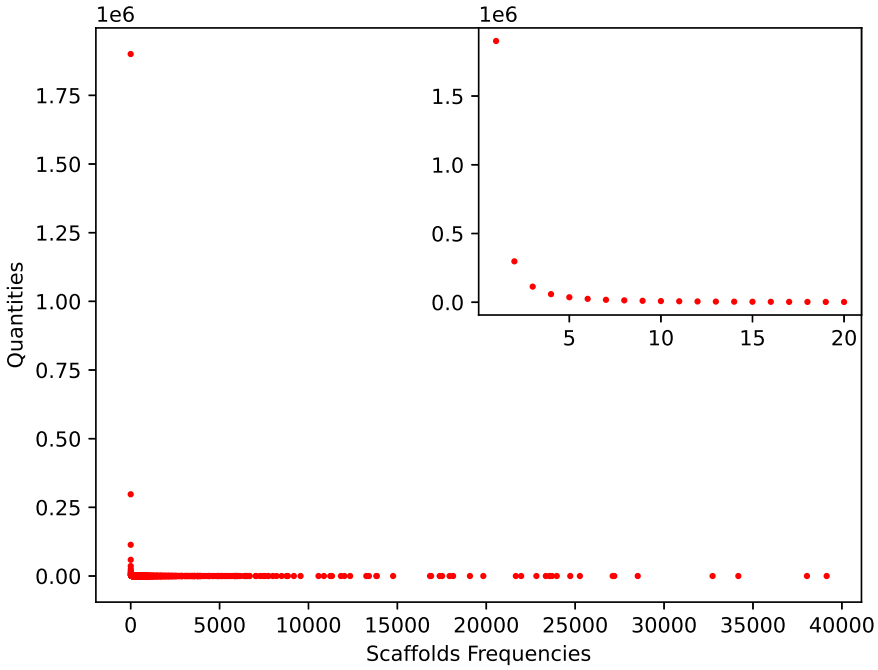


Fig. S1: Frequency statistics of molecular scaffolds in the initial dataset. The horizontal axis represents the frequency of occurrence of molecular scaffolds, while the vertical axis represents the quantity of occurrences at different frequencies.

TABLE SI: Sub-datasets in MoleculeNet used for the downstream evaluation. We list the dataset split method, evaluation metric, the number of tasks it has, and the number of molecules included in each sub-dataset.

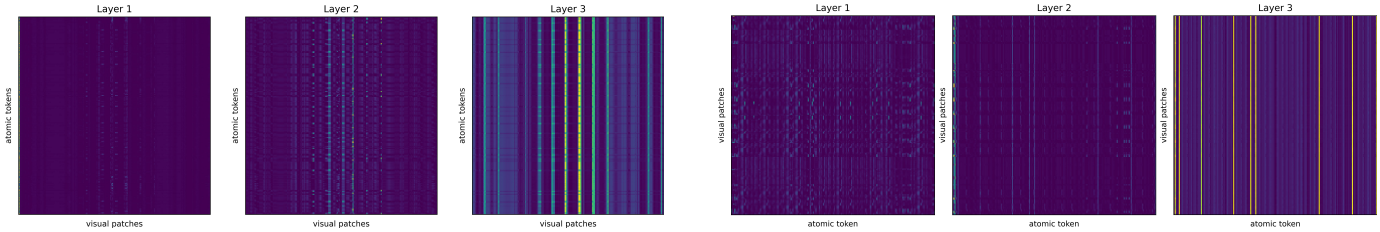
ADMET	Datasets	Describe
Absorption	Pgp-substrate	Modulating P-glycoprotein mediated transport of its substrate has significant pharmacokinetic implications for Pgp substrates, which can be either leveraged for specific therapeutic advantages or lead to contraindications.
	HIA	Human intestinal absorption. The human intestinal absorption of an oral drug is the essential prerequisite for its apparent efficacy.
	F20%	One of the most important pharmacokinetic parameters for any drug administered is its oral bioavailability.This parameter serves as an indicator of the efficiency with which the drug is delivered to the systemic circulation.The human oral bioavailability 20%.
	F30%	The human oral bioavailability 30%.
Metabolism	CYP1A2-sub	The substrates of P450. The human cytochrome P450 family contains 57 isozymes and these isozymes metabolize approximately two-thirds of known drugs in human with 80% of this attribute to five isozymes–1A2, 3A4, 2C9, 2C19 and 2D6.
	CYP2C19-sub	
	CYP2C9-sub	
	CYP2D6-sub	
	CYP3A4-sub	
Excretion	T12	Reliable estimates of clearance and volume of distribution are crucial in determining the half-life of a drug, as this hybrid concept relies on both properties.
Toxicology	DILI	The evaluation of whether human liver could be injured by a compound.
	FDAMDD	The maximum daily dose that is recommended serves as an approximation of the threshold for toxic doses of chemicals in humans.
	Skin Sensitization	The evaluation of whether a compound that comes into contact with the skin can induce allergic contact dermatitis.
	Respiratory Toxicity	The evaluation of whether a compound is toxic to the human respiratory tract.

TABLE SII: The search method and search range of hyper-parameters during fine-tuning stage.

