bonds, and the relative distance between atoms, etc.

Image-based models. We selected several latest molecular image-based models as comparison methods, which are Chemception, ADMET-CNN and QSAR-CNN respectively. We find that our ImageMol has high performance and outperforms the state-of-the-art methods on HIV and Tox21 datasets with an ROC-AUC improvement ranging from 7.4% to 9.2% (Figure **2.b** and Figure **S9**), CYP isoforms training sets (PubChem Data Set I) with an average ROC-AUC improvement of 8.5% ranging from 6.3% to 10.3% (Figure **S10**) and CYP isoforms validation sets (PubChem Data Set II) with an average ROC-AUC improvement of 11.2% ranging from 3.6% to 14.0% (Figure **2.c**). It is worth noting that the advantage of ImageMol is statistically significant compared to all image-based methods (Table **S15**). Especially, we also observed a similar performance between ImageMol_NonPretrain and Chemception, which is 73.2% vs. 72.2% and 73.4% vs. 75.2% on the HIV and Tox21 datasets respectively (Figure **S9**). However, after pre-training on 10 million molecular images, our ImageMol showed a significant improvement on the HIV (an increase of 8.2%) and Tox21 (an increase of 9.2%) with an average increase from -0.4% to 8.3%, which proves the effectiveness and

superiority of our pre-training strategies for molecular images.

C.3 Supplementary Results on anti-SARS-CoV-2 targets

The Table **S19-S21** showed the experimental results of anti-SARS-CoV-2 activities estimation task. In Table **S19**, we obtain the results of Jure's GNN by running its public source code (https://github.com/snap-stanford/pretrain-gnns) on our SARS-CoV-2 dataset. In Table **S20**, REDIAL-2020 provided the dataset they used (https://doi.org/10.5281/zenodo.4606720), including training set, validation set and test set. Therefore, we run ImageMol under the same experimental settings as theirs. In Table **S21**, the results of sequence-based models (e.g. RNN_LR, TRFM_LR, RNN_MLP, TRFM_MLP, RNN_RF, TRFM_RF [27] and CHEM-BERT [28]) and graph-based models (e.g. $\text{MolCLR}_{\text{GIN}}$, $\text{MolCLR}_{\text{GCN}}$ [13] and GROVER [16]) are obtained by running their public source code on these SARS-CoV-2 datasets with three different seeds.

C.4 Results on the virtual screening anti-SARS-CoV-2 drugs

Table **S22** shows virtual screening results of approved drugs in DrugBank for 3CL inhibitors. In particular, we performed virtual screening using the 3CL model with 83.7% ROC-AUC in Table **S19**. The predicted label and probability for each drug is in the columns pred_labels, non-inhibitor_probs, inhibitor_probs of Table **S22**. The distribution histogram of predicted 3CL

inhibitors is depicted in Figure **S19**. Additionally, to further validate the effectiveness of our approach, we also screened drugs for SARS-CoV-2 from approved drugs. We use HEK293's model for virtual screening because it models larger data volumes and has good performance. The screening results are shown in Table **S24**. The 15 of the top 20 drugs were verified by different literatures, demonstrating the great potential of ImageMol. We also tested the accuracy of the model on the external validation set (Table **S25**), which is provided by and has 122 inhibitors of the SARS-CoV-2 (shown in https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-022-04482-x/MediaObjects/41586_2022_4482_MOESM1_ESM.pdf). Since we focus on virtual screening of small molecule drugs, we took the intersection of these 122 inhibitors and drugs in DrugBank and finally got 70 small molecules for testing. Of these 70 drugs, we successfully predicted 42 potential drugs, demonstrating the potential of ImageMol as a novel drug discovery tool.

C.5 Supplementary Discussion on Ablation Studies

Impact of pre-training: The robustness of the model to hyperparameter is important because the initialization of different parameters can affect the performance of the model[45]. Here, we explore the impact of pre-training strategies on the hyperparameter tuning of ImageMol. As shown in Figure **S25**, ImageMol is more robust than ImageMol_NonPretrained, with an average performance standard deviation of 0.5% versus 8.9% on

classification task and 0.654 versus 1.68 on regression task. Therefore, pre-training strategies improve the robustness of ImageMol to initialization parameters. In addition, the difference in performance between pre-training and no pre-training (ROC-AUC improvement ranging from 9.0% to 32.4% with an average improvement of 20.2% on classification task and RMSE improvement ranging from 0.482 to 1.472 with an average improvement of 0.879 on regression task) also indicated that the pre-training process significantly improved the model performance.

Impact of pre-training data scale: To explore the impact of pre-training with different data scales, we first use 0 million (no pre-training), 0.2 million, 0.6 million, 1 million, and 8 million drug-like compounds to pretrain ImageMol respectively and then evaluate their performance. We found that the average ROC-AUC performance increased from 1.2% to 10.2% as the pre-trained data size increases (Figure **S26**). Thus, ImageMol can be further improved as the more drug-like molecules can be pre-trained.

Impact of different pretext tasks: We investigated the impact of different pretext tasks using multi-granularity chemical clusters classification (MG3C), jigsaw puzzle prediction (JPP), and MASK-based contrastive learning (MCL) (*cf.* Methods), respectively. We found that each pretext task improves the mean AUC value of ImageMol from 0.7% to 4.9%: without pre-text task (75.7%), JPP (78.8%), MG3C (80.6%) and MCL (76.4%) (Figure **S27**). The best performance was achieved by assembling all 3 pretext tasks

for pre-training (AUC = 85.9%, Figure **S27**). Thus, pre-training tasks of ImageMol are well compatible and jointly improve model performance.

Impact of data augmentation: We applied three data augmentation strategies in the pre-training and fine-tuning of ImageMol, including RandomHorizontalFlip, RandomGrayscale and RandomRotation. Table **S29** illustrates several examples of data augmentation visualizations. We observed that the data augmentation did not change the original structure of the molecules. Meanwhile, the similarity of the embedding vectors exceeds 99%, indicating that ImageMol captures the invariance of data augmentation to improve the generalization of the model. We further conducted an ablation study. Table **S28** shows that data augmentation can synergistically improve the performance of ImageMol. In detail, each data augmentation strategy improves ImageMol performance compared to without any data augmentation strategy. At the same time, the performance after using multiple data augmentation also exceeds the performance using a single data augmentation strategy, which shows the effectiveness of multiple data augmentation strategies. Altogether, the performance of ImageMol can be improved using data augmentation.

Supplementary Figures

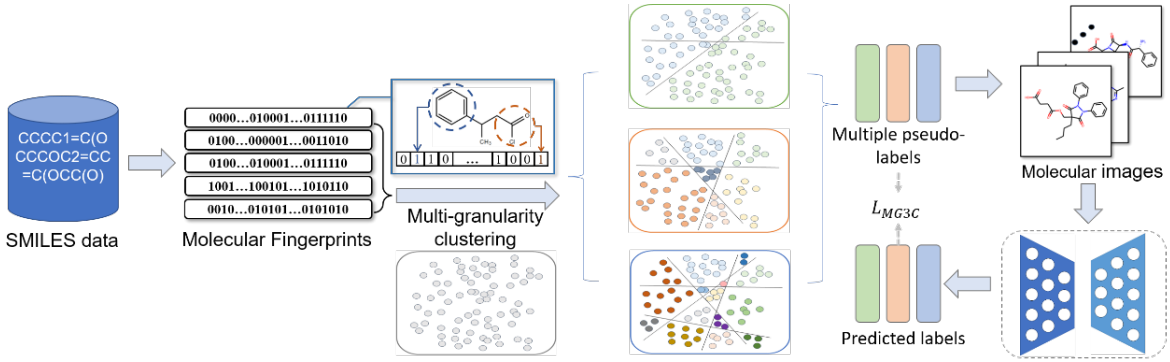


Figure S1: The architectural details of the Multi-Granularity Chemical Clusters Classification (MG3C) task. Firstly, the molecular fingerprints are extracted from SMILES and input into unsupervised multi-granularity clustering to produce clusters with different granularity. Then, these clusters are uniquely numbered as pseudo-labels of molecular images. Finally, the molecular encoder and structural classifier are jointly used to predict the labels of molecular images and optimizing the loss between pseudo-labels and predicted labels in pre-training.

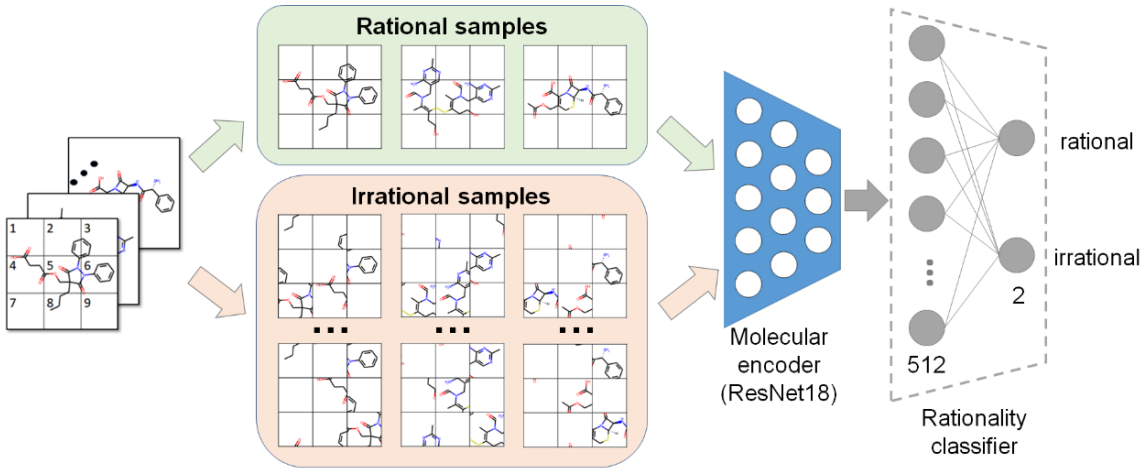


Figure S2: The architectural details of the Molecular Rationality Discrimination (MRD) task. In order to construct an irrational molecular image, we first disrupt the molecular structure, which uses a 3x3 grid to decompose the molecular image into 9 patches and randomly shuffle them to form an irrational image. Then, these rational and irrational molecules will be input to the molecular encoder to extract visual features. Finally, these features are forward propagated to rationality classifier for rationality judgment.

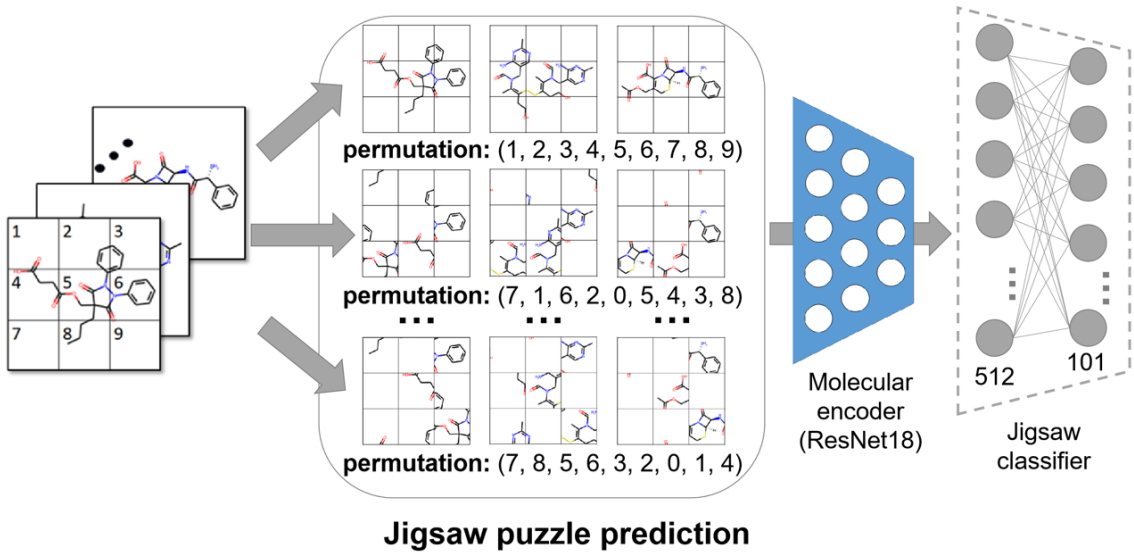


Figure S3: The architectural details of the Jigsaw Puzzle Prediction (JPP) task. We first use a 3x3 grid to decompose the molecular image into 9 patches and assign numbers from 1 to 9. Then, we use different permutations to reorganize the image. Finally, the reorganized images are fed into the molecular encoder and jigsaw classifier to predict the corresponding permutations.

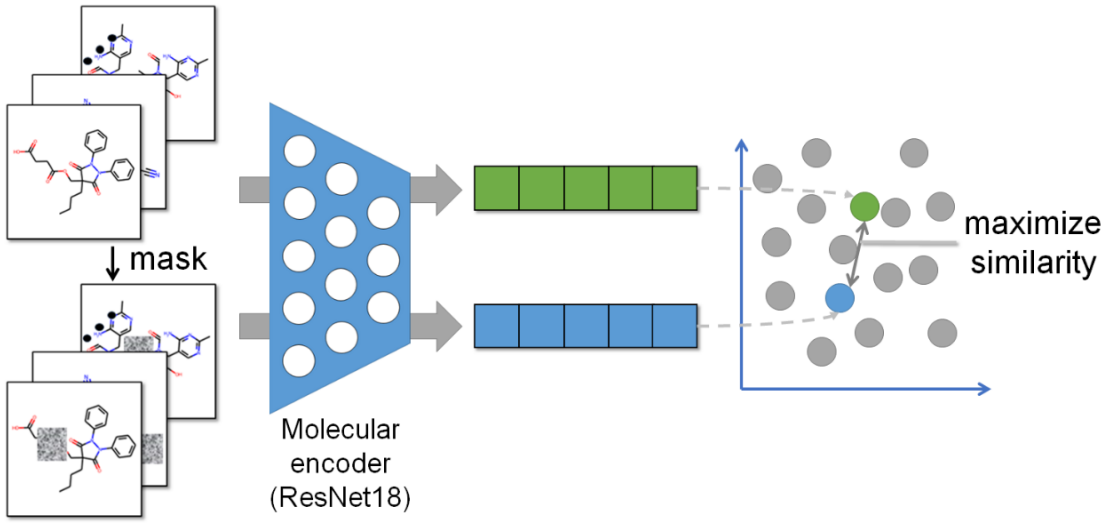


Figure S4: The architectural details of the MASK-based Contrastive Learning (MCL) task. We first randomly mask a 16×16 area to obtain a masked image. Then a pair of images (original image and the masked image) are simultaneously fed into the molecular encoder to extract latent features. Finally, we optimize the molecular encoder by maximizing the similarity among the latent feature pairs.

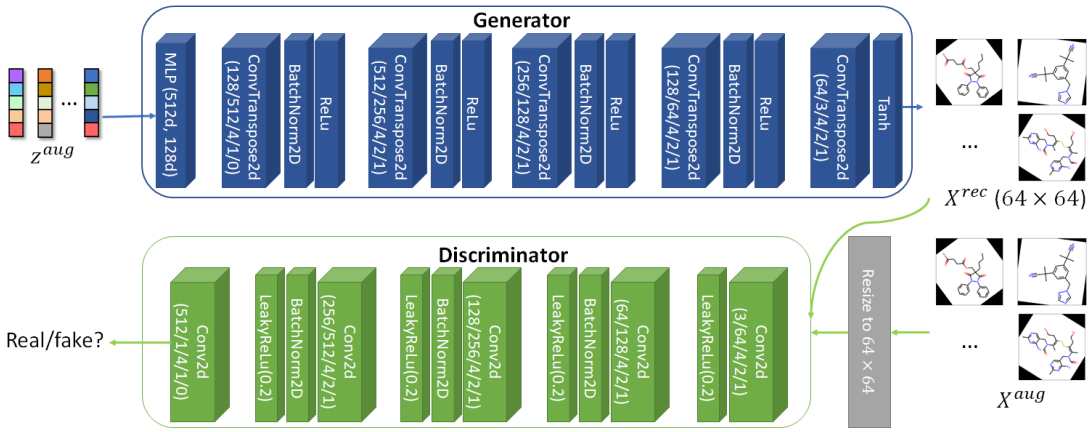


Figure S5: The architectural details of the Molecular Image Reconstruction (MIR) task. The generator is used to reconstruct latent features z^{aug} back into 64×64 molecular images X^{rec}. The discriminator accepts the generated image X^{rec} and the real image X^{aug} and discriminates their real and fake.

