Supplementary information for "Molecular Motif Learning as a pretraining objective for molecular property prediction"

Ziyang Liu[1], Chaokun Wang[1*], Shuwen Zheng[1], Cheng Wu[1], Hao Feng[1], Li Xu[2], Yue Zheng[2], Liang Rong[2], Peng Li[2,3*]

[1]School of Software, Tsinghua University, Beijing 100084, China.
[2]Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China.
[3]Tianjian Laboratory of Advanced Biomedical Sciences, Zhengzhou University, Zhengzhou 450001, China.
[*]Corresponding author.

*Corresponding author(s). E-mail(s): chaokun@tsinghua.edu.cn; li-peng@mail.tsinghua.edu.cn;
Contributing authors: liu-zy21@mails.tsinghua.edu.cn; zsw22@mails.tsinghua.edu.cn; wuc22@mails.tsinghua.edu.cn; fh20@mails.tsinghua.edu.cn; xulilulu@tsinghua.edu.cn; zheng-y@mail.tsinghua.edu.cn; rongl23@mails.tsinghua.edu.cn;

A Related work

A.1 Diffusion model

Diffusion models have emerged as a powerful class of generative models, gaining significant attention in recent years due to their impressive performance in various generative tasks, including image generation [1, 2], text generation [3, 4], and molecular design [5–7]. The core idea behind diffusion models is to gradually add noise to the data in a forward process, and then learn a reverse process that progressively denoises the data to generate high-quality samples. Recent advancements have focused on improving the efficiency and scalability of diffusion models. For instance, research has explored ways to accelerate the sampling process, which is traditionally slow due to the iterative nature of diffusion models. Techniques such as improved sampling methods and the use of fewer steps in the reverse process have been proposed to address this issue [8, 9]. Additionally, there has been significant progress in applying diffusion models to a broader range of domains, including the successful adaptation of these models to generate complex molecular structures and optimize drug-like properties [10, 11]. Moreover, diffusion models have been integrated with other generative frameworks, such as GANs and VAEs, to enhance their expressiveness and training stability. This hybrid approach has shown promising results in generating more diverse and realistic data [12, 13].

One of the inherent strengths of diffusion models is their role as Gaussian denoisers, which naturally endows them with robustness to noise perturbations [14–17]. This characteristic is particularly beneficial in the context of molecular contrastive learning, where the molecule graph augmentation can leave models

vulnerable to the effects of noise perturbations in the molecular graph. Diffusion models can effectively compensate for it by maintaining stability and performance in noisy environments. This robustness to noise is often overlooked by current augmentation-free contrastive learning methods [18–20], highlighting a significant advantage of integrating diffusion models into molecular contrastive learning (MCL).

A.2 Graph representation learning on small molecules

Representation learning on small molecular graphs has gained significant traction in recent years, driven by advancements in deep learning and graph neural networks (GNN). Molecular graphs, where atoms are represented as nodes and chemical bonds as edges, provide a natural and effective way to model molecular structures, enabling the application of GNN for various chemical and biological tasks [21–23].

Methods such as MPNN [24], DMPNN [25],CMPNN [26], and GEM [21] have been extensively applied to the small molecule tasks such as molecular property prediction. MPNN [24] is a foundational framework in graph neural networks that propagates and aggregates information through nodes and edges in a graph, enabling the model to learn rich node and graph-level representations. DMPNN [25] extends the MPNN framework by incorporating directionality in the message passing process, which allows for more precise modeling of directional relationships between atoms in molecular graphs. CMPNN [26] enhances molecular representations by reinforcing the interactions between nodes and edges through a communicative kernel. GEM [21] models the atom-bond-angle relationships to create a geometry-based GNN that effectively encodes both the topological and geometric information of molecules.

Self-supervised learning has evolved and been successfully applied to molecular tasks, encompassing two main approaches: predictive self-supervised learning and contrastive self-supervised learning. Predictive self-supervised learning methods in molecular modeling focus on predicting specific aspects of molecular structures to learn useful representations without relying on labeled data. N-GRAM [27] draws inspiration from natural language processing, modeling the molecular graph as a sequence of substructures (n-grams) and training the model to predict the presence of these substructures in a given context. MGSSL [28] uses a masked graph modeling approach, where certain parts of the molecular graph are masked and the model is tasked with predicting the masked components, thereby learning rich representations that capture both local and global molecular features. Contrastive self-supervised learning has emerged as a powerful paradigm for learning robust molecular representations by contrasting positive and negative molecule pairs. GraphMVP [29] employs a multiview approach, leveraging both 2D molecular graphs and 3D molecular structures to capture complementary information for more accurate predictions. MolCLR [22] introduces three molecular graph augmentation strategies—atom masking, bond deletion, and subgraph removal—to create contrastive pairs, enabling the learning of molecular graph representations using general GNN backbones. KANO [23] integrates domain knowledge with contrastive learning, enhancing the model's ability to distinguish between subtle structural variations by incorporating chemically relevant augmentations and functional prompt. The negative effects of molecular graph augmentation have hindered the advancement of MCL. The limitations of molecular graph augmentations, such as their tendency to disrupt semantic consistency, have been shown to hinder the progress of MCL, motivating the development of augmentation-free approaches [19, 30, 31]. For example, Zhao et al. [18] introduce an augmentation-free MCL using a graph transformer, but they overlook the model's robustness to the noise perturbations in the molecular graph, which also limits the model's effectiveness.