Detail	**Description**	
Method	Grid search strategy	
Parameters Ranges	batch size	[8,16,32]
	learning rate	[5e-6, 1e-5, 5e-5, 1e-4]
	dropout rate	[0.1,0.2]
	max steps	[1500, 2000, 2500, 3000]
The total number of hyperparameter configurations	96	

TABLE SIII: Comparing the significance of each representative method within their respective categories, we found that the results conform to a normal distribution. Therefore, we employed a Student's t-test for statistical comparisons. We bolded the p-values that were less than 0.05 to indicate statistical significance.

p-value	**Graph-based**	**SMILES-based**	**Image-based**
Datasets	CD-MVGNN vs ISMol	ChemBERTa-77M-MTR vs ISMol	ImageMol vs ISMol
Pgp-sub	0.763	**0.004**	0.410
HIA	**0.033**	0.369	0.591
F(30%)	**0.026**	0.228	0.309
CYP1A2	0.743	0.164	0.742
CYP2C19	0.066	**0.009**	**0.039**
CYP2C9	0.077	**0.049**	0.104
CYP2D6	0.724	0.550	0.749
CYP3A4	**0.012**	**0.001**	**0.0002**
DILI	0.152	0.253	0.797
SkinSen	**0.002**	**0.005**	**0.016**
Respiratory	**0.047**	0.697	0.674



(a) SMILES Cross-attention (b) Visual Cross-attention

Fig. S2: Visualization of complete cross-attention maps regarding molecule 'O=[P+](O)c1ccc(Cl)cc1' in the multimodality fusion module of ISMol (contains blank patches).

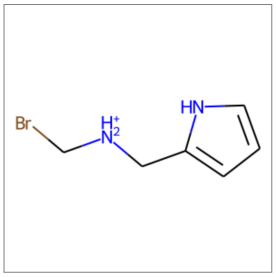
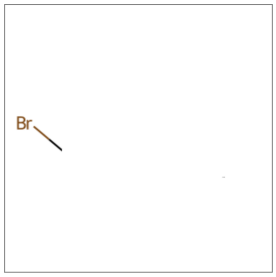
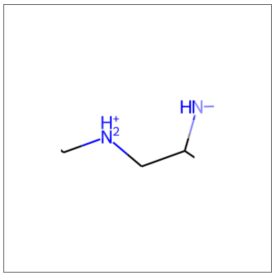
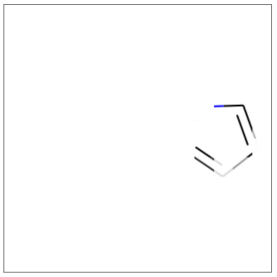
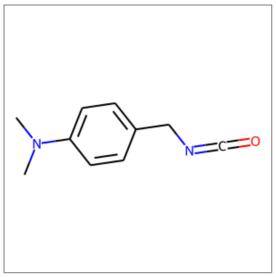
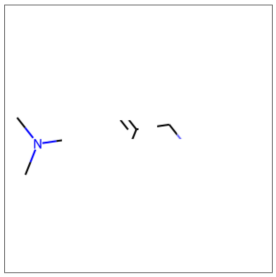
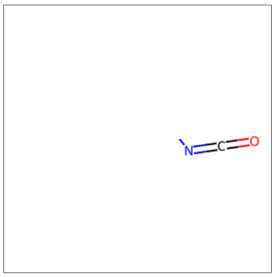
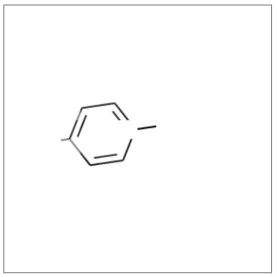
TABLE SIV: The quantity and proportion of scaffolds appearing in the 14 downstream ADMET tasks dataset among all scaffolds in the pre-training dataset. Where Number of scaffolds is the number of scaffolds types in the downstream dataset, and totalCnts is the number of scaffolds in the downstream dataset that have appeared in the pre-training dataset. In addition, Percent is totalCnts divided by Number of scaffolds, indicating the proportion of downstream dataset scaffolds that ISMol has seen in the pre-training.

Datasets	**Number of scaffold**	**3.5M data used in ISMol**		**10M data used in ImageMol**	
		totalCnts	**Percent**	**totalCnts**	**Percent**
Pgp-sub	1185	406	0.343	638	0.538
HIA	1160	776	0.669	950	0.819
F(20%)	992	624	0.629	793	0.799
F(30%)	992	624	0.629	793	0.799
FDAMDD	1197	817	0.683	972	0.812
CYP1A2	366	226	0.617	283	0.773
CYP2C19	258	188	0.729	217	0.841
CYP2C9	811	533	0.657	656	0.809
CYP2D6	877	578	0.659	712	0.812
CYP3A4	979	592	0.605	737	0.753
T12	1219	732	0.600	982	0.806
DILI	467	325	0.696	378	0.809
SkinSen	405	248	0.612	269	0.664
Respiratory	1388	706	0.509	924	0.666
Average	878.3	526.8	0.617	664.6	0.764

TABLE SV: Statistical results of a dataset split over time for small molecules targeting the popular target EGFR.

Split Way		**Train**	**Valid**	**Test**
prospect	Date	2004.09.16-2005.08.08	2005.08.09-2005.09.12	2005.09.13-2005.10.07
	Total	3356	1257	1020
	Positive:Negative	621:2735	365:892	279:741
retrospect	Date	2005.07.15-2005.10.07	2005.03.27-2005.07.14	2004.09.16-2005.03.26
	Total	3304	1282	1047
	Positive:Negative	886:2418	186:1096	193:854

TABLE SVI: Visualization of more molecular image examples aligned with SMILES substrings. We highlight the queried substrings for emphasis with the **focus** style.

Original image	**SMILES** (with **focus**)		
	BrC[NH2+]Cc1ccc[nH]1	Br**C[NH2+]C**c1ccc**[nH]**1	BrC[NH2+]C**c1ccc**[nH]1
			
	CN(C)c1ccc(CN=C=O)cc1	CN(C)**c1ccc**(CN=C=O)**cc1**	CN(C)c1ccc(**CN=C=O**)cc1
			
	C#CC([NH2+]C)c1ccccc1OC	C#C**C([NH2+]C)**c1ccccc1OC	C#CC([NH2+]C)**c1ccccc1OC**