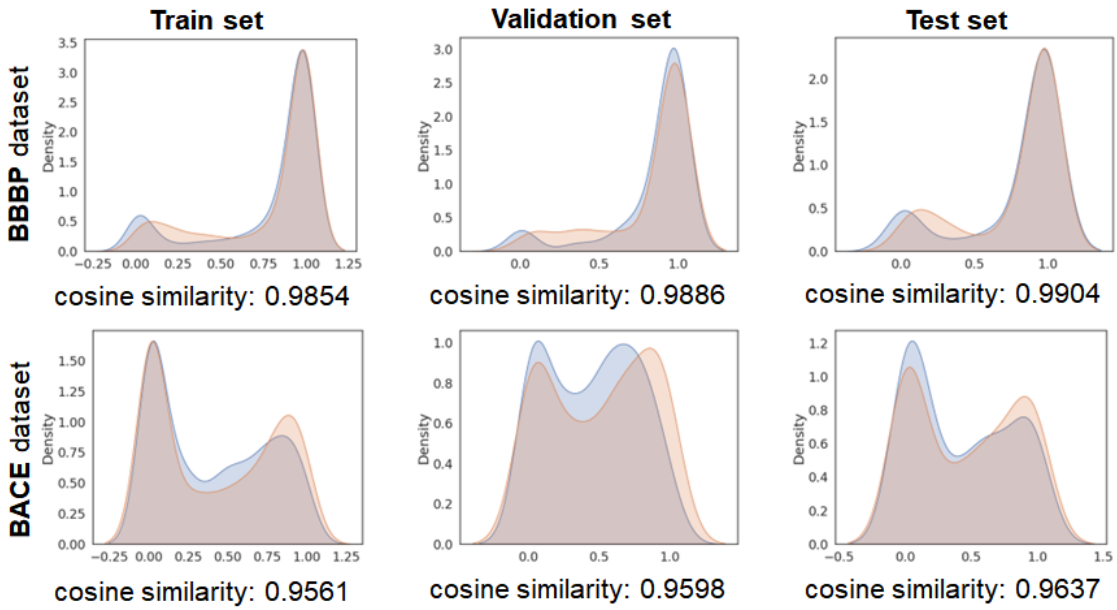


Figure S6: Probability distributions of models in different random seeds on train, validation, and test sets with random scaffold split. Different colors represent probability distributions obtained by models with different random seeds. The first and second rows represent the BBBP and BACE datasets, respectively. The first to third columns represent the probability distributions of the training set, validation set, and test set, respectively.

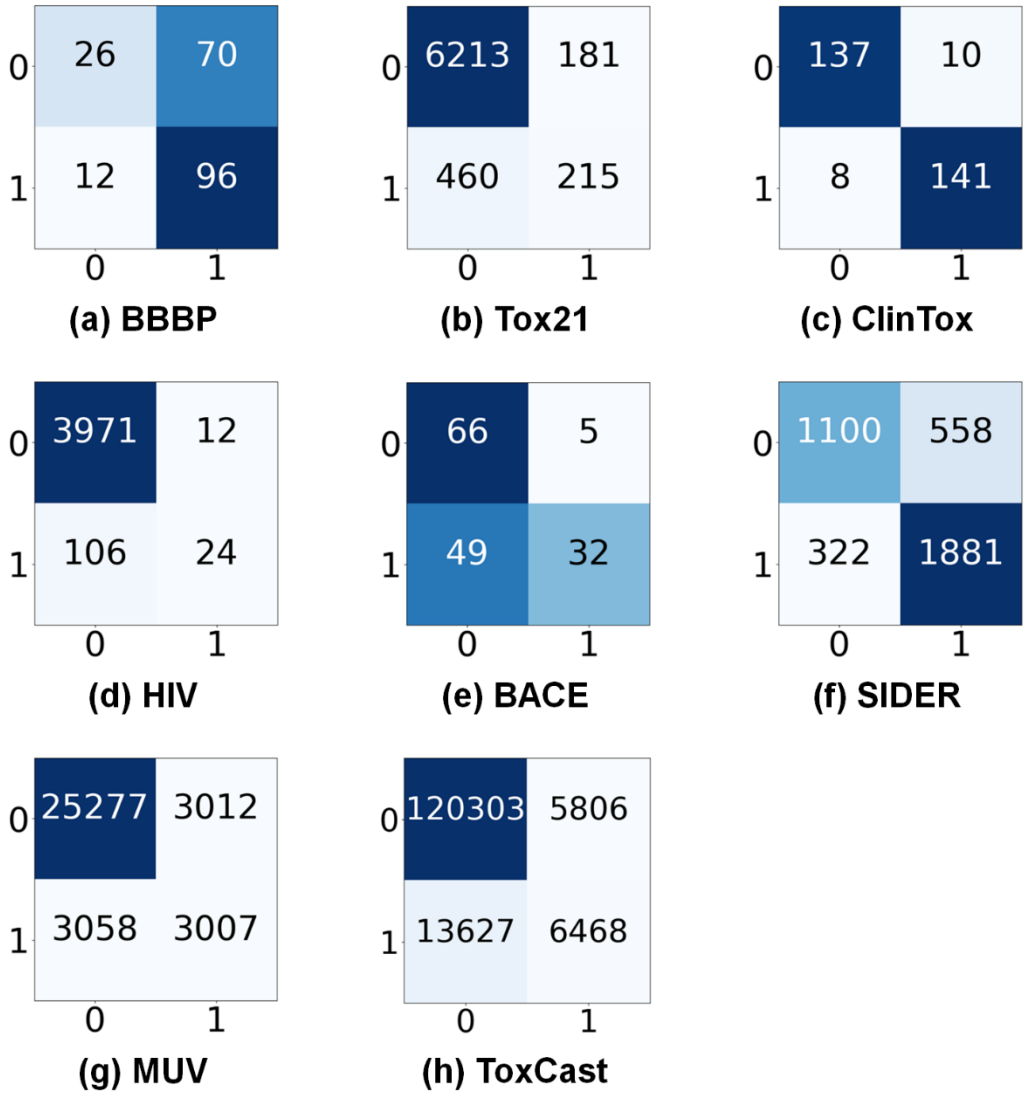


Figure S7: The confusion matrix on 8 molecular property prediction datasets with scaffold split. Results are obtained from the best model in 3 runs.

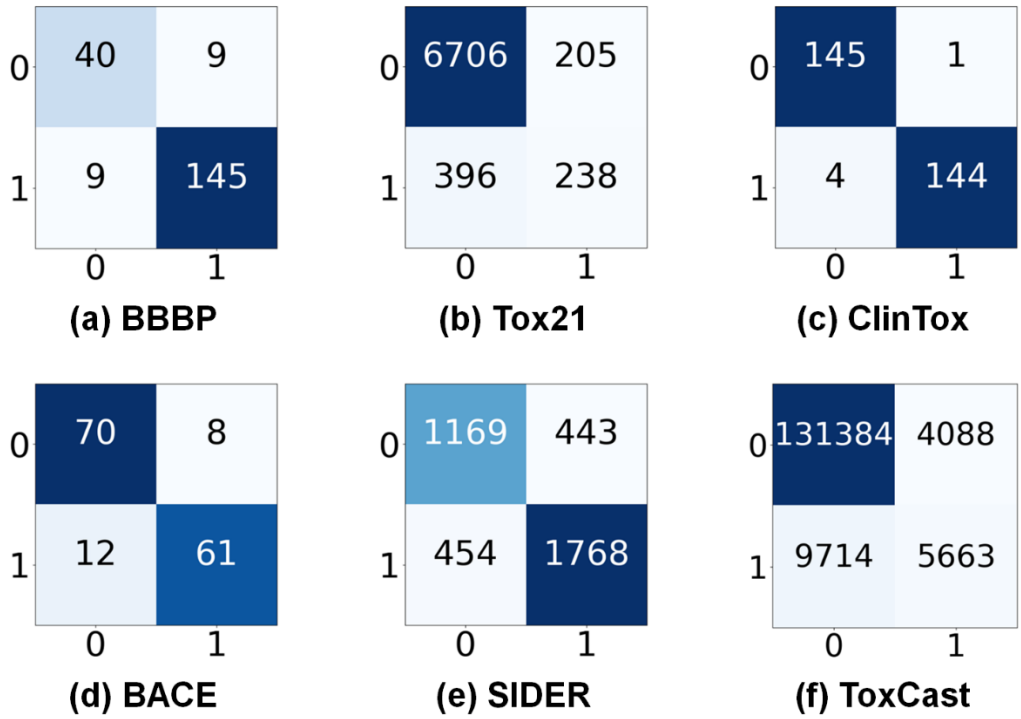
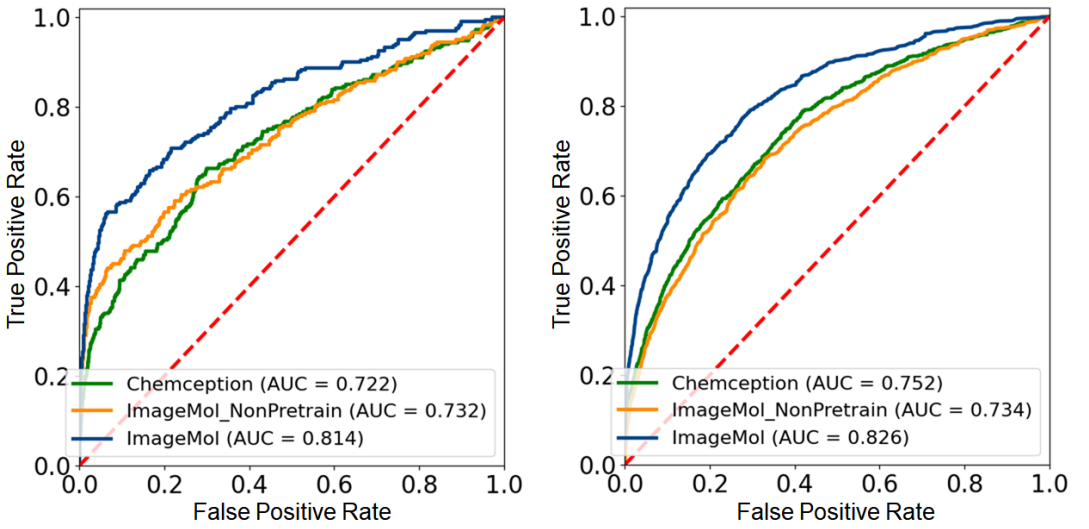


Figure S8: The confusion matrix on 6 molecular property prediction datasets with random scaffold split. Results are obtained from the best model in 3 runs.



(a) The ROC-AUC curve on HIV dataset (b) The ROC-AUC curve on Tox21 dataset

Figure S9: Receiver operating characteristic (ROC) curves of Chemception, ImageMol_NonPretrained and ImageMol on Tox21 and HIV datasets. Chemception is the method based on molecular image to predict the molecular property. ImageMol_NonPretrained is the ResNet18 trained from scratch without any pre-training. ImageMol is our pre-trained model based on 10 million molecular images. The standard deviations of Chemception, ImageMol_NonPretrain and ImageMol are 0.012, 0.016 and 0.003 on HIV dataset and 0.007, 0.011 and 0.001 on Tox21 dataset by running 3 times with different random seeds.

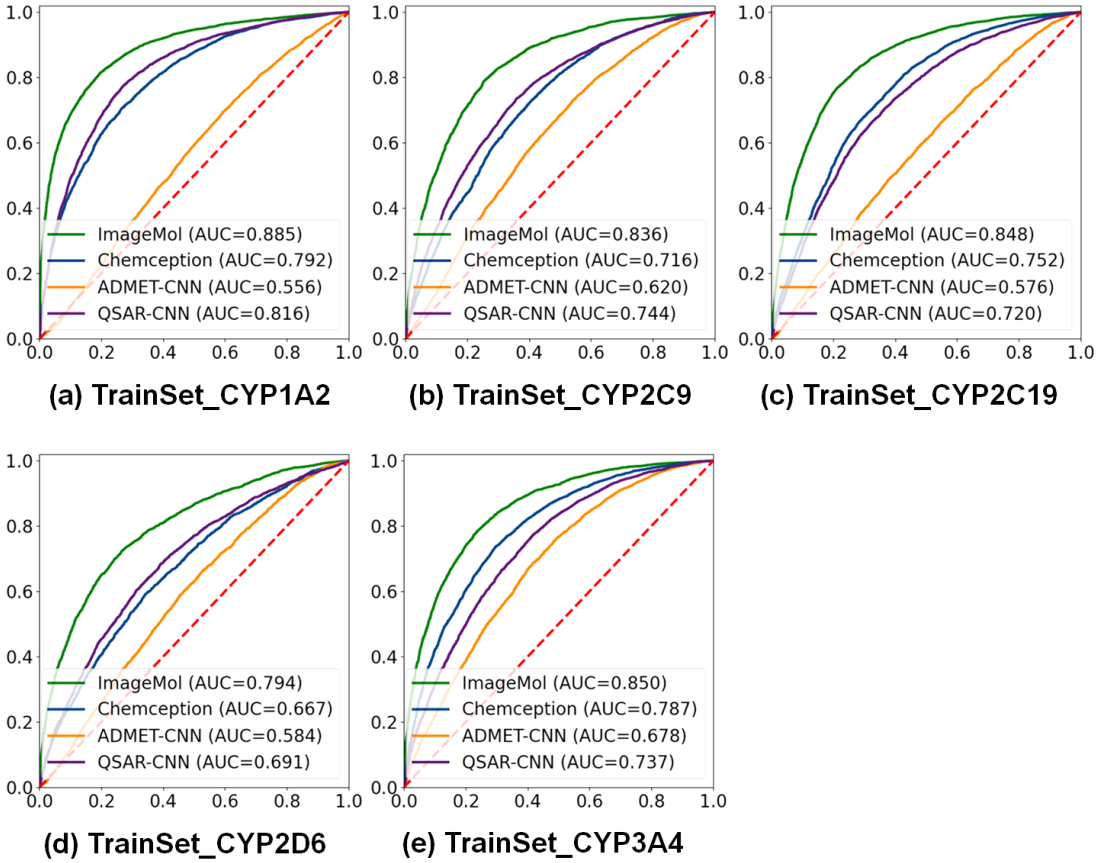


Figure S10: Receiver operating characteristic (ROC) curves of ADMET-CNN[25], QSAR-CNN[26] and ImageMol on five CYP450 isoforms training sets (PubChem Data Set I) with 5-fold cross-validation.