Existing works have explored various ways to incorporate molecular motifs into representation learning. Specifically, MGSSL [28] uses BRICS-derived motifs as generation targets in an autoregressive framework, while DGPM [32] discovers motifs via edge pooling to align subgraph kernels in a pretraining task. MOTOR [33] employs motifs as relational descriptors for drug-drug interaction prediction, and HM-GNN [34] integrates motifs as heterogeneous graph nodes to facilitate message passing between molecules. MICRO-Graph [35], on the other hand, clusters motif-like subgraphs to guide contrastive learning. While these methods demonstrate the utility of motifs for tasks like generation, pretraining, or interaction modeling, they share a key limitation: motifs serve as fixed tools or intermediate constructs rather than explicitly learned, transferable representations. In contrast, MotiL repositions motifs as the

central representation units, optimizing them for semantic consistency and alignment across diverse molecular contexts. Unlike prior works that use motifs to assist molecule-level learning, MotiL directly learns generalizable, interpretable motif embeddings that can transfer across tasks and scales, ranging from small molecules to macromolecules.

A.3 Representation learning on proteins

Protein representation learning has garnered increasing attention in protein modeling and structural bioinformatics, playing a crucial role in biological research. Given that proteins are sequences of amino acids, methods such as 1D CNNs [36], LSTMs [37], and Transformers [37] have been widely employed for sequence-based protein representation learning. Beyond sequence information, 3D geometric structures, including amino acid or atom 3D coordinates and protein surfaces, have been utilized to enhance protein representations [38–40]. Additionally, several methods have been proposed to integrate both 1D sequential order and 3D geometric coordinates of amino acids [41–45]. For example, CDConv [45] models regular sequences and irregular coordinates in protein data through discrete and continuous representations, respectively. However, these methods are primarily based on supervised learning, which limits their ability to learn from unlabeled data. Graph self-supervised learning, particularly graph contrastive learning, which can extract representations directly from protein graph data without relying on labeled proteins, remains underexplored in this context. This limitation constrains the full potential of these approaches such as CDConv in capturing the complex nature of protein structures.

A.4 Sequence-based and SMILES-based molecular representation learning

In addition to graph-based representations, molecular property prediction has also benefited from linear representations such as SMILES strings and amino acid sequences. These formats treat molecules and proteins as sequences, enabling the application of natural language processing (NLP) techniques to molecular modeling. For instance, models like ChemBERTa [46] and SMILES-BERT [47] leverage transformer-based architectures to learn contextual embeddings from large-scale unlabeled SMILES corpora. Similarly, the models such as ESM [48] utilize masked language modeling on millions of amino acid sequences to capture biological semantics. A key advantage of these sequence-based approaches is their scalability: the simplicity of SMILES and FASTA formats allows access to massive datasets and efficient pretraining pipelines. However, this comes with trade-offs. Unlike molecular graphs, which naturally encode topological and stereochemical information, SMILES strings often linearize or obscure such structural nuances. Therefore, while sequence models offer practical benefits in terms of data volume and pretraining efficiency, they may struggle to preserve fine-grained spatial or functional substructures critical to downstream molecular tasks. Our work focuses on graph-based models as they offer a more faithful structural representation, particularly advantageous for capturing and aligning motifs in molecular graphs.

B Overview of experimental datasets

We use 14 benchmark datasets for the small molecular property prediction task, as detailed in Supplementary Table 1. These datasets, sourced from MoleculeNet [64], encompass a diverse range of molecular characteristics and are categorized into four primary domains: physiology, biophysics, physical chemistry, and quantum mechanics. The physiology datasets (BBBP, Tox21, ToxCast, SIDER, and ClinTox) focus on the behavior of molecules within biological systems, while the biophysics datasets (BACE, MUV, and HIV) apply physical principles to study biological phenomena. The physical chemistry datasets (ESOL, FreeSolv, and Lipophilicity) explore chemical behaviors and principles through a physics-based lens, and the quantum mechanics datasets (QM7, QM8, and QM9) delve into the fundamental electronic properties at the atomic level. Each dataset is characterized by its task type, evaluation metrics, number of tasks, number of compounds, and splitting strategies, offering a comprehensive framework for evaluating various molecular properties.

Task Type	Metric	Category	Dataset	# Tasks	# Compounds	Split
Classification	ROC-AUC	Physiology	BBBP [49]	1	2,039	Scaffold split
		Physiology	Tox21 [50]	12	7,831	Scaffold split
		Physiology	ToxCast [51]	617	8,597	Scaffold split
		Physiology	SIDER [52]	27	1,427	Scaffold split
		Physiology	ClinTox [53]	2	1,478	Scaffold split
		Biophysics	BACE [54]	1	1,513	Scaffold split
		Biophysics	MUV [55]	17	93,087	Scaffold split
		Biophysics	HIV [56]	1	41,127	Scaffold split
Regression	RSME	Physical chemistry	ESOL [57]	1	1,128	Scaffold split
		Physical chemistry	FreeSolv [58]	1	642	Scaffold split
		Physical chemistry	Lipophilicity [59]	1	4,200	Scaffold split
	MAE	Quantum mechanics	QM7 [60]	1	6,830	Scaffold split
		Quantum mechanics	QM8 [61]	12	21,786	Scaffold split
		Quantum mechanics	QM9 [62]	3	133,885	Random split

Supplementary Table 1 Summary of the experimental datasets on small molecules. # Tasks indicates the number of binary classification or regression tasks within each dataset.

Prediction objective	Metric	Dataset	# Classes	# Train	# Valid	# Test
Gene ontology term	F_{max}	Biological process (BP) [63]	1,943	29,898	3,322	3,415
		Molecular function (MF) [63]	489			
		Cellular component (CC) [63]	320			
Enzyme commission number	F_{max}	Enzyme Commission (EC) [63]	538	15,550	1,729	1,919

Supplementary Table 2 Summary of the experimental datasets on protein macromolecules.

We selected the MoleculeNet benchmark suite for several key reasons, which we summarize here for completeness. First, MoleculeNet has become a widely recognized standard for evaluating molecular property prediction models, particularly in the context of graph neural networks and self-supervised learning. Many state-of-the-art methods, including MolCLR [22], GROVER [65], KANO [23], and MGSSL [28], have been extensively evaluated on these datasets using consistent splits and protocols, establishing a clear basis for comparative studies. Second, the diversity of molecular tasks in MoleculeNet enables a comprehensive assessment of model generalization across different property types, including physicochemical, biological, and pharmacokinetic domains. This diversity aligns with our goal of validating whether motif-level contrastive learning can uncover transferable substructures relevant to distinct molecular functions. Third, MoleculeNet provides standardized scaffold splits and evaluation metrics, facilitating fair comparisons and reproducible experiments. The scaffold split, in particular, is well-suited for testing the ability of models to generalize beyond known chemical scaffolds, which is an essential consideration for real-world molecular discovery tasks. Finally, as our work builds on the self-supervised pretraining paradigm, it is critical to benchmark against datasets and splits adopted in prior contrastive learning frameworks. It ensures that observed improvements can be attributed to our architectural contributions (such as motif-level alignment and diffusion-based pretraining) rather than differences in dataset selection or experimental design.

Detailed descriptions of the small molecule classification datasets are provided below:

- BBBP [49] captures data on whether compounds have the ability to cross the blood-brain barrier, indicating their permeability properties.
- Tox21 [50] is a publicly accessible database that assesses compound toxicity, utilized notably in the 2014 Tox21 Data Challenge.
- ToxCast [51] provides a wealth of toxicity labels derived from high-throughput screening tests conducted on thousands of chemicals.

- SIDER [52] contains information on marketed drugs along with their associated adverse drug reactions, documented in the Side Effect Resource.
- ClinTox [53] compares drugs that have been approved by the FDA with those that were withdrawn during clinical trials due to toxicity concerns.
- BACE [54] compiles data on compounds that have been identified as inhibitors of human β-secretase (BACE-1) over recent years.
- MUV [55] is a curated subset of PubChem BioAssay, designed using a refined nearest neighbor analysis to validate virtual screening techniques.
- HIV [56] records the experimentally measured efficacy of over 40,000 molecules in inhibiting HIV replication.

Detailed descriptions of small molecule regression datasets are provided below:

- ESOL [57] is a compact dataset that records the solubility of various compounds.
- FreeSolv [58] is derived from the Free Solvation Database, documenting the hydration free energy of small molecules in water, based on both experimental data and alchemical free energy calculations.
- Lipophilicity [59] is extracted from the ChEMBL database[59], containing experimental results of the octanol-water partition coefficient, an indicator of molecular solubility.
- QM7 [60] is a subset of the GDB-13 database, providing data on the spatial structures of molecules. It includes various electronic properties that are stable and synthetically accessible, such as HOMO and LUMO levels, calculated using ab-initio density functional theory (DFT), along with atomization energies.
- QM8 [61] applies multiple quantum mechanics methodologies to determine the electronic spectra and excited state energies of small molecules.
- QM9 [62] offers comprehensive data on the geometry, energy, electronic, and thermodynamic properties of small molecules, all calculated using DFT.