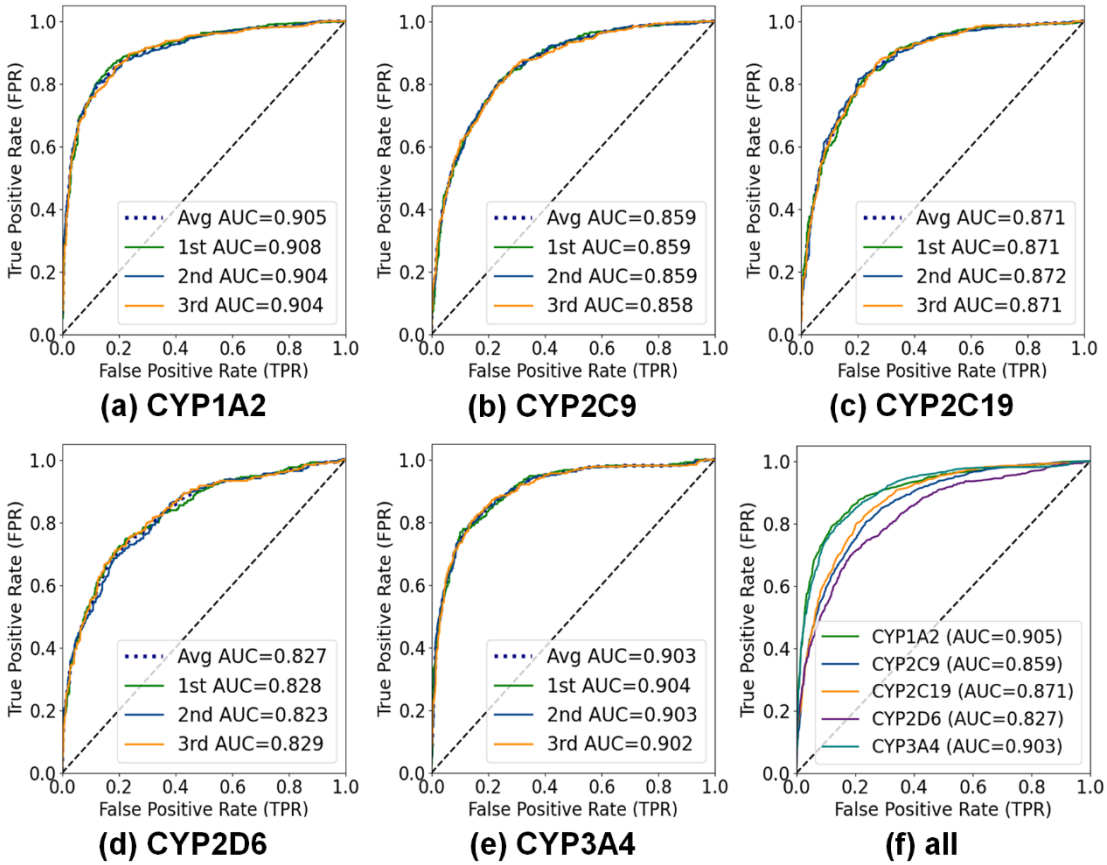


Figure S11: The ROC-AUC curve on 5 CYP450 datasets with balanced scaffold split. 1st AUC, 2nd AUC and 3rd AUC represent the results of the first, second and third random runs, respectively. Avg AUC means macro-averaged AUC on three random runs. (f) represents the average AUC curve of 5 CYP450 datasets.

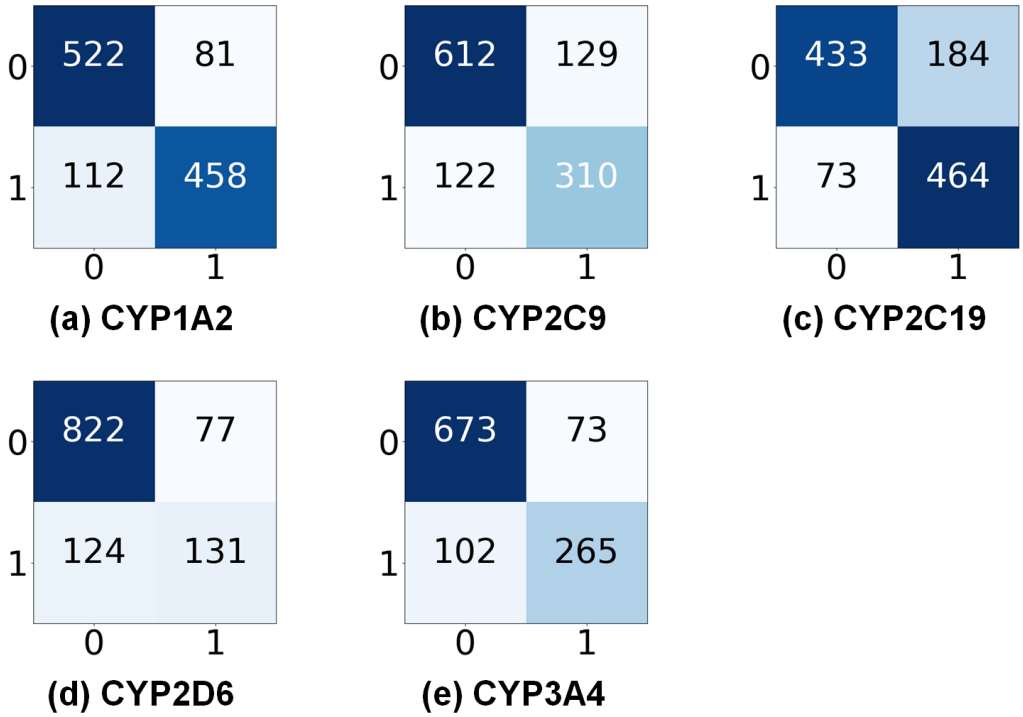


Figure S12: The confusion matrix on 5 CYP450 datasets with balanced scaffold split. Results are obtained from the best model in 3 runs.

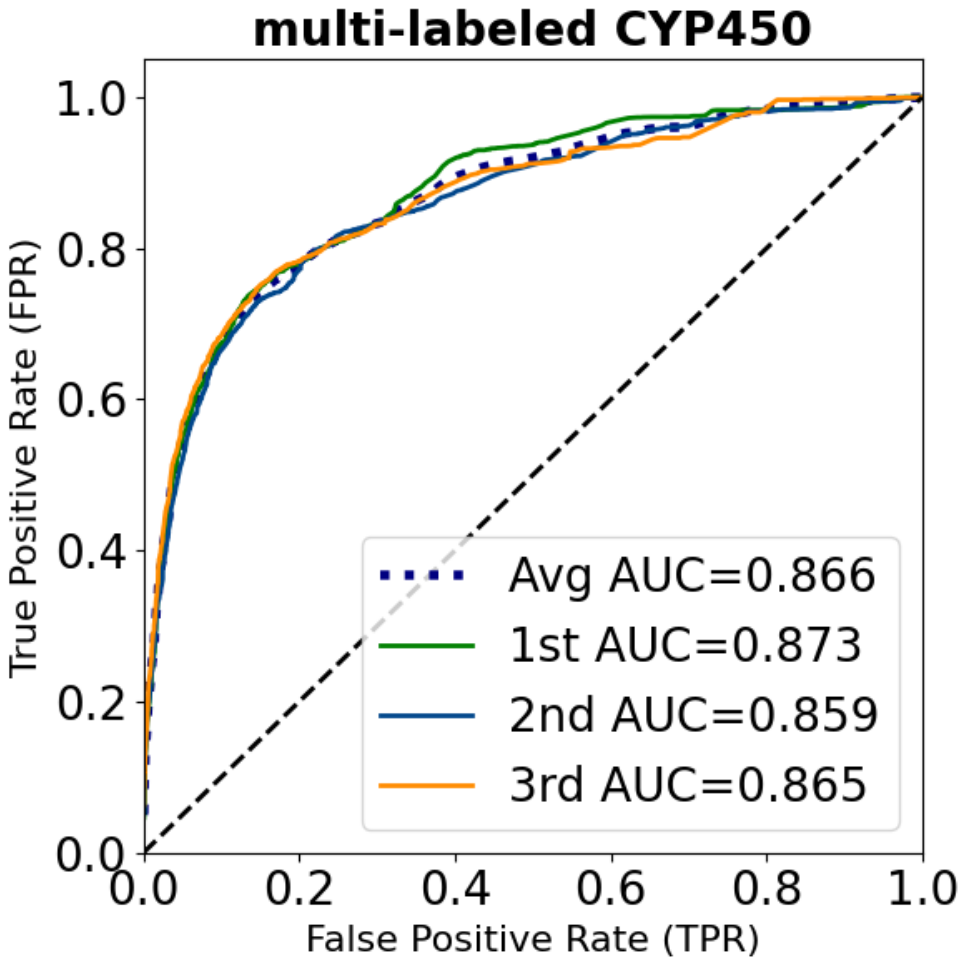


Figure S13: The ROC-AUC curve on the multi-labeled CYP450 dataset. 1st AUC, 2nd AUC and 3rd AUC represent the results of the first, second and third random runs, respectively. Avg AUC means macro-averaged AUC on three random runs.

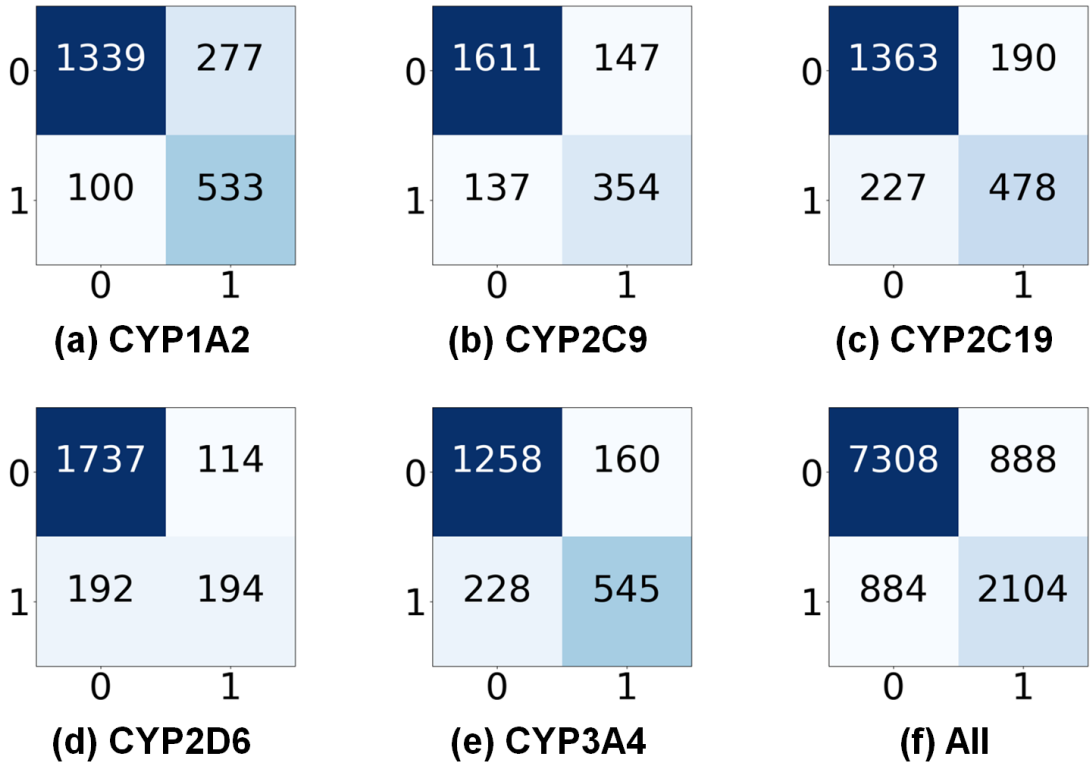


Figure S14: The confusion matrix on multi-labeled CYP450 datasets with balanced scaffold split. Results are obtained from the best model in 3 runs.

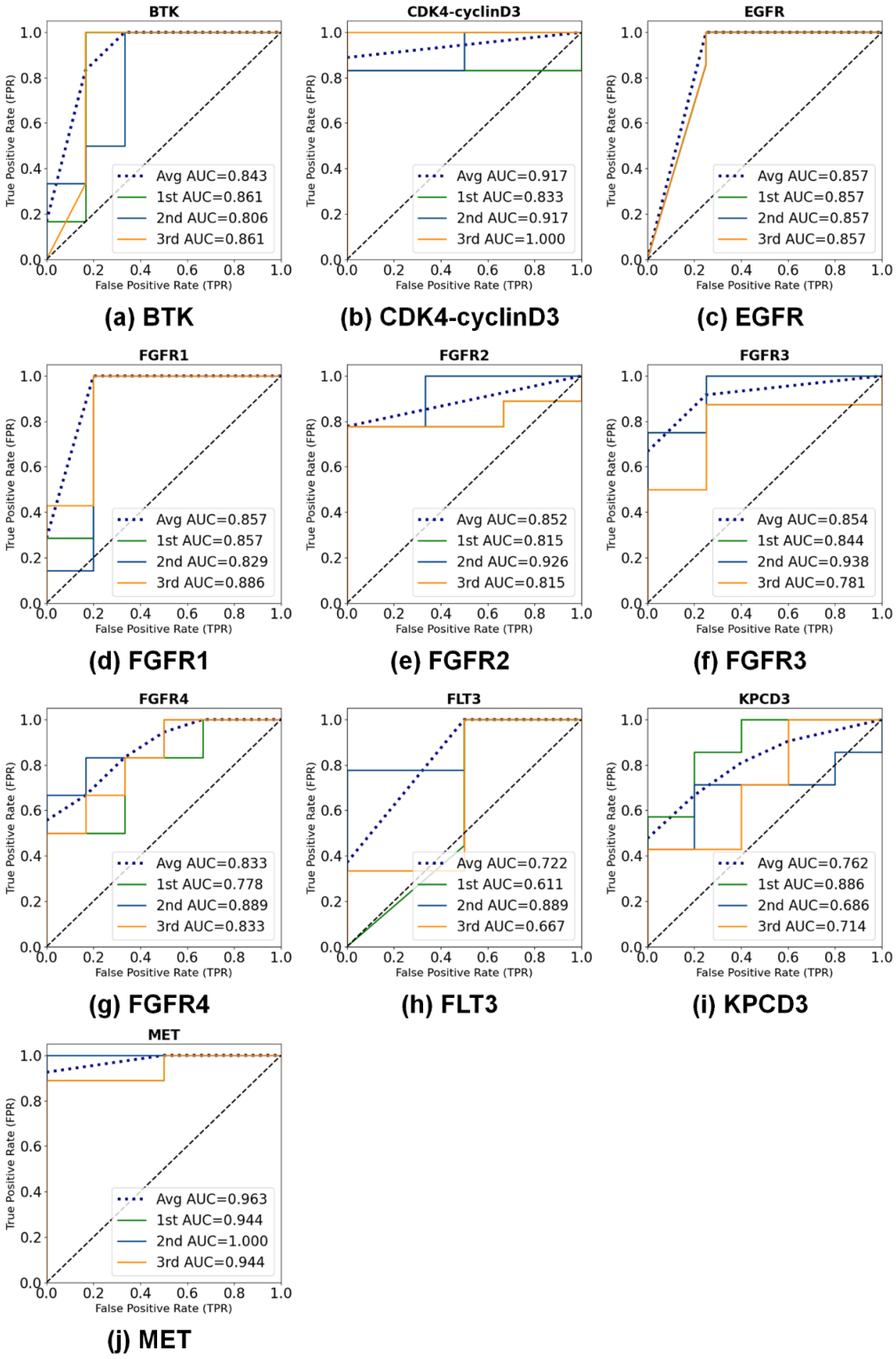


Figure S15: The ROC-AUC curve on the drug-protein binding activity datasets from KinomeScan. 1st AUC, 2nd AUC and 3rd AUC represent the results of the first, second and third random runs, respectively. Avg AUC means macro-averaged AUC on three random runs.