As detailed in Supplementary Table 1, we employ scaffold splitting [66] for all datasets except QM9. Scaffold splitting divides molecules based on their distinct two-dimensional structural frameworks, creating subsets that provide a more challenging and realistic scenario since the molecules in the testing set are structurally different from those in the training set. For QM9, we follow standard practice by applying random splitting, consistent with the protocols in related studies [22, 23, 27]. For the above benchmark experiments, we trained MotiL using three independent runs for each dataset, with different random seeds for initialization. The same dataset splits and model configurations were used across runs to ensure comparability. We report the average performance and the unbiased standard deviation across the three runs to account for variability arising from stochastic optimization and data sampling.

In addition, we use two protein datasets for the protein macromolecular property prediction, as detailed in Supplementary Table 2:

- Gene ontology (GO): This task involves predicting the functions of proteins based on multiple GO terms, making it a multi-label classification challenge. In accordance with ref. [63], proteins are categorized into hierarchically related functional classes within three ontologies: biological process (BP) with 1,943 classes, molecular function (MF) with 489 classes, and cellular component (CC) with 320 classes. The dataset comprises 29,898 proteins for training, 3,322 for validation, and 3,415 for testing.
- Enzyme commission (EC): This task focuses on predicting the three- and four-level Enzyme Commission (EC) numbers across 538 classes. We adopt the training, validation, and testing splits from ref. [63], which include 15,550 proteins for training, 1,729 for validation, and 1,919 for testing. The prediction of EC numbers is also framed as a multi-label classification task.

C Evaluation metrics

For the classification task on small molecules, we use the area under the receiver operating characteristic curve (ROC-AUC) as the evaluation metric. ROC-AUC is a performance measurement for classification

problems at various threshold settings. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at different threshold values. The area under the curve (AUC) represents the degree or measure of separability achieved by the model. It quantifies the overall ability of the model to distinguish between classes. An ROC-AUC value of 1.0 indicates perfect classification, while an ROC-AUC value of 0.5 suggests the model's performance is equivalent to random guessing, i.e., it has no discriminative power.

For the regression task on small molecules, we use root mean square error (RMSE) and mean absolute error (MAE) as evaluation metrics. RMSE is a commonly used metric to measure the differences between values predicted by a model and the values actually observed. It provides a way to quantify the accuracy of continuous predictions. RMSE is the square root of the average of the squared differences between predicted and actual values. Mathematically, it is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n is the total number of observations. Lower RMSE values indicate better model performance. MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It is a straightforward metric that provides an interpretation of the average error between predicted and actual values. MAE is defined as the mean of the absolute differences between predicted and actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{2}$$

MAE indicates how wrong predictions typically are, with lower values indicating more accurate predictions.

For the macromolecule task, we use $\text{F}_{\max}$ as evaluation metrics. $\text{F}_{\max}$ is used to evaluate the balance between precision and recall in multi-label classification tasks, particularly in the context of protein function prediction [63]. It provides a comprehensive measure of a model's performance across different decision thresholds. Specifically, the F-score is calculated based on the precision and recall for each protein at a given threshold λ. The precision for protein i at threshold λ is defined as:

$$\text{Precision}_i(\lambda) = \frac{\sum_j \left((p_i^j \geq \lambda) \cap b_i^j\right)}{\sum_j (p_i^j \geq \lambda)} \tag{3}$$

and the recall for protein i at threshold λ is given by:

$$\text{Recall}_i(\lambda) = \frac{\sum_j (p_i^j \geq \lambda)}{\sum_j b_i^j}. \tag{4}$$

Here, p_i^j represents the predicted score for the j-th function of protein i, and b_i^j is the binary indicator of the actual presence of that function. The average precision and recall across all proteins are then calculated as:

$$\text{Precision}(\lambda) = \frac{\sum_{i=1}^{N}\text{Precision}_i(\lambda)}{\sum_{i=1}^{N}\left(\sum_j (p_i^j \geq \lambda) \geq 1\right)}, \quad \text{Recall}(\lambda) = \frac{\sum_{i=1}^{N}\text{Recall}_i(\lambda)}{N} \tag{5}$$

where N is the total number of proteins. Finally, $\text{F}_{\max}$ is defined as the maximum F-score over all tested thresholds λ within the range [0, 1]:

$$\text{F}_{\max} = \max_{\lambda \in [0,1]} \left(\frac{2 \times \text{Precision}(\lambda) \times \text{Recall}(\lambda)}{\text{Precision}(\lambda) + \text{Recall}(\lambda)}\right). \tag{6}$$

The metric of F_{max} captures the best trade-off between precision and recall across different thresholds, providing a robust measure of model performance.

Supplementary Table 3 ROC-AUC (%) performance comparison on small molecules. Results are presented as mean ± standard deviation over three independent runs. Statistical significance was evaluated using a two-sided t-test. Exact *p*-values are reported in the last row. No adjustments were made for multiple comparisons.

Category	**Physiology**					**Biophysics**		
Metric	ROC-AUC (%) ↑					ROC-AUC (%) ↑		
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	MUV	HIV
# molecules	2,039	7,831	8,575	1,427	1,478	1,513	93,807	41,127
# tasks	1	12	617	27	2	1	17	1
GCN [38]	71.8±0.9	70.9±0.3	65.0±6.1	53.6±0.3	62.5±2.8	71.6±2.0	71.6±4.0	74.0±3.0
GIN [67]	65.8±4.5	74.0±0.8	66.7±1.5	57.3±1.6	58.0±4.4	70.1±5.4	71.8±2.5	75.3±1.9
MPNN [24]	91.3±4.1	80.8±2.4	69.1±3.0	59.5±3.0	87.9±5.4	81.5±1.0	75.7±1.3	77.0±1.4
DMPNN [25]	91.9±3.0	75.9±0.7	63.7±0.2	57.0±0.7	90.6±0.6	85.2±0.6	78.6±1.4	77.1±0.5
CMPNN [26]	95.7±2.8	80.4±0.3	70.4±0.8	63.4±2.0	87.6±0.8	85.6±3.5	76.4±3.9	82.5±2.3
N-GRAM [27]	91.2±0.3	76.9±2.7	-	63.2±0.5	87.5±2.7	79.1±1.3	76.9±0.7	78.7±0.4
Hu et. al [68]	70.8±1.5	78.7±0.4	65.7±0.6	62.7±0.8	72.6±1.5	84.5±0.7	81.3±2.1	79.9±0.7
MGSSL [28]	70.5±1.1	76.4±0.4	64.1±0.7	61.8±0.8	80.7±2.1	79.7±0.8	78.7±1.5	79.5±1.1
GEM [21]	72.4±0.4	78.1±0.1	69.2±0.4	63.2±0.4	90.1±1.3	85.6±1.1	81.7±0.5	80.6±0.9
GROVER [65]	86.8±2.2	80.3±2.0	56.8±3.4	61.2±2.5	70.3±13.7	82.4±3.6	67.3±1.8	68.2±1.1
GraphMVP [29]	72.4±1.6	75.9±0.5	63.1±0.4	63.9±1.2	79.1±2.8	81.2±0.9	77.7±0.6	77.0±1.2
MolCLR [22]	73.3±1.0	74.1±5.3	65.9±2.1	61.2±3.6	89.8±2.7	82.8±0.7	78.9±2.3	77.4±0.6
$\text{MolCLR}_{\text{CMPNN}}$ [22]	72.4±0.7	78.4±2.6	69.1±1.2	59.7±3.4	88.0±4.0	85.0±2.4	74.5±2.1	77.8±5.5
KANO [23]	96.0±1.6	*83.7±1.3*	*73.2±1.6*	*65.2±0.8*	*94.4±0.3*	*93.1±2.1*	*83.7±2.3*	*85.1±2.2*
MotiL	**96.9±0.3**	**84.0±0.2**	**73.5±0.6**	**65.6±0.9**	**95.2±0.8**	**93.9±1.0**	**85.0±1.0**	**85.6±0.2**
p-value	1.2e-4	1.1e-7	5.9e-6	1.9e-4	2.8e-4	6.6e-7	3.5e-7	2.7e-6

Supplementary Table 4 Performance comparison on small molecules using RMSE or MAE, where lower values indicate better performance. Results are presented as mean ± standard deviation over three independent runs. Statistical significance was evaluated using a two-sided t-test. Exact *p*-values are reported in the last row. No adjustments were made for multiple comparisons.