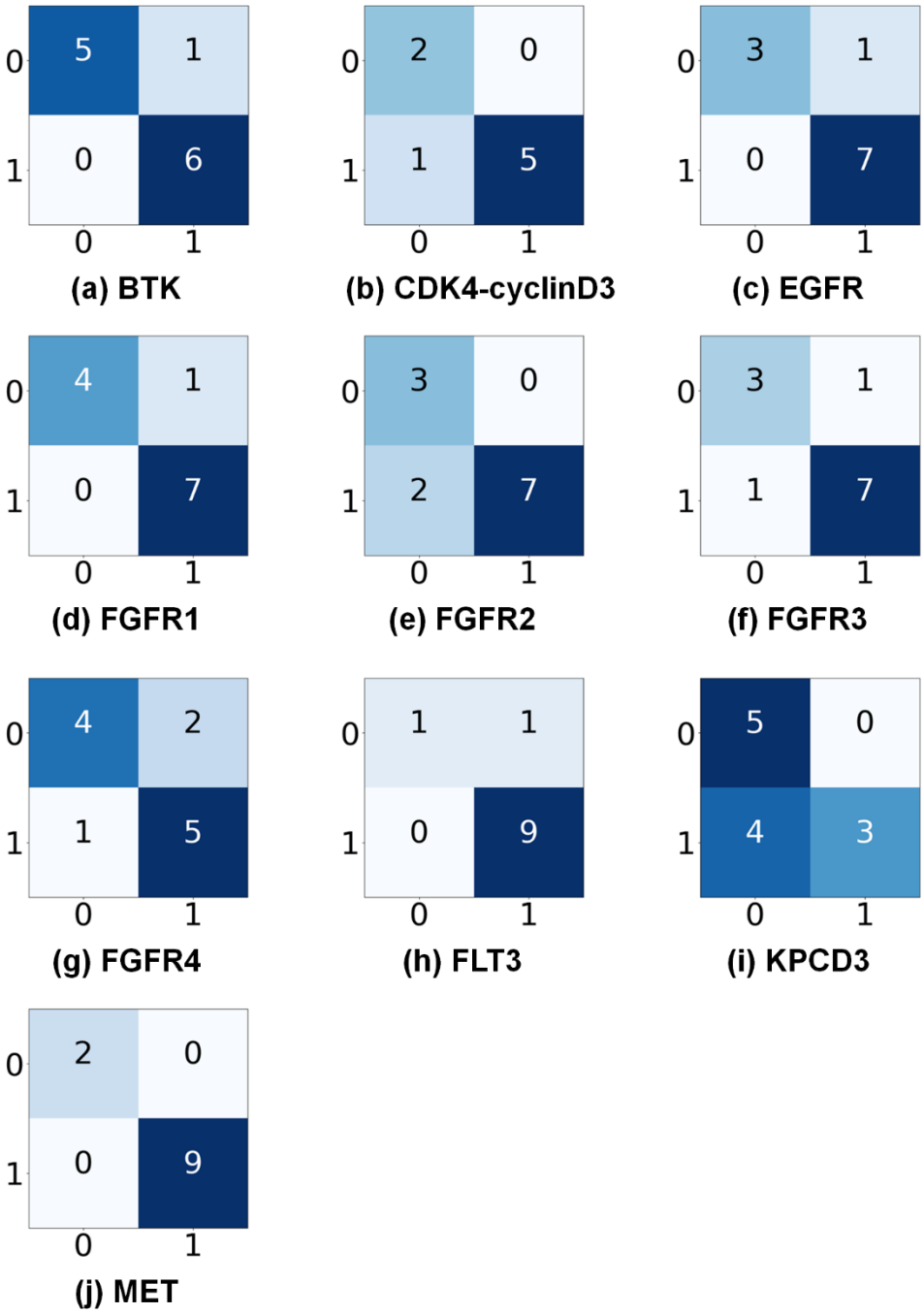


Figure S16: The confusion matrix on drug-protein binding activity datasets from KinomeScan with balanced scaffold split. Results are obtained from the best model in 3 runs.

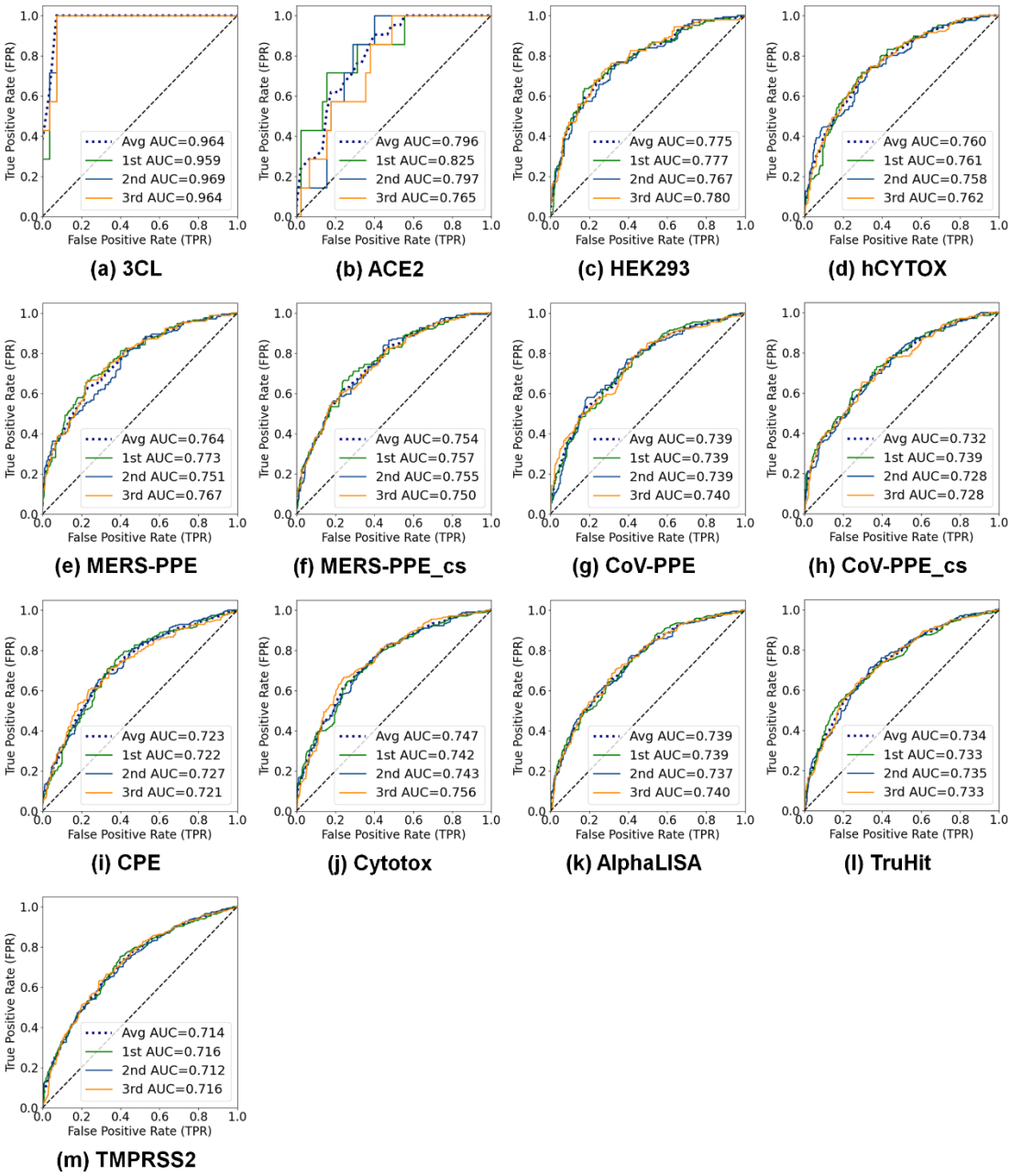


Figure S17: The ROC-AUC curves on the 13 SARS-CoV-2 datasets with balanced split. 1st AUC, 2nd AUC and 3rd AUC represent the results of the first, second and third random runs, respectively. Avg AUC means macro-averaged AUC on three random runs. The title is an abbreviation for dataset.

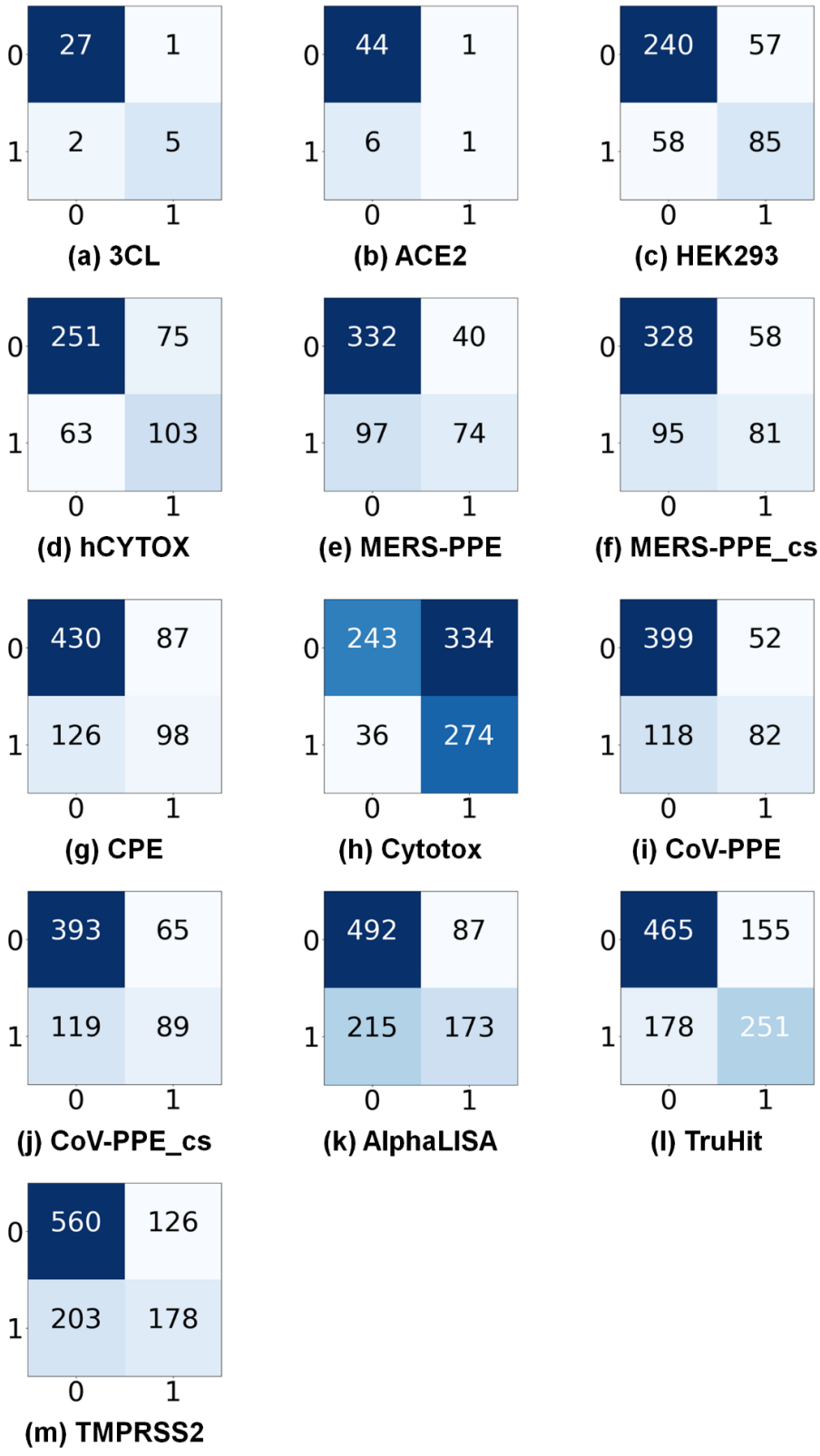


Figure S18: The confusion matrix on 13 SARS-CoV-2 datasets with balanced scaffold split. Results are obtained from the best model in 3 runs.

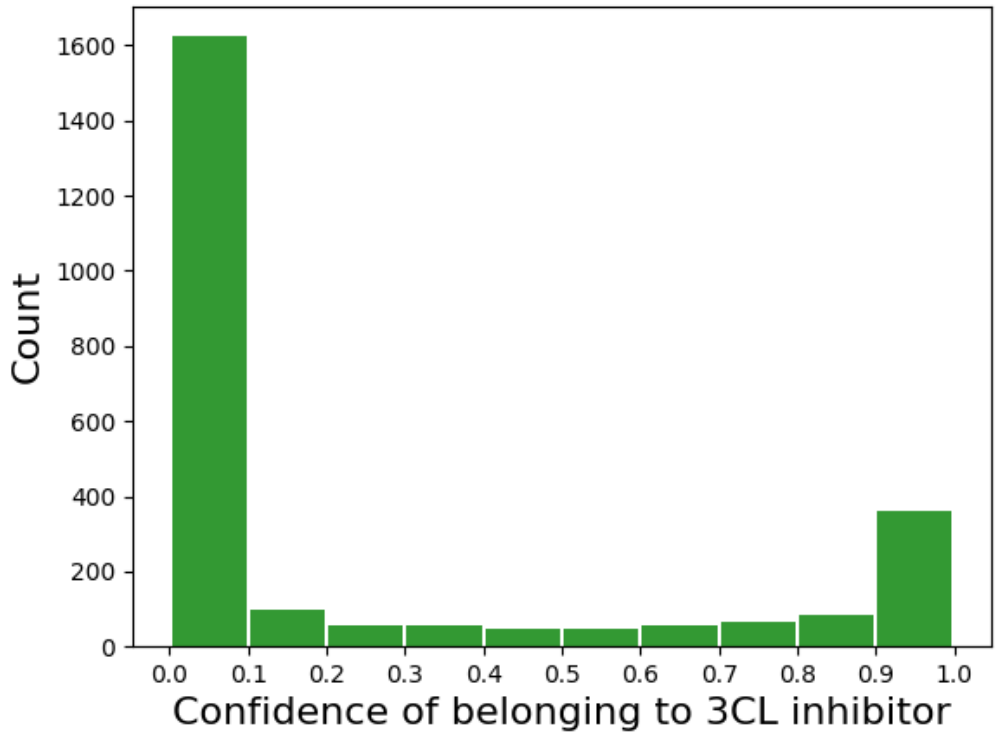
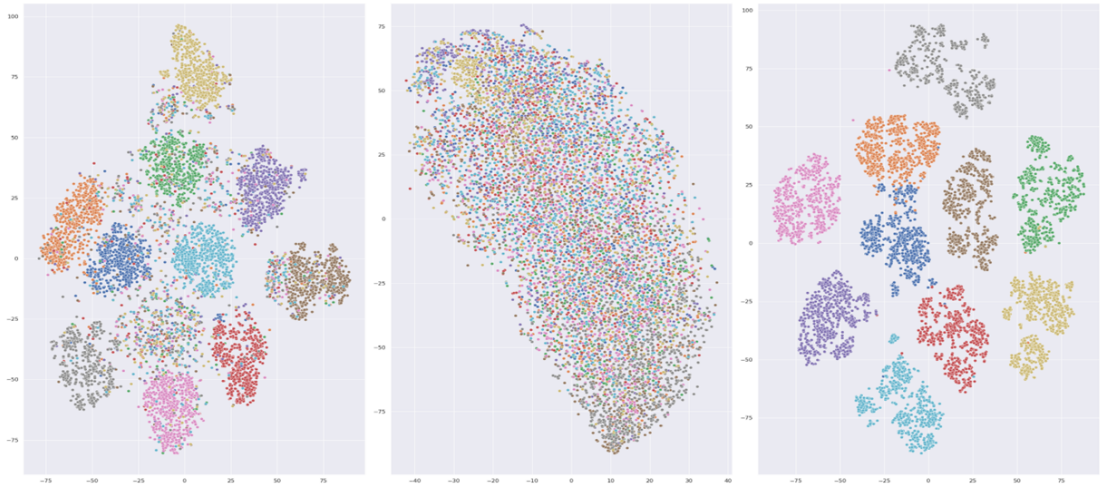


Figure S19: The distribution histogram of drugs that can be used as 3CL inhibitors. The x-axis represents the confidence that the drug is a 3CL inhibitor, and the y-axis represents the number of drugs within a certain confidence interval. The range of confidence is 0 to 1. The higher the confidence, the more likely the corresponding drug is to be a 3CL inhibitor.



(a) MACCS molecular fingerprints
DB index: 2.93

(b) ImageMol_NonPretrained
DB index: 19.40

(c) ImageMol
DB index: 1.92

Figure S20: t-SNE visualization of the representations learned by different models. Different colors indicate different clusters. Davies Bouldin (DB) index is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. The lower the DB index value, the better the clustering result.