Category	**Physical chemistry**			**Quantum mechanics**		
Metric	RSME ↓			MAE ↓		
Dataset	ESOL	FreeSolv	Lipophilicity	QM7	QM8	QM9
# molecules	1,128	642	4,200	7,160	21,786	133,885
# tasks	1	1	1	1	12	3
GCN [38]	1.431±0.050	2.870±0.135	0.712±0.049	122.9±2.2	0.0366±0.000	0.00835±0.00001
GIN [67]	1.452±0.020	2.765±0.180	0.850±0.071	124.8±0.7	0.0371±0.001	0.00824±0.00004
MPNN [24]	1.167±0.430	1.621±0.952	0.672±0.051	111.4±0.9	0.0148±0.001	0.00522±0.00003
DMPNN [25]	1.050±0.008	1.673±0.082	0.683±0.016	103.5±8.6	0.0156±0.001	0.00514±0.00001
CMPNN [26]	0.798±0.112	1.570±0.442	0.614±0.029	75.1±3.1	0.0153±0.002	0.00405±0.00002
N-GRAM [27]	1.100±0.030	2.510±0.191	0.880±0.121	125.6±1.5	0.0320±0.003	0.00964±0.00031
Hu et.al [68]	1.100±0.006	2.764±0.002	0.739±0.003	113.2±0.6	0.0215±0.001	0.00922±0.00004
GEM [21]	0.813±0.028	1.748±0.114	0.674±0.022	60.0±2.7	0.0163±0.001	0.00562±0.00007
GROVER [65]	1.423±0.288	2.947±0.615	0.823±0.010	91.3±1.9	0.0182±0.001	0.00719±0.00208
MolCLR [22]	1.113±0.023	2.301±0.247	0.789±0.009	90.9±1.7	0.0185±0.013	0.00480±0.00003
$\text{MolCLR}_{\text{CMPNN}}$ [22]	0.911±0.082	2.021±0.133	0.875±0.003	89.8±6.3	0.0179±0.001	0.00475±0.00001
KANO [23]	*0.670±0.019*	*1.142±0.258*	*0.566±0.007*	*56.4±2.8*	*0.0123±0.000*	*0.00320±0.00001*
MotiL	**0.644±0.014**	**1.097±0.001**	**0.548±0.000**	**50.7±1.0**	**0.0117±0.000**	**0.00299±0.00001**
p-value	5.3e-7	2.0e-4	5.1e-7	9.2e-5	3.9e-6	4.0e-6

D Complete performance comparison on small molecules and macromolecules

Supplementary Tables 3 and 4 provide a comprehensive performance comparison between MotiL and baselines on 14 small molecule datasets, including the mean and standard deviation after running three

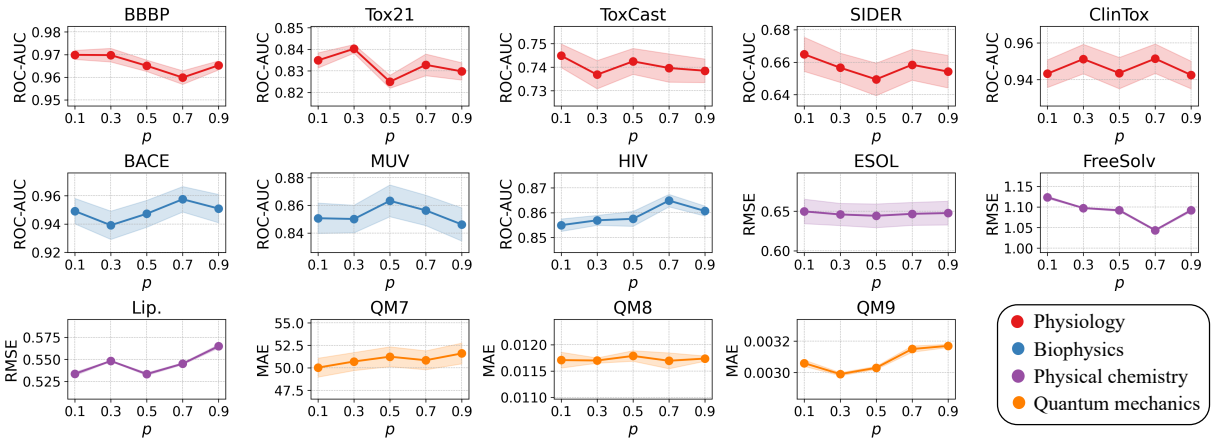
Supplementary Table 5 Performance comparison on protein macromolecules. All methods, including baselines and MotiL, use identical data splits: 29,898/3,322/3,415 proteins for training, validation, and test on the GO dataset, and 15,550/1,729/1,919 proteins on the EC dataset. Following the multi-cutoff split strategy in [63], the test set contains only PDB chains with sequence identity no more than 95% to those in the training set, ensuring biologically meaningful generalization evaluation. Results are reported as mean ± standard deviation over three independent runs with different random seeds. Statistical significance was evaluated using a two-sided t-test. Exact p-values are reported in the last row. No adjustments were made for multiple comparisons.

Dataset	**GO-BP**	**GO-MF**	**GO-CC**	**EC**
Metric		$F_{max}(\%)\uparrow$		
# proteins		36,635		19,198
CNN [36]	26.3±0.5	23.0±0.6	33.7±1.7	35.9±2.2
ResNet [37]	26.4±0.6	22.4±0.7	33.3±2.0	19.4±0.7
LSTM [37]	24.2±0.3	23.6±0.4	31.3±0.8	21.2±0.5
Transformer [37]	24.9±1.4	22.9±0.9	35.2±4.0	20.0±0.3
GCN [38]	27.2±0.3	23.7±0.1	38.5±0.5	31.6±1.0
GAT [39]	26.0±0.3	21.8±0.5	37.2±0.6	11.0±2.9
3D CNN [40]	25.9±0.2	21.0±1.1	32.5±0.8	23.4±2.7
GraphQA [41]	25.9±0.2	23.8±0.3	34.1±0.1	16.3±1.0
GVP [42]	26.5±0.3	20.7±0.1	33.4±2.2	16.9±1.2
$\text{IEConv}_{\text{residue level}}$ [43]	29.5±1.0	38.6±0.8	36.7±0.8	45.5±1.4
GearNet [44]	31.3±0.6	51.5±0.6	34.6±0.6	64.0±1.6
$\text{GearNet}_{\text{IEConv}}$ [44]	34.6±0.5	52.4±0.6	37.4±0.5	64.6±0.7
$\text{GearNet}_{\text{Edge}}$ [44]	35.5±0.5	53.2±0.3	37.5±0.4	66.2±0.8
$\text{GearNet}_{\text{Edge-IEConv}}$ [44]	36.2±0.5	55.0±0.8	41.0±0.6	69.3±0.5
CDConv [45]	*42.2*±0.7	*61.8*±0.5	*42.0*±0.3	*86.1*±0.4
MotiL	**44.4**±0.2	**64.0**±0.3	**48.9**±0.3	**87.9**±0.3
p-value	4.5e-2	4.7e-3	5.2e-4	3.7e-2

independent experiments. Based on the average performance, we can see that MotiL consistently outperforms all baseline methods. On certain datasets, MotiL markedly outperforms the second-best method. For example, on the MUV dataset, MotiL achieves a 1.6% higher ROC-AUC relative to the runner-up KANO, and on the QM7 dataset, it shows a substantial improvement of 10.1% over the runner-up KANO. To assess whether the observed performance gains of MotiL over the strongest baseline (KANO) are statistically significant, we performed paired two-tailed t-tests across all MoleculeNet benchmark datasets using the results from three independent runs. The corresponding p-values are summarized in Supplementary Tables 3 and 4. All values are below the 0.05 threshold, confirming that MotiL's improvements are statistically significant and robust across different datasets. Supplementary Table 5 presents a comprehensive performance comparison between MotiL and the baselines on macromolecules. In our macromolecular experiments, we follow the established dataset split used in prior protein property prediction studies [44, 45, 63]: the data are partitioned into training, validation, and test sets in an approximately 8:1:1 ratio, with the critical constraint that all test-set PDB chains share $\leq 95\%$ sequence identity with any training example. This split ensures a biologically meaningful assessment of generalization. Standard k-fold cross-validation undermines this constraint required for fair evaluation because in the repeated folding process, highly similar sequences inevitably end up in both the training and validation sets in at least one fold. Therefore, we instead fix the split exactly as in previous work [44, 45, 63] and evaluate model robustness via three independent runs with different random seeds, reporting mean ± standard deviation across runs. This protocol enables fair comparison while accounting for stochasticity in training (e.g., random data augmentation on the protein's 3D coordinates, mini-batch shuffling, and dropout). The results shown in Supplementary Table 5 demonstrate that MotiL significantly outperforms the best baseline, CDConv, on the GO and EC datasets.

E Sensitivity analysis

We conduct the sensitivity analysis to understand the impact of the retaining probability p in network pruning on the MotiL's performance. The retaining probability p is a key parameter that controls the ratio of network pruning on GNN: smaller values of p lead to more pruning, whereas larger values retain a



Supplementary Figure 1 Performance of MotiL with different retaining probability p over $n = 3$ independent experiments. Source data are provided as a Source Data file.