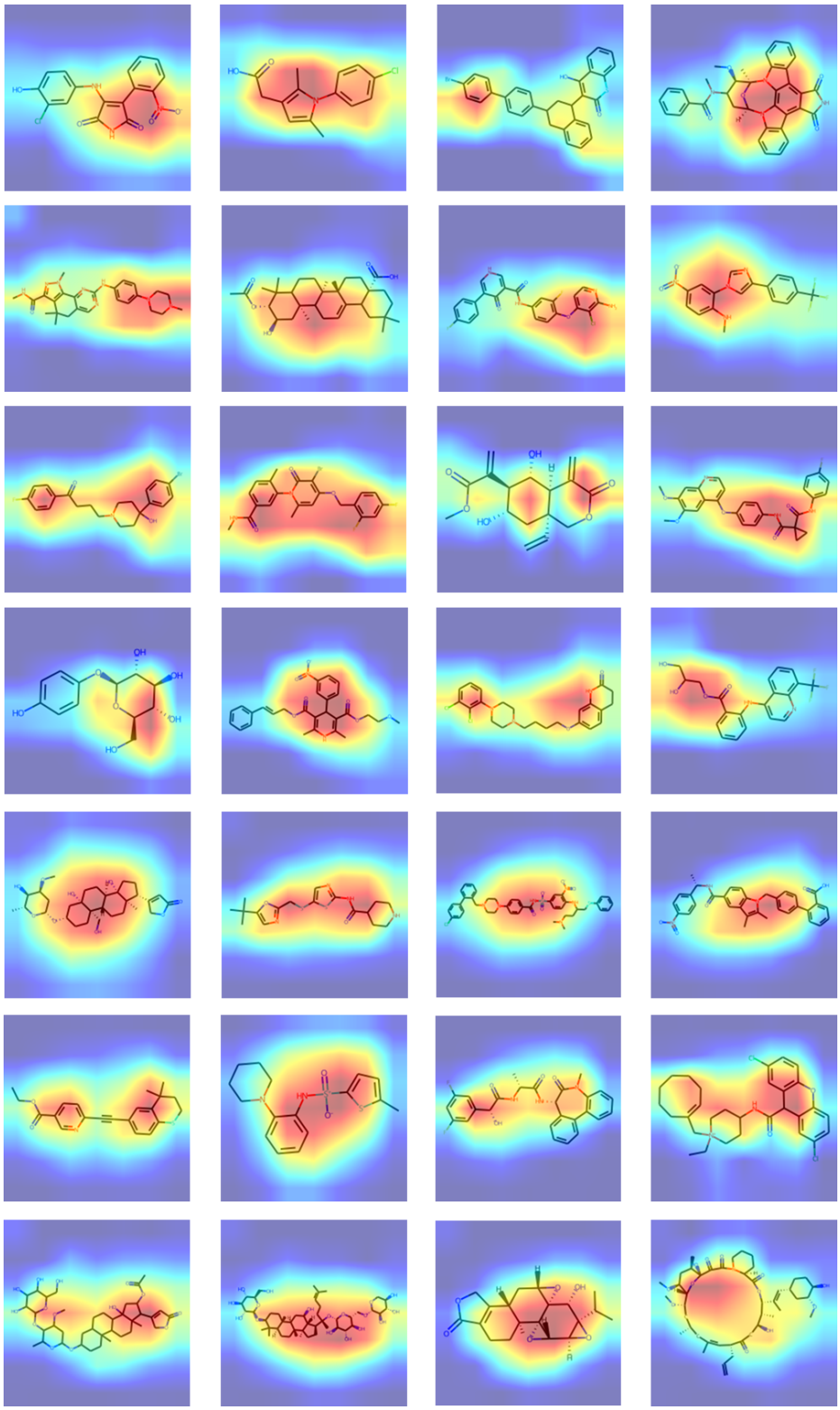


Figure S21: The heat maps of attention to global structural information, which are highlighted by Gradient-weighted Class Activation Mapping (Grad-CAM) [46]. The warmer color, the higher attention; the cooler color, the lower attention. Obviously, all meaningful structural regions are highlighted by ImageMol.

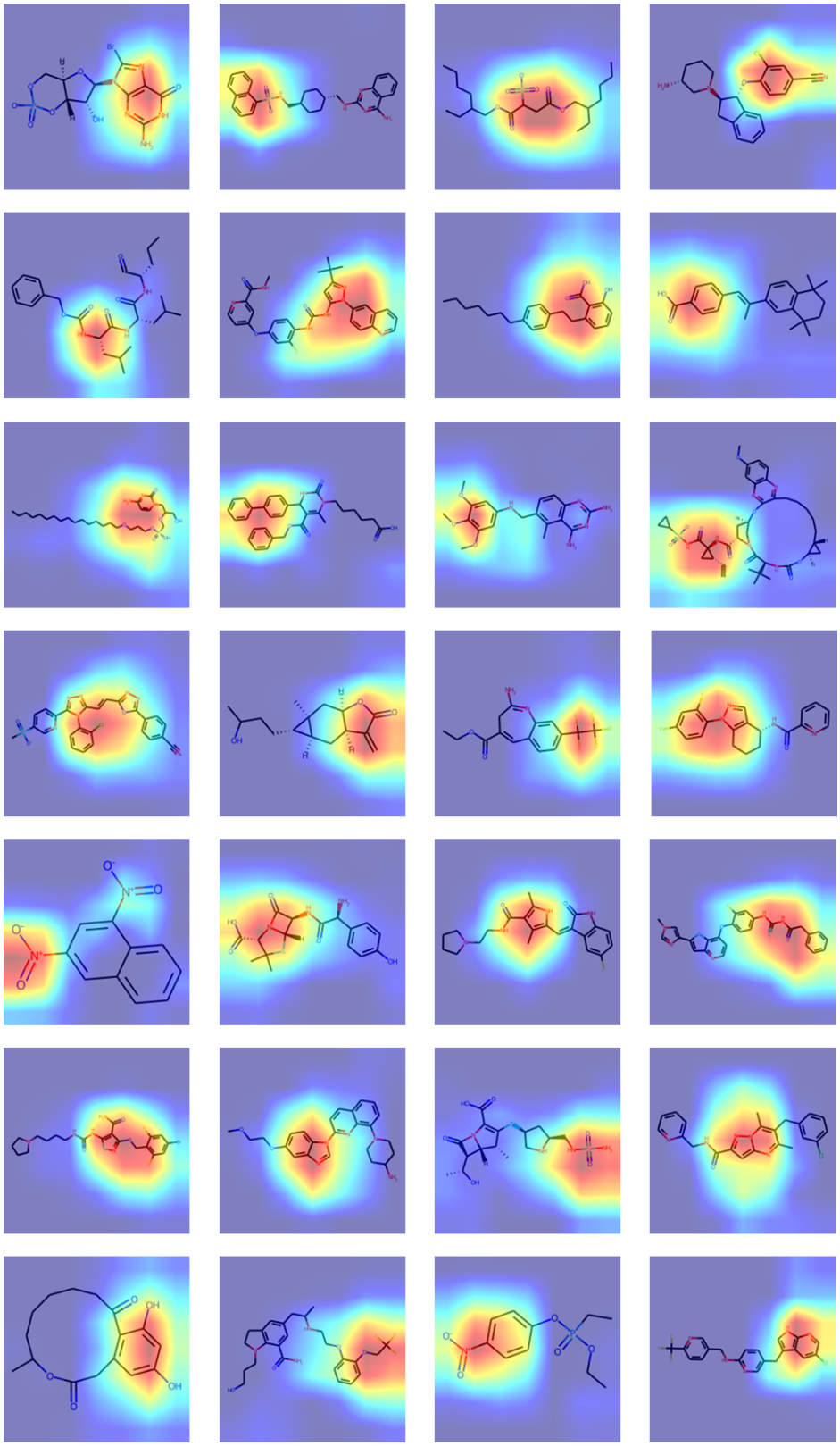


Figure S22: The heat maps of attention to local structural information, which are highlighted by Gradient-weighted Class Activation Mapping (Grad-CAM) [46]. The warmer color, the higher attention, and the cooler color, the lower attention. Obviously, ImageMol captures more local regions for inference.

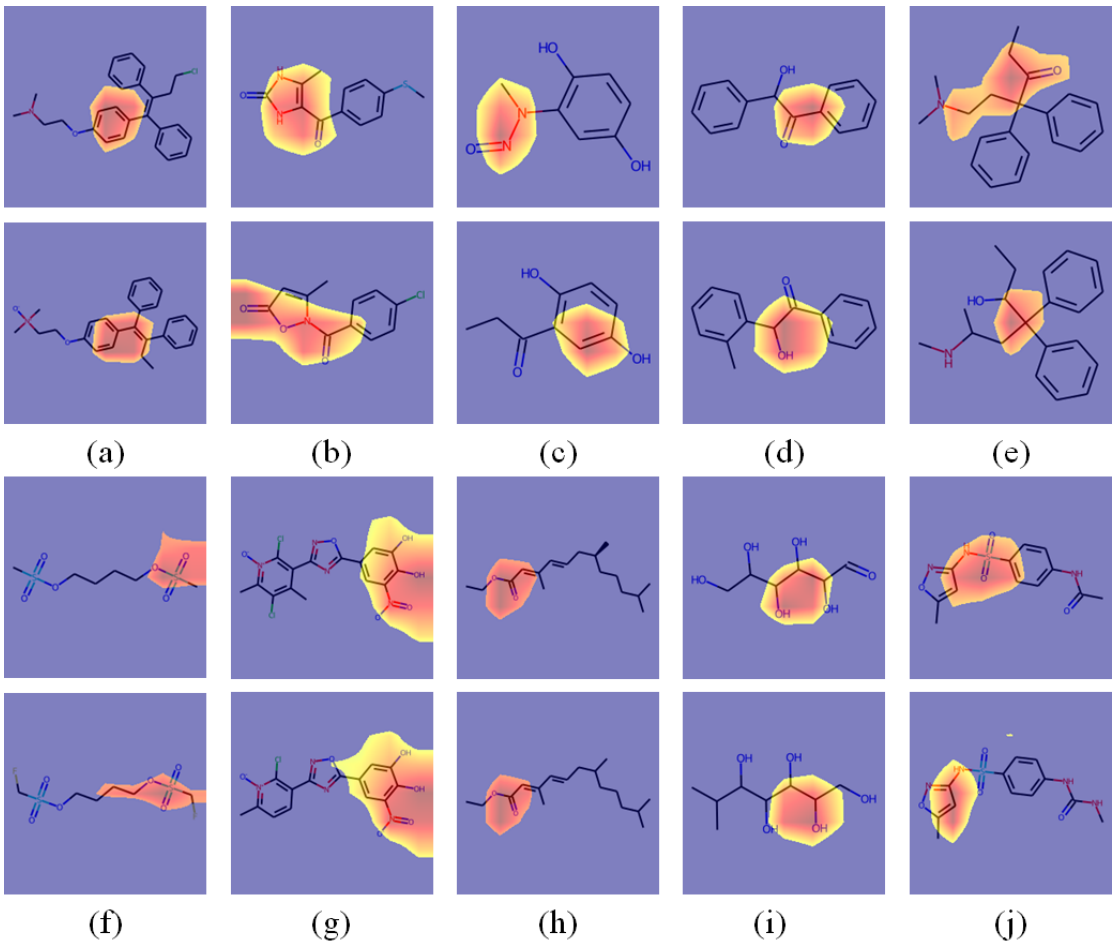


Figure S23: Examples of similar molecular structures but different biological activities. Take 3CL protease inhibitors by utilizing the ImageMol (3CL) model as an example. The first and second rows in subfigures represent 3CL inhibitors and 3CL non-inhibitors, respectively. This may help discover key structures that inhibit the 3CL protease.

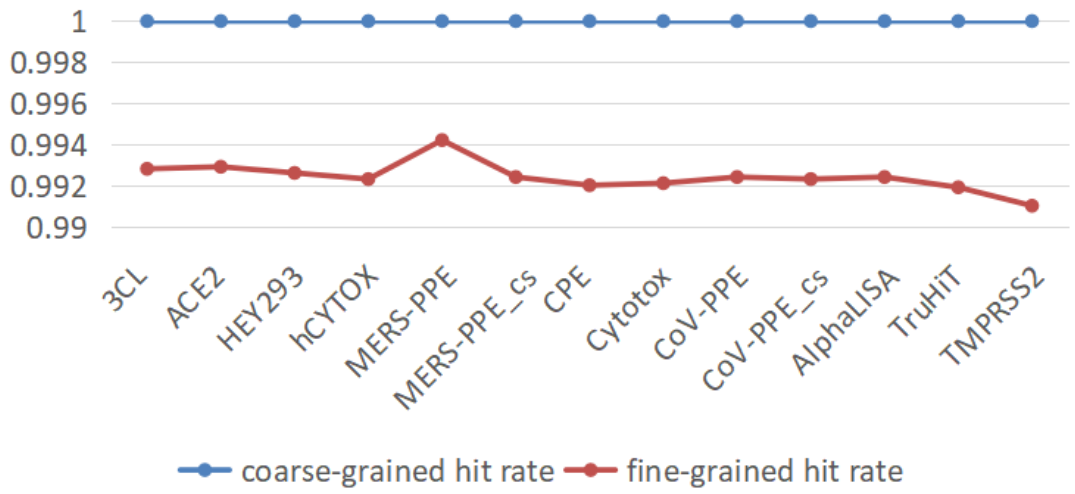


Figure S24: Quantitative information about the molecular structure area that is focused on by ImageMol. The coarse-grained hit rate represents the proportion of the molecular structure in each molecular image that was noticed by ImageMol, and fine-grained hit rate represents the proportion of the molecular structure area that was noticed by ImageMol to the total molecular structure area.

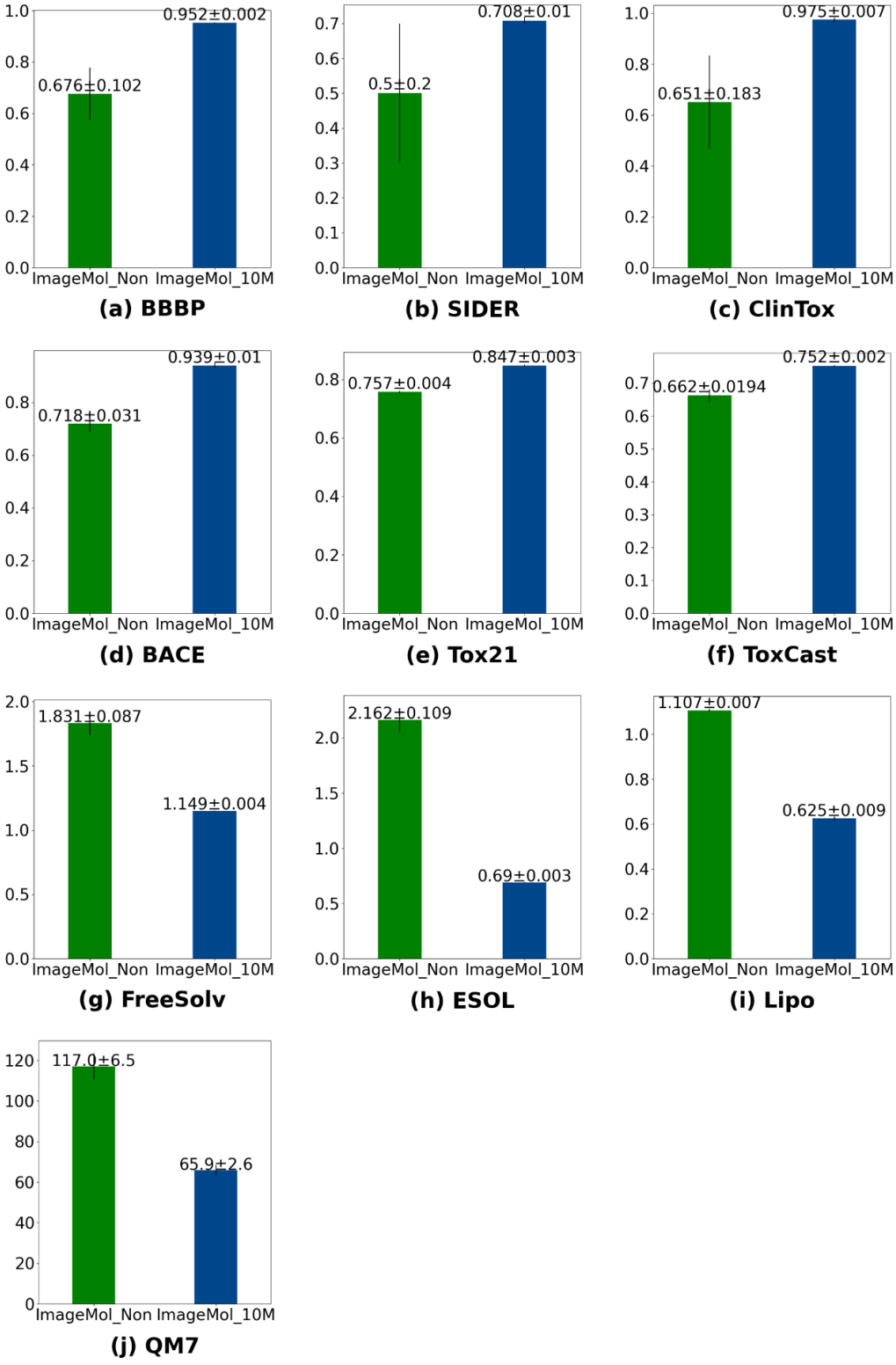


Figure S25: Ablation experiments with or without pretraining on molecular property prediction tasks. ImageMol_10M means pre-training on 10 million datasets, ImageMol_Non means no pre-training. The metrics of (a)-(f), (g)-(i) and (j) are ROC-AUC, RMSE and MAE, respectively.