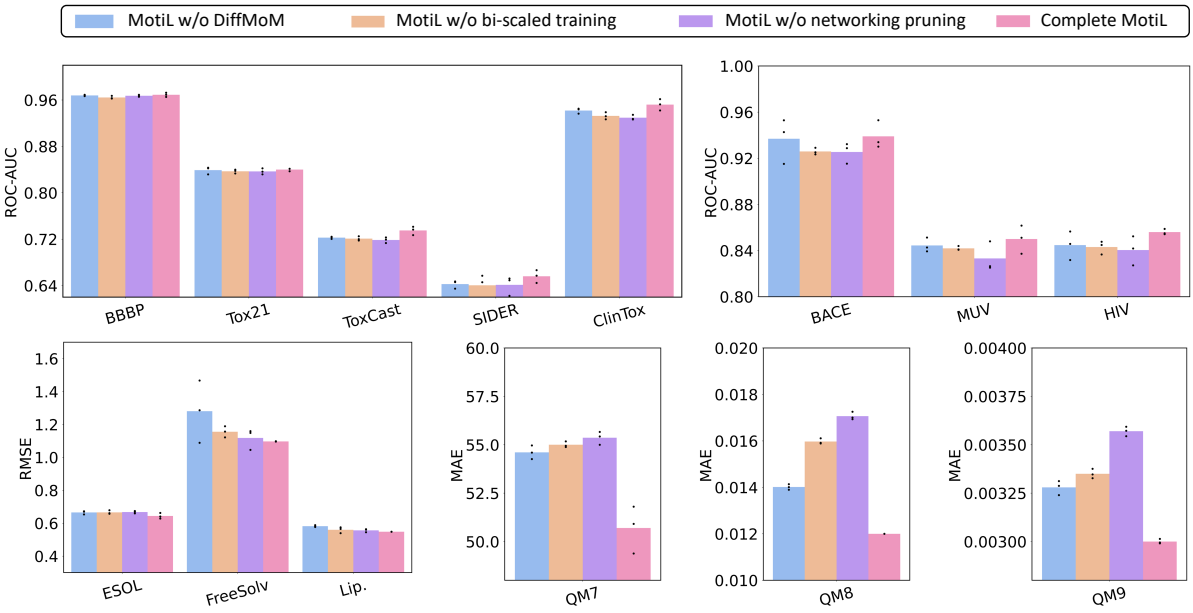
higher proportion of the neural network structure. We evaluate the performance of MotiL across different values of p set to 0.1, 0.3, 0.5, 0.7, and 0.9. As expected, a lower p indicates greater divergence between the representations obtained from the two encoders, thereby increasing the difficulty of bi-scaled training. Conversely, a higher p leads to more similar representations, reducing the learning space for GNN and making it easier to converge but potentially limiting the richness of the learned features. The results are shown in Supplementary Fig. 1. Based on these results, the setting of $p = 0.7$ performs well across most datasets, with particularly strong performance on the ClinTox, BACE, HIV, and FreeSolv datasets. Therefore, we recommend setting $p = 0.7$ in the experiments.

F Ablation study

We conduct an ablation study on MotiL by designing three variant methods: 1) MotiL without DiffMoM, 2) MotiL without bi-scaled training, and 3) MotiL without network pruning. The results of the complete MotiL and its variants are shown in Supplementary Fig. 2. The following conclusions can be drawn from the figure:

- The complete MotiL method performs best across all datasets and outperforms the variant methods markedly on some datasets, such as MUV, QM7, and QM9. It indicates that each of the three components contributes positively to the performance of MotiL.
- Among the three components, DiffMoM has the greatest impact. For instance, on the ClinTox, FreeSolv, and QM7 datasets, the removal of DiffMoM leads to a noticeable decline in performance. It suggests that the DiffMoM model in the diffusion priming phase plays a crucial role in enhancing the GNN's robustness to noise perturbations, which is essential for the MotiL's classification or regression performance in downstream tasks.
- The impact of bi-scaled training on MotiL's performance is moderate. Particularly, on the ToxCast, SIDER, and MUV datasets, MotiL without bi-scaled training consistently underperforms. These datasets have a large number of classification tasks, with ToxCast containing 617 tasks, SIDER 27 tasks, and MUV 17 tasks. Functional groups are critical substructures within molecules, and the more tasks a dataset contains, the more important these fine-grained substructures become. Therefore, bi-scaled training, which includes both graph-level and motif-level training, has a significant impact on those datasets with a large number of tasks.
- Network pruning has the smallest impact on MotiL's performance in all components. Removing network pruning reduces the number of positive molecule pairs in the bi-scaled training phase, which weakens

the capability of bi-scaled training to capture the relationship between similar molecules, and thus lowers the MotiL's performance in downstream tasks.



Supplementary Figure 2 Performance of MotiL and its three variants over $n = 3$ independent experiments. Source data are provided as a Source Data file.

G Robustness evaluation to graph structural noise

To evaluate the robustness of molecular representation models under structural noise, we introduce varying levels of edge deletion to the molecular graphs in the BBBP and Solubility datasets. Supplementary Fig. 8 shows the ROC-AUC or RMSE results for MotiL, MotiL w/o diffusion priming, and three representative baselines (MolCLR, MGSSL, and KANO). MotiL exhibits significantly better resistance to noise perturbations, with a flatter error curve and superior RMSE across all noise levels. It demonstrates the effectiveness of the diffusion priming phase in enhancing the stability of learned representations under noisy graph conditions.

H Case analysis: t-SNE visualization of micromolecular representations