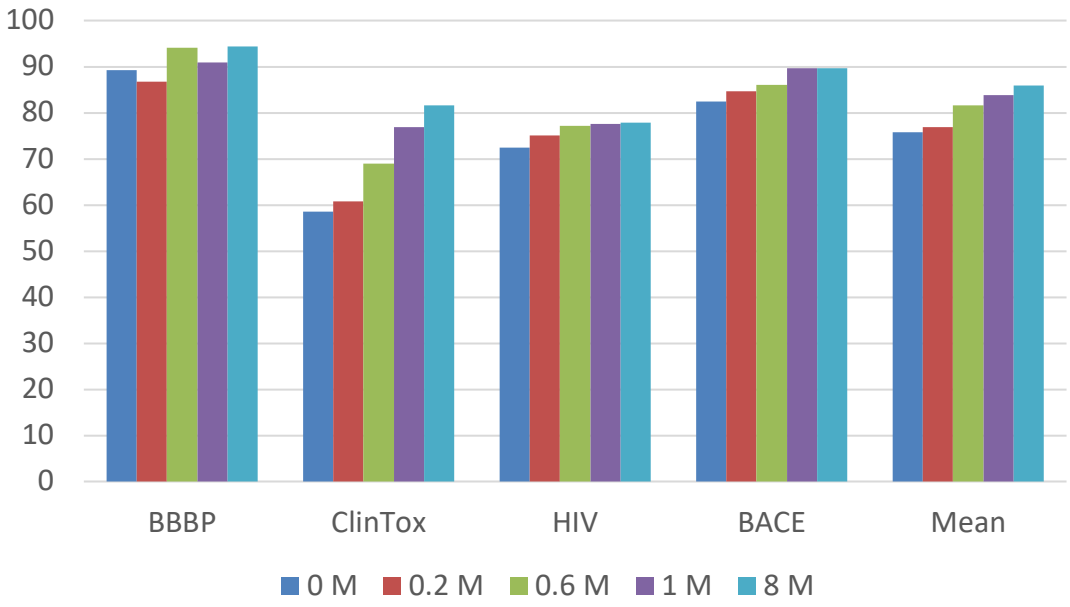


Figure S26: The performance of ImageMol pre-trained under the data scale of 0M (no pre-training), 0.2M, 0.6M, 1M and 8M with random scaffold split on four property prediction datasets. “Mean” represents the average performance of ImageMol pre-trained with different data scales on four datasets.

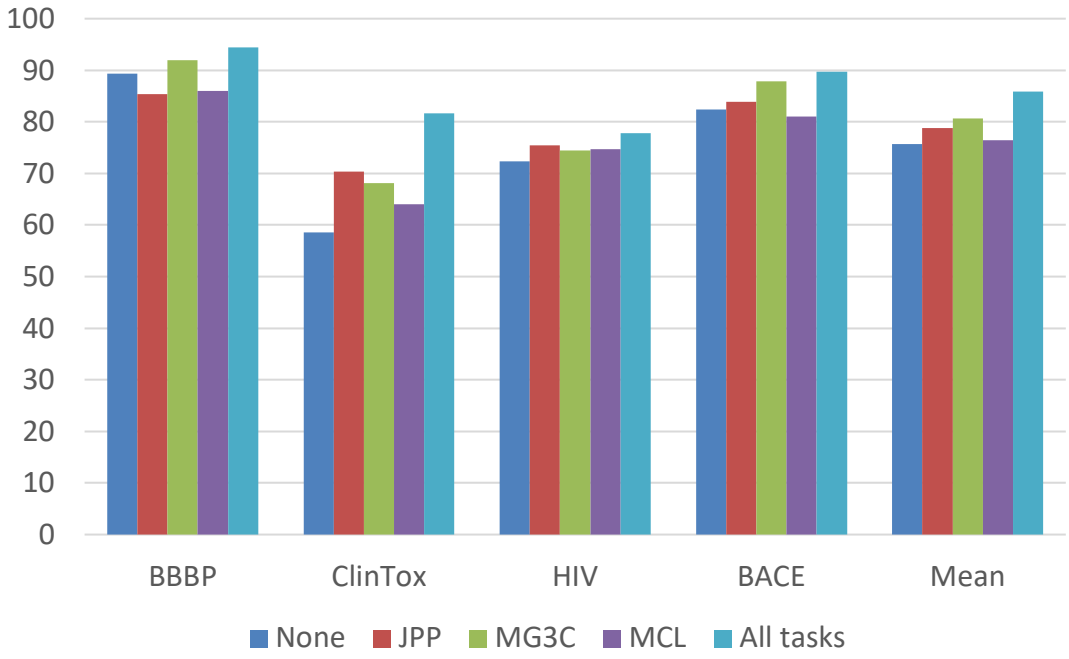
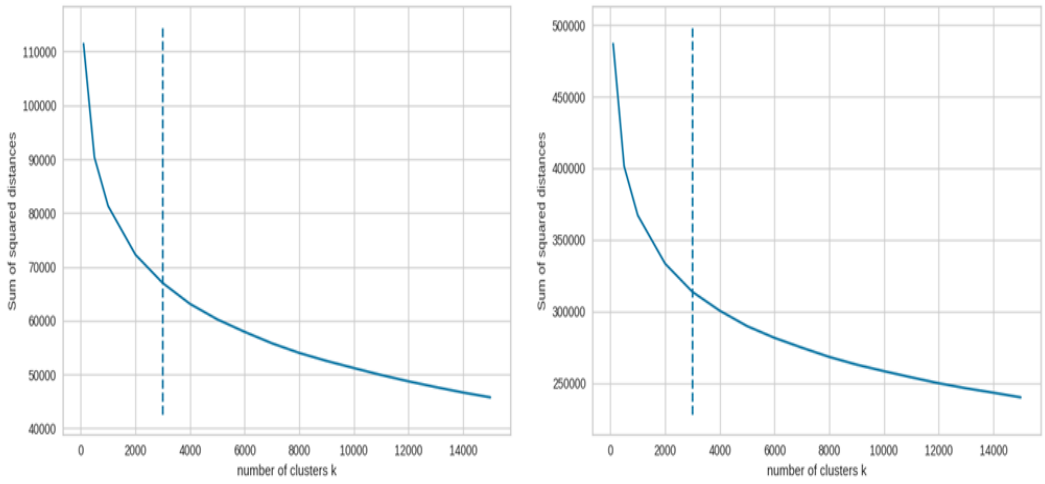


Figure S27: The ablation study of the pretext task in ImageMol, which uses ROC-AUC evaluation with random scaffold split on four property prediction datasets. “Mean” represents the average performance of ImageMol pre-trained with different pre-task on four datasets.



(a) The “elbow” point of ChEMBL dataset

(b) The “elbow” point of ZINC dataset

Figure S28: The “elbow” point of two datasets with respect to the number of clusters. The x-axis represents the number of clusters, and the y-axis represents the sum of Euclidean distances between samples in the clusters. We find the "elbow", which is a value corresponding to the point of maximum curvature in an elbow curve, by using knee point detection algorithm [39].

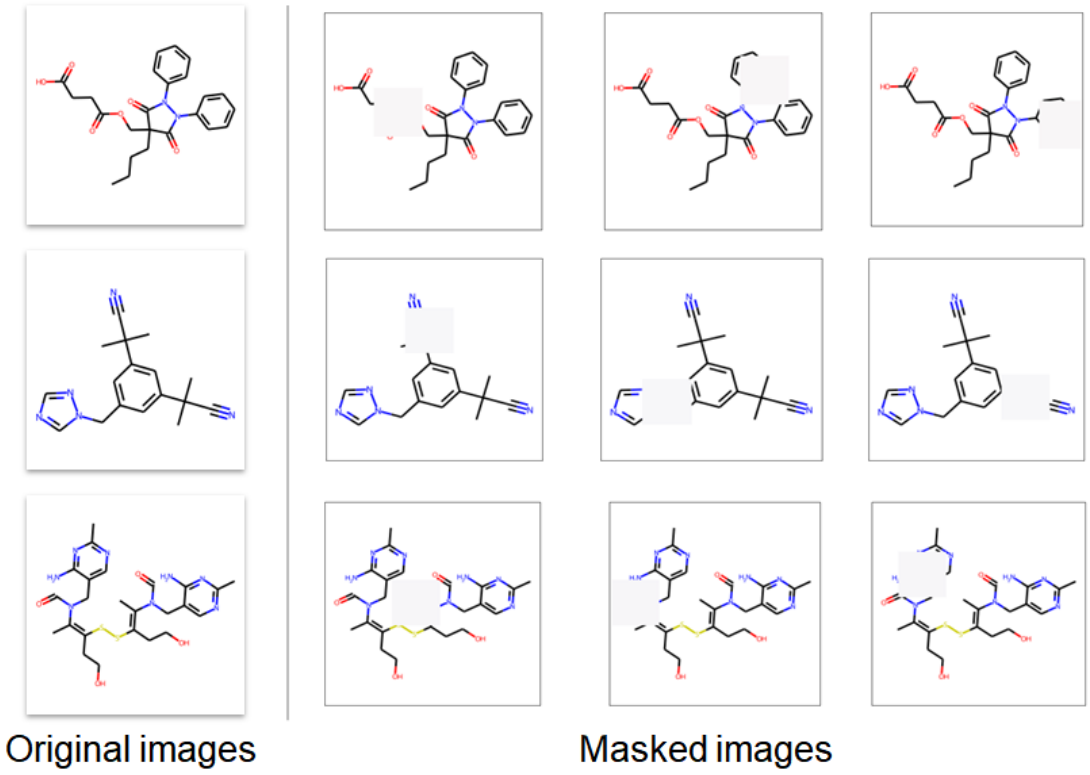


Figure S29: Examples of masked images. The first column represents the original images, and the last three columns represent the masked images.

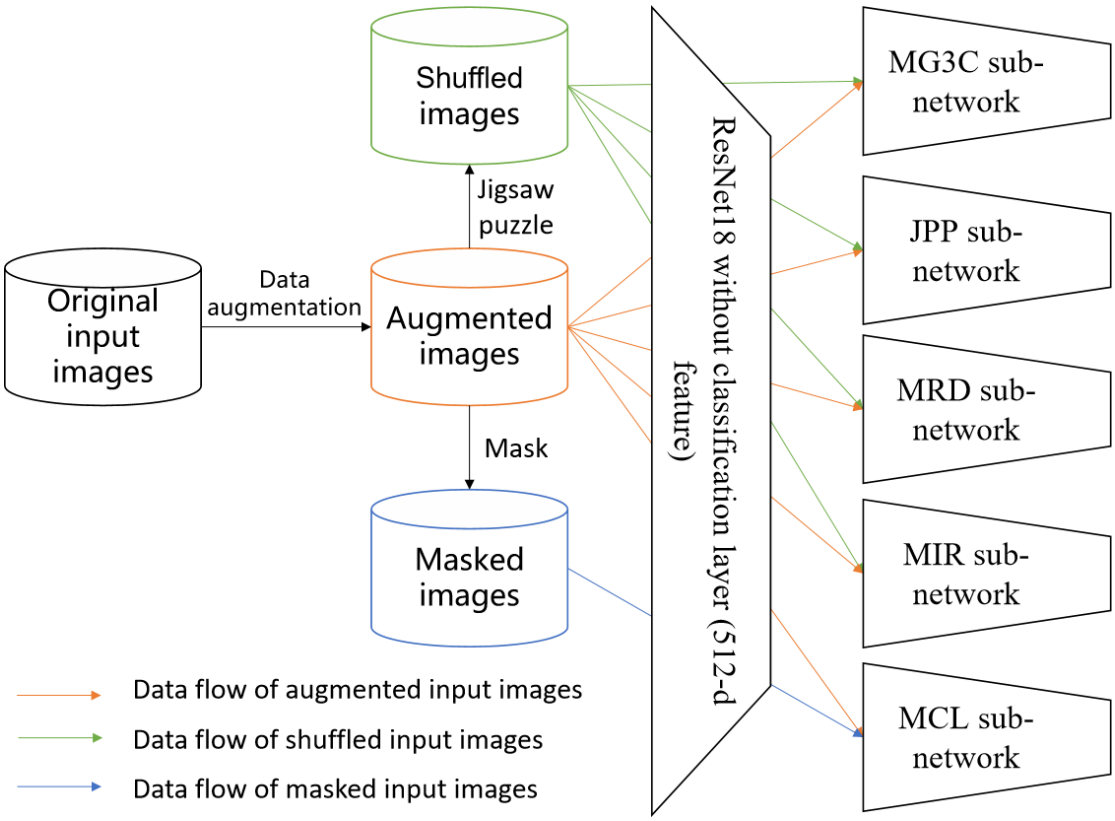
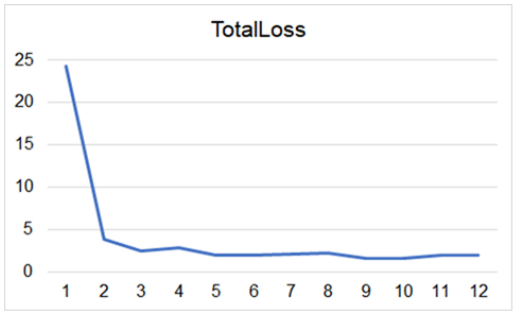
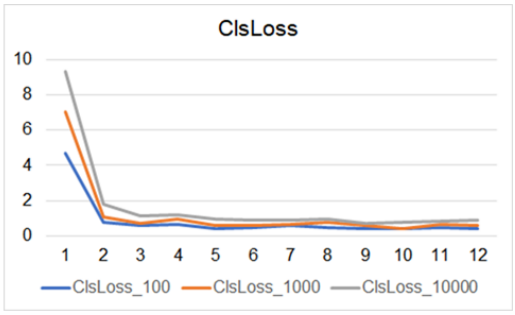


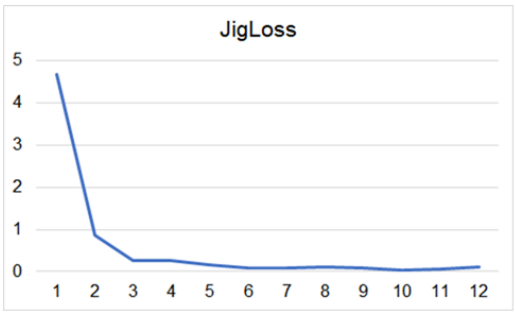
Figure S30: The data flow of the forward propagation of ImageMol framework in pre-training. Data augmentation techniques are first used to extract different augmentations of the original input images and further permutation and masking to obtain shuffled images and masked images, respectively. These images are then fed into ResNet18 to extract latent features. Finally, augmented images are used for five tasks. Shuffled images are used in four tasks. Masked images are used in MCL task.



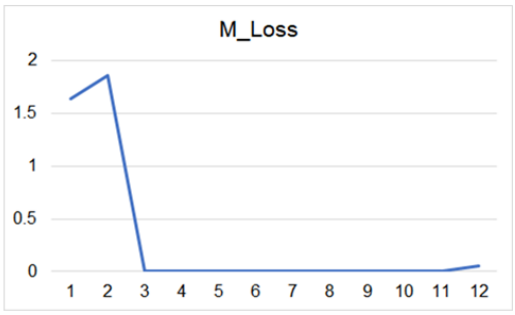
(a) The total loss of ImageMol, which is the sum of the losses of pretext task 1, 2, 3, 4.



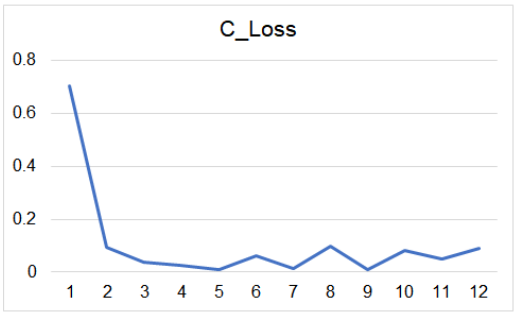
(b) The loss of the pretext task1 during pretraining. ClsLoss_100, ClsLoss_1000 and ClsLoss_10000 represent 100 categories, 1000 categories and 10000 categories.



(c) The loss of the pretext task2 during pretraining.



(d) The loss of the pretext task3 during pretraining.



(e) The loss of the pretext task4 during pretraining.

Figure S31: Pre-training details for ImageMol. The x-axis and y-axis represent epoch number and loss value respectively. For simplicity, clustering pseudo-label classification task, jigsaw puzzle prediction task, molecular rationality discrimination task and MASK-based contrastive learning are simplified to pretext task1, pretext task2, pretext task3 and pretext task4 in this group of subfigures.

Supplementary Tables

Table S1: Basic statistical information of benchmark datasets. Molecules represents the number of molecules. Binary prediction tasks represents the number of binary prediction task. Metric indicates a metric for evaluating model performance. Type indicates the task type of the dataset, including classification task and regression task.

Dataset	Molecules	Binary prediction tasks	Metric	Type
BBBP	2039	1	ROC-AUC	Classification
Tox21	7831	12		
ClinTox	1478	2		
HIV	41127	1		
BACE	1513	1		
SIDER	1427	27		
MUV	93087	17		
ToxCast	8575	617		
FreeSolv	642	1	RMSE	Regression
ESOL	1128	1		
Lipo	4200	1		
QM7	21786	1	MAE	
QM9	133885	8		

Table S2: The statistical information of drug metabolism datasets across 5 main types of CYP450 metabolism enzymes.

Datasets	CYP isoforms	Number of inhibitors	Number of noninhibitors	Total
PubChem Data Set I	1A2	5,663	6,436	12,099
	2C9	4,369	7,761	12,130
	2C19	5,322	6,563	11,885
	2D6	2,516	9,365	11,881
	3A4	4,637	6,899	11,536
PubChem Data Set II	1A2	1,752	1,052	2,804
	2C9	609	1,970	2,579
	2C19	719	1,972	2,691
	2D6	544	2,316	2,860
	3A4	2,070	4,955	7,025

Table S3: The statistical information of ligand-GPCR binding activity datasets across 10 main types of GPCRs with the largest number of reported ligands from the ChEMBL (https://www.ebi.ac.uk/chembl/) database.

Dataset (Assay)	Number of samples	Type
AA1R	3,408	Regression
5HT1A	3,568	
5HT2A	3,079	
AA2AR	3,866	
AA3R	3,306	
CNR2	3,079	
DRD2	5,771	
DRD3	3,945	
HRH3	3,206	
OPRM	2,977	

Table S4: The statistical information of compound-kinase binding activity datasets across 10 main types of biochemical kinase from KinomeScan with percent of control. All datasets are classification tasks. We use control as the criterion of activity, control=100% is inactive (non-inhibitor) and control<100% is active (inhibitor).

Dataset (Assay)	Number of samples	Number of positive samples (percentage)
BTK	106	69 (61.6%)
CDK4-cyclinD3	80	59 (52.7%)
EGFR	105	75 (67.0%)
FGFR1	108	76 (67.9%)
FGFR2	109	75 (67.0%)
FGFR3	107	67 (59.8%)
FGFR4	107	59 (52.7%)
FLT3	110	79 (70.5%)
KPCD3	109	68 (60.7%)
MET	105	70 (62.5%)

Table S5: The ROC-AUC (%) performance of different methods on benchmark datasets with scaffold split. All results are reported as mean ± standard deviation. The yellow and blue backgrounds represent methods without and with pretraining, respectively. The light green background represents the results of our method. Rank represents the ranking of ImageMol in the comparison. (.pdf)

Table S6: The ROC-AUC performance of different methods on benchmark datasets with random scaffold split. All results are reported as mean ± standard deviation. The yellow and blue backgrounds represent methods without and with pretraining, respectively. The light green background represents the results of our method. Rank represents the ranking of ImageMol in the comparison. (.pdf)

Table S7: The comprehensive performance evaluation (including accuracy, AUC, AUPR, F1, precision, recall and kappa) of ImageMol on molecular property prediction task with scaffold split. (.pdf)

Table S8: The comprehensive performance evaluation (including accuracy, AUC, AUPR, F1, precision, recall and kappa) of ImageMol on molecular property prediction task with random scaffold split. (.pdf)

Table S9: The accuracy and ROC-AUC value of five major CYP Isoforms from PubChem Data Set II. ImageMol_NonPretrained and ImageMol indicate that ImageMol is not pre-trained and ImageMol is pre-trained on 10M molecular images. "MACCS-" represents the method based on MACCS molecular fingerprint, "FP4-" represents the method based on FP4 molecular fingerprint. "CC-" represents combined algorithm (ensemble learning). For details on other methods, see [24]. ImageMol_NonPretrained and ImageMol are fine-tuned on PubChem Data Set I and evaluated on PubChem Data Set II. The best and second best result for each dataset is bolded and underlined, respectively. The methods in blue background represent image-based methods. The “3 times” represents the mean ± standard deviation of 3 runs with different random seeds. (.pdf)

Table S10: The performance of different methods on single task CYP450 datasets with balanced scaffold split[16]. (.pdf)

Table S11: The performance of different methods on multi-labeled CYP450 datasets with balanced scaffold split. (.pdf)

Table S12: The performance of different methods on the drug-protein binding efficiency datasets with balanced scaffold split. (.pdf)

Table S13: The performance of different methods on the drug-protein binding activity datasets from KinomeScan with balanced scaffold split. (.pdf)

Table S14: Results of statistical significance tests between ImageMol and GROVER-10M, MPG-10M, X-MOL, MolCLRGIN on the 8 molecular property prediction datasets, where numbers represent the p-values (one-sided significance level) of the McNemar's test between the models in the table (GROVER-10M, MPG-10M, X-MOL and $MolCLR_{GIN}$) and ImageMol. The numbers in green background indicate statistically different models, using a significance threshold of 0.05. (.pdf)

Table S15: Results of statistical significance tests between ImageMol and Image-based methods (Chemception, ADMET-CNN, QSAR-CNN) on five CYP450 isoforms training sets (PubChem Data Set I) and validation sets (PubChem Data Set II), where numbers represent the p-values (one-sided significance level) of the McNemar's test. The numbers in green background indicate statistically different models, using a significance threshold of 0.05. 0 indicates statistical significance less than E-100. (.pdf)

Table S16: Results of statistical significance tests on the 5 CYP450 datasets with balanced scaffold split, where numbers represent the p-values (one-sided

significance level) of the McNemar's test between the models in the table (CHEM-BERT, GROVER, MolCLR$_{GIN}$, MolCLR$_{GCN}$, RNN_LR, RNN_MLP, RNN_RF, TRFM_LR, TRFM_MLP and TRFM_RF) and ImageMol. The numbers in green background indicate statistically different models, using a significance threshold of 0.05. (.pdf)

Table S17: Results of statistical significance tests on the 13 SARS-CoV-2 datasets with balanced scaffold split, where numbers represent the p-values (one-sided significance level) of the McNemar's test between the models in the table (CHEM-BERT, GROVER, MolCLR$_{GIN}$ and MolCLR$_{GCN}$, RNN_LR, RNN_MLP, RNN_RF, TRFM_LR, TRFM_MLP and TRFM_RF) and ImageMol. The numbers in green background indicate statistically different models, using a significance threshold of 0.05. (.pdf)

Table S18: The overview of all binary classification tasks. These datasets are obtained from SARS-CoV-2 assays in NCATS OpenData. Abbreviation represents shorthand for dataset and target category represents the classification type of assay. The number of samples and the number of positive samples are calculated using only samples with AC_{50}.

Dataset (Assay)	Abbreviation	Target Category	Number of samples	Number of positive samples (percentage)
3C-like (3CL) enzymatic activity	3CL	Viral replication	344	63 (18.3%)
angiotensin converting enzyme 2 (ACE2) enzymatic activity	ACE2	Viral entry	508	92 (18.1%)
human embryonic kidney 293 cell line toxicity	HEK293	Counterscreen	4,392	1,753 (33.7%)
Human fibroblast toxicity	hCYTOX	Counterscreen	4,912	1,724 (35.1%)
middle east respiratory syndrome (MERS) Pseudotyped particle entry	MERS-PPE	In vitro infectivity	5,419	1,851 (34.2%)
MERS Pseudotyped particle entry (Huh7 tox counterscreen)	MERS-PPE_cs	Counterscreen	5,606	1,884 (33.6%)
SARS-CoV Pseudotyped particle entry	CoV-PPE	In vitro infectivity	6,496	2,145 (33.0%)
SARS-CoV Pseudotyped particle entry (VeroE6 tox counterscreen)	CoV-PPE_cs	Counterscreen	6,651	2,191 (32.9%)
SARS-CoV-2 cytopathic effect (CPE)	CPE	Live virus infectivity	7,404	2,464 (33.3%)
SARS-CoV-2 cytopathic effect (host tox counterscreen)	Cytotox	Counterscreen	8,865	3,302 (37.3%)
Spike-ACE2 protein-protein interaction (AlphaLISA)	AlphaLISA	Viral entry	9,664	3,570 (36.9%)
Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	TruHit	Counterscreen	10,477	4,027 (38.4%)
transmembrane protease serine 2 (TMPRSS2) enzymatic activity	TMPRSS2	Viral entry	10,658	4,051 (38.0%)

Table S19: The experimental results of Jure's GNN, ImageMol_NonPretrained and ImageMol for anti-SARS-CoV-2 activities estimation on several *SARS-CoV-2 assay* datasets from the National Center for Advancing Translational Sciences (NCATS) COVID-19 portal. The evaluation metrics include the AUC and AUPR. ImageMol_NonPretrained is the ResNet18 randomly initialized weights. ImageMol is our pre-trained model based on PubChem dataset. The best results are bolded and the second best results are underlined. The red value represents the number of performance improvement compared with ImageMol_NonPretrained.

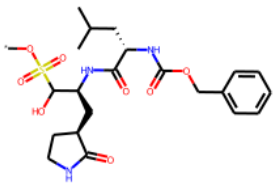
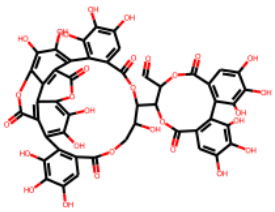
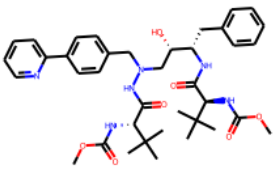
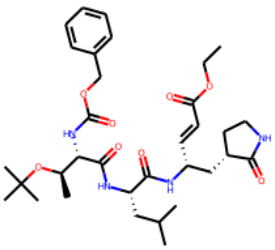
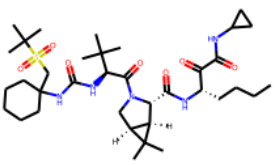
	AUC			**AUPR**		
Dataset	Jure's GNN	ImageMol_NonPretrained	ImageMol	Jure's GNN	ImageMol_NonPretrained	ImageMol
3CL	70.43	<u>76.07</u>	**83.72**$^{+7.65}$	33.96	<u>49.66</u>	**61.99**$^{+12.33}$
ACE2	<u>73.66</u>	65.45	**83.75**$^{+18.30}$	<u>33.56</u>	30.13	**55.69**$^{+25.56}$
HEK293	<u>73.87</u>	70.26	**76.35**$^{+6.09}$	<u>64.32</u>	60.83	**67.37**$^{+6.54}$
hCYTOX	<u>76.19</u>	72.26	**76.40**$^{+4.14}$	<u>66.19</u>	63.52	**66.72**$^{+3.20}$
MERS-PPE	<u>71.71</u>	70.82	**75.78**$^{+4.96}$	<u>60.73</u>	58.78	**65.03**$^{+6.25}$
MERS-PPE_cs	71.19	<u>72.81</u>	**75.41**$^{+2.60}$	57.39	<u>60.96</u>	**66.25**$^{+5.29}$
CPE	69.19	<u>72.06</u>	**74.61**$^{+2.55}$	55.96	<u>58.13</u>	**61.20**$^{+3.07}$
cytotox	72.38	<u>73.88</u>	**76.09**$^{+2.21}$	63.45	<u>65.70</u>	**69.20**$^{+3.50}$
CoV-PPE	<u>70.78</u>	69.95	**74.12**$^{+4.16}$	57.58	<u>57.66</u>	**62.64**$^{+4.98}$
CoV-PPE_cs	70.45	<u>71.60</u>	**74.51**$^{+2.91}$	55.62	<u>58.31</u>	**63.44**$^{+5.13}$
AlphaLISA	69.25	<u>71.28</u>	**74.35**$^{+3.07}$	59.30	<u>62.88</u>	**65.97**$^{+3.09}$
TruHit	67.33	<u>70.19</u>	**72.66**$^{+2.47}$	58.73	<u>61.84</u>	**63.78**$^{+1.94}$
TMPRSS2	67.79	<u>70.09</u>	**72.57**$^{+2.48}$	58.69	<u>61.52</u>	**63.92**$^{+2.40}$

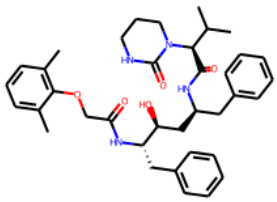
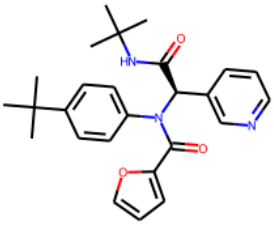
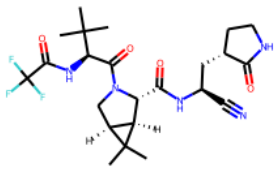
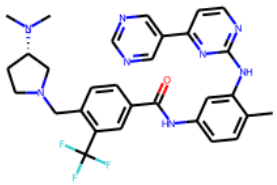
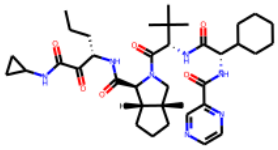
Table S20: The experimental results of REDIAL-2020, ImageMol_NonPretrained and ImageMol for anti-SARS-CoV-2 activities estimation. ACC, accuracy; F1, F1 score; SEN, sensitivity; PREC, precision; AUC, area under the receiver operating characteristic curve. ImageMol (3 times) represents the mean ± standard deviation of the results of 3 runs with different random seeds. (.pdf)

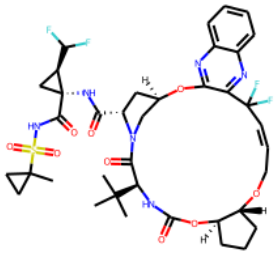
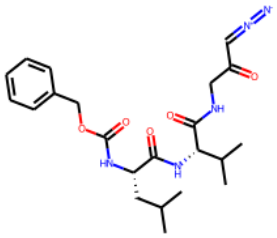
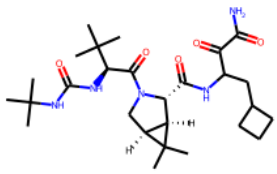
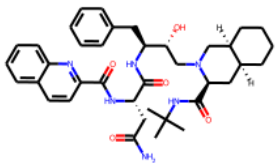
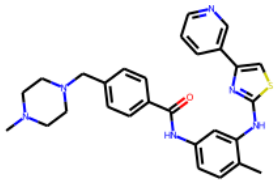
Table S21: The performance of different methods on 13 SARS-CoV-2 datasets with balanced scaffold split[16]. (.pdf)

Table S22: Screening results of approved drugs in DrugBank for 3CL inhibitors via ImageMol. (.pdf)

Table S23: Screening results from 16 known 3CL inhibitors.

Index	Drug name	Structure	Probability	Evidence Citations
1	**GC376**		**0.99692**	47
2	**Punicalagin**		**0.98766**	48
3	**Atazanavir**		**0.97574**	48
4	**Compound 4**		**0.95191**	49
5	**Narlaprevir**		**0.82554**	50

6	**Lopinavir**		**0.81443**	50
7	**ML188**		**0.79253**	51
8	**Nirmatrelvir**		**0.73093**	52
9	**Bafetinib**		**0.68171**	53
10	**Telaprevir**		**0.62670**	50

11	**Glecaprevir**		**0.14985**	50
12	**Z-LVG-CHN2**		**0.14434**	54
13	**Boceprevir**		**0.09992**	55
14	**Saquinavir**		**0.00398**	50
15	**Masitinib**		**0.00001**	53

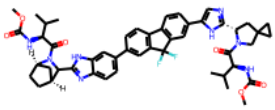
16	**Ledipasvir**		**0.00001**	50

Table S24: Screening results of approved drugs in DrugBank for SARS-CoV-2. (.pdf)

Table S25: Virtual screening of 70 validated anti-SARS-CoV-2 small molecule drugs. These drugs were validated in Calu-3 cells [56]. (.pdf)

Table S26: Examples of using high attention value to control the highlighted area of GradCAM, which can better inform chemists.

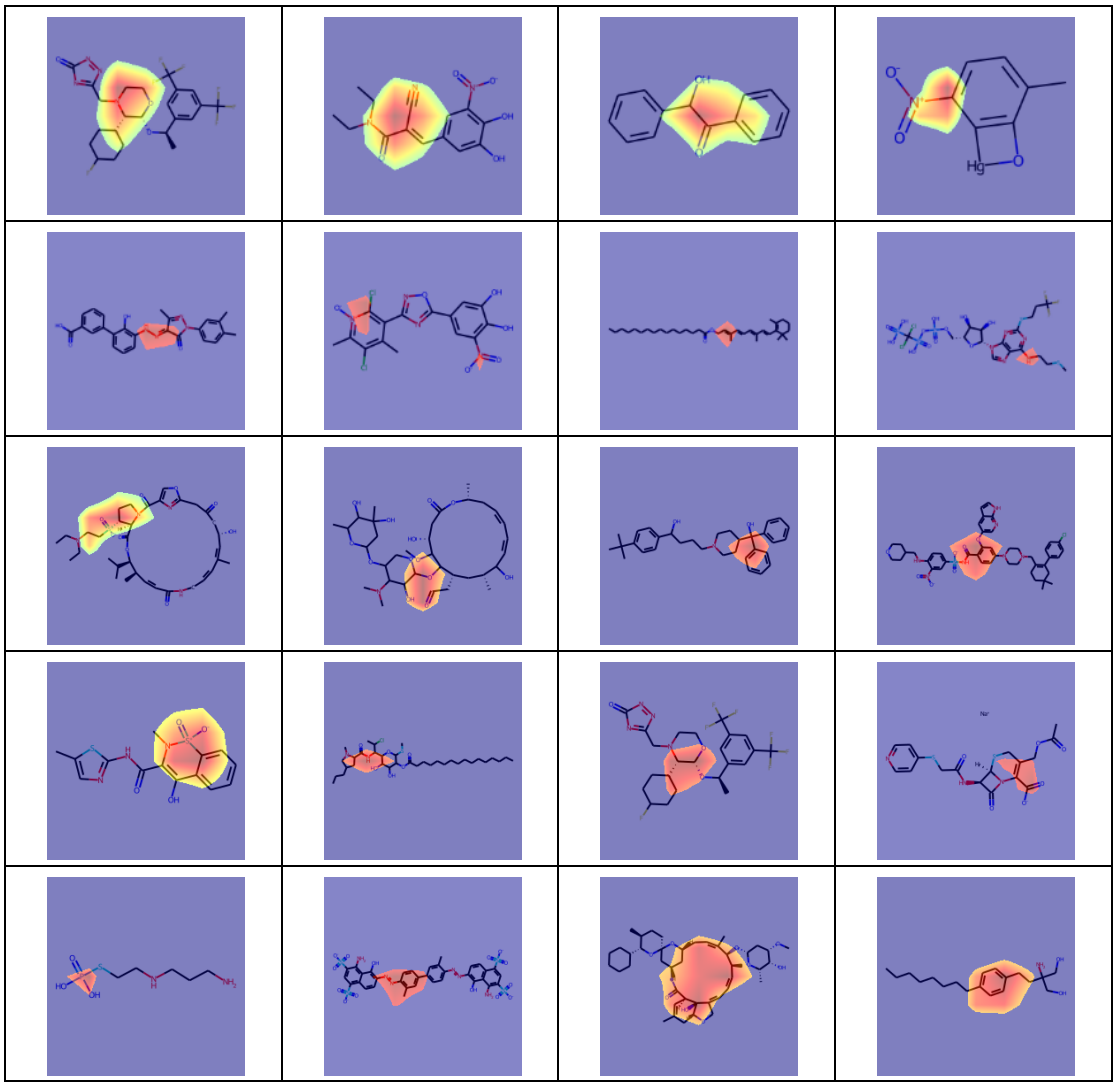


Table S27: The statistical information of anti-SARS-CoV-2 activity datasets[38]. Sparsity refers to the proportion of blank areas in an image to the entire image.

Datasets	Number of actives	Number of inactives	Sparsity (%)
3CL	81	3,330	94.93
CPE	44	2,913	94.89
ACE2	70	1,192	94.88
Cytotox	193	2,764	94.86
AlphaLISA	143	1,119	94.86
TruHit	134	1,128	94.86
CoV-PPE	43	881	94.91
CoV-PPE_cs	247	1,085	94.91
hCYTOX	81	1,306	94.87
MERS-PPE	104	1,024	94.88
MERS-PPE_cs	46	1,082	94.88

Table S28: Ablation results of data augmentation on benchmark datasets with ROC-AUC metric and random scaffold split. All results are reported as mean ± standard deviation. Bold font indicates best result. (.pdf)

Table S29: The several examples of data augmentation. The first column represents the original image, and the last three columns represent different data augmentation strategies, namely flip, grayscale and rotation. The similarity represents the cosine similarity between the augmented image and the original image in the embedding vector, which is obtained from the pretrained ImageMol.

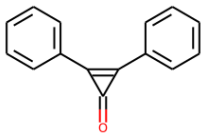
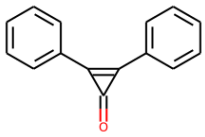
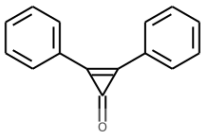
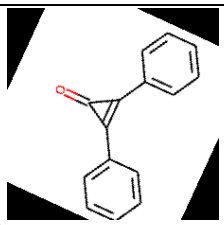
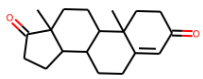
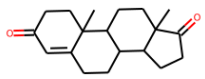
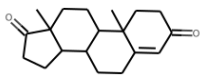
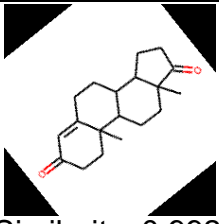
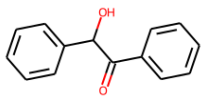
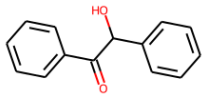
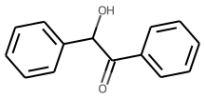
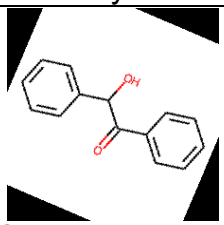
Original image	Flip	Grayscale	Rotation
	 Similarity: 0.999	 Similarity: 0.996	 Similarity: 0.9729
	 Similarity: 0.996	 Similarity: 0.999	 Similarity: 0.992
	 Similarity: 0.999	 Similarity: 0.999	 Similarity: 0.996

Table S30: Hyperparameters for pre-training and finetuning ImageMol.

Hyperparameters	Pre-training	Fine-tuning
Learning rate	0.01	5e-4~0.5
Batch Size	256	8,16,32,64,128
Weight Decay	1e-5	1e-5
Max Epochs	15	10~60
Learning Rate Decay	Linear	Linear
Image Size	224×224×3	224×224×3
Classification Layer Number	0	1, 2

Table S31: Hyperparameter optimization for MPG, GROVER, X-MOL, MolCLR, CHEM-BERT, SMILES_Transformer, Chemception, ADMET-CNN and QSAR-CNN. (.pdf)

Supplementary References

1. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513-530 (2018).
2. Xiong, Z. et al. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **63**, 8749-8760 (2020).
3. Ramsundar, B. et al. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* (2015).
4. Xue, D. et al. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. *Sci. Bull.* **67**, 899-902 (2022).
5. Welling, M. & Kipf, T.N. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations* (ICLR 2017).
6. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595-608 (2016).
7. Schütt, K. et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems* **30** (MIT Press, 2017).
8. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. & Dahl, G.E. Neural message passing for quantum chemistry. *International conference on machine learning*, 1263-1272 (PMLR, 2017).
9. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Informat. Model.* **59**, 3370-3388 (2019).
10. Lu, C. et al. Molecular property prediction: A multilevel quantum interactions modeling perspective. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 1052-1060 (AAAI, 2019).
11. Hu, W. et al. Strategies For Pre-training Graph Neural Networks. *International Conference on Learning Representations (ICLR)* (ICLR, 2020).
12. Liu, S., Demirel, M.F. & Liang, Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems* **32** (MIT Press, 2019).
13. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279-287 (2022).
14. Qiu, J. et al. Gcc: Graph contrastive coding for graph neural network pre-training. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1150-1160 (KDD, 2020).
15. Hu, Z., Dong, Y., Wang, K., Chang, K.-W. & Sun, Y. Gpt-gnn: Generative pre-training of graph neural networks. *Proceedings of the*

26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1857-1867 (2020).
16. Rong, Y. et al. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems* **33**, 12559-12571 (MIT Press, 2020).
17. Zhang, Z., Liu, Q., Wang, H., Lu, C. & Lee, C.-K. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. *Advances in Neural Information Processing Systems* **34** (MIT Press, 2021).
18. Stärk, H. et al. 3D Infomax improves GNNs for Molecular Property Prediction. *NeurIPS 2021 AI for Science Workshop* (MIT Press, 2021).
19. Xu, M., Wang, H., Ni, B., Guo, H. & Tang, J. Self-supervised graph-level representation learning with local and global structure. *International Conference on Machine Learning*, 11548-11558 (PMLR, 2021).
20. You, Y. et al. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems* **33**, 5812-5823 (MIT Press, 2020).
21. Liu, S. et al. Pre-training Molecular Graph Representation with 3D Geometry. *International Conference on Learning Representations* (ICLR, 2021).
22. Li, P. et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Brief. Bioinformat.* **22**, bbab109 (2021).
23. Goh, G.B., Siegel, C., Vishnu, A., Hodas, N.O. & Baker, N.J.a.p.a. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. arXiv:1706.06689 (2017).
24. Cheng, F. et al. Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers. *J. Chem. Informat. Model.* **51**, 996-1011 (2011).
25. Shi, T. et al. Molecular image-based convolutional neural network for the prediction of ADMET properties. *Chemometrics and Intelligent Laboratory Systems* **194**, 103853 (2019).
26. Zhong, S., Hu, J., Yu, X. & Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **408**, 127998 (2021).
27. Honda, S., Shi, S. & Ueda, H.R.J.a.p.a. SMILES transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv:1911.04738 (2019).
28. Kim, H., Lee, J., Ahn, S. & Lee, J.R. A merged molecular representation learning for molecular properties prediction with a web-based service. *Sci. Rep.* **11**, 1-9 (2021).

29. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700-4708 (IEEE, 2017).
30. Medsker, L.R. & Jain, L. Recurrent neural networks. *Design and Applications* **5**, 64-67 (2001).
31. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems*, 5998-6008 (MIT Press, 2017).
32. LaValley, M.P. Logistic regression. *Circulation* **117**, 2395-2399 (2008).
33. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **2**, 183-197 (1991).
34. Breiman, L. Random forests. *Machine Learning* **45**, 5-32 (2001).
35. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186 (ACL, 2019).
36. Sterling, T. & Irwin, J.J. ZINC 15–ligand discovery for everyone. *J. Chem. Informat. Model.* **55**, 2324-2337 (2015).
37. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *arXiv:1810.00826* (2018).
38. Bocci, G. et al. A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nat. Mach. Intell.*, 1-9 (2021).
39. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. *2011 31st international conference on distributed computing systems workshops*, 166-171 (IEEE, 2011).
40. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.J.J.o.c.i. & sciences, c. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42**, 1273-1280 (2002).
41. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B. & Tommasi, T. Domain generalization by solving jigsaw puzzles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229-2238 (IEEE, 2019).
42. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A.A. Context encoders: Feature learning by inpainting. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536-2544 (IEEE, 2016).
43. Arús-Pous, J. et al. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminformat.* **12**, 1-18 (2020).
44. Goss, K.-U. & Schwarzenbach, R.P. Rules of thumb for assessing equilibrium partitioning of organic compounds: successes and pitfalls.

J. Chem. Educ. **80**, 450 (2003).
45. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on Machine Learning* **28**, 1139-1147 (PMLR, 2013).
46. Selvaraju, R.R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618-626 (IEEE, 2017).
47. Ma, C. et al. Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res.* **30**, 678-692 (2020).
48. Du, R. et al. Discovery of chebulagic acid and punicalagin as novel allosteric inhibitors of SARS-CoV-2 3CLpro. *Antiviral Res.* **190**, 105075 (2021).
49. Iketani, S. et al. Lead compounds for the development of SARS-CoV-2 3CL protease inhibitors. *Nat. Commun.* **12**, 1-7 (2021).
50. Sun, Q. et al. Bardoxolone and bardoxolone methyl, two Nrf2 activators in clinical trials, inhibit SARS-CoV-2 replication and its 3C-like protease. *Signal Transduction and Targeted Therapy* **6**, 1-3 (2021).
51. Jacobs, J. et al. Discovery, synthesis, and structure-based optimization of a series of N-(tert-butyl)-2-(N-arylamido)-2-(pyridin-3-yl) acetamides (ML188) as potent noncovalent small molecule inhibitors of the severe acute respiratory syndrome coronavirus (SARS-CoV) 3CL protease. *J. Med. Chem.* **56**, 534-546 (2013).
52. Vandyck, K. & Deval, J. Considerations for the Discovery and Development of 3-Chymotrypsin-Like Cysteine Protease Inhibitors Targeting SARS-CoV-2 Infection. *Current Opinion in Virology* (2021).
53. Drayman, N. et al. Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2. *Science* **373**, 931-936 (2021).
54. Riva, L. et al. A large-scale drug repositioning survey for SARS-CoV-2 antivirals. *BioRxiv,* DOI: 10.1101/2020.04.16.044016 (2020).
55. Hu, F. et al. A novel framework integrating AI model and enzymological experiments promotes identification of SARS-CoV-2 3CL protease inhibitors and activity-based probe. *Brief. Bioinformat.* **22**, bbab301 (2021).
56. Schultz, D.C. et al. Pyrimidine inhibitors synergize with nucleoside analogues to block SARS-CoV-2. *Nature*, 604, 134-140 (2022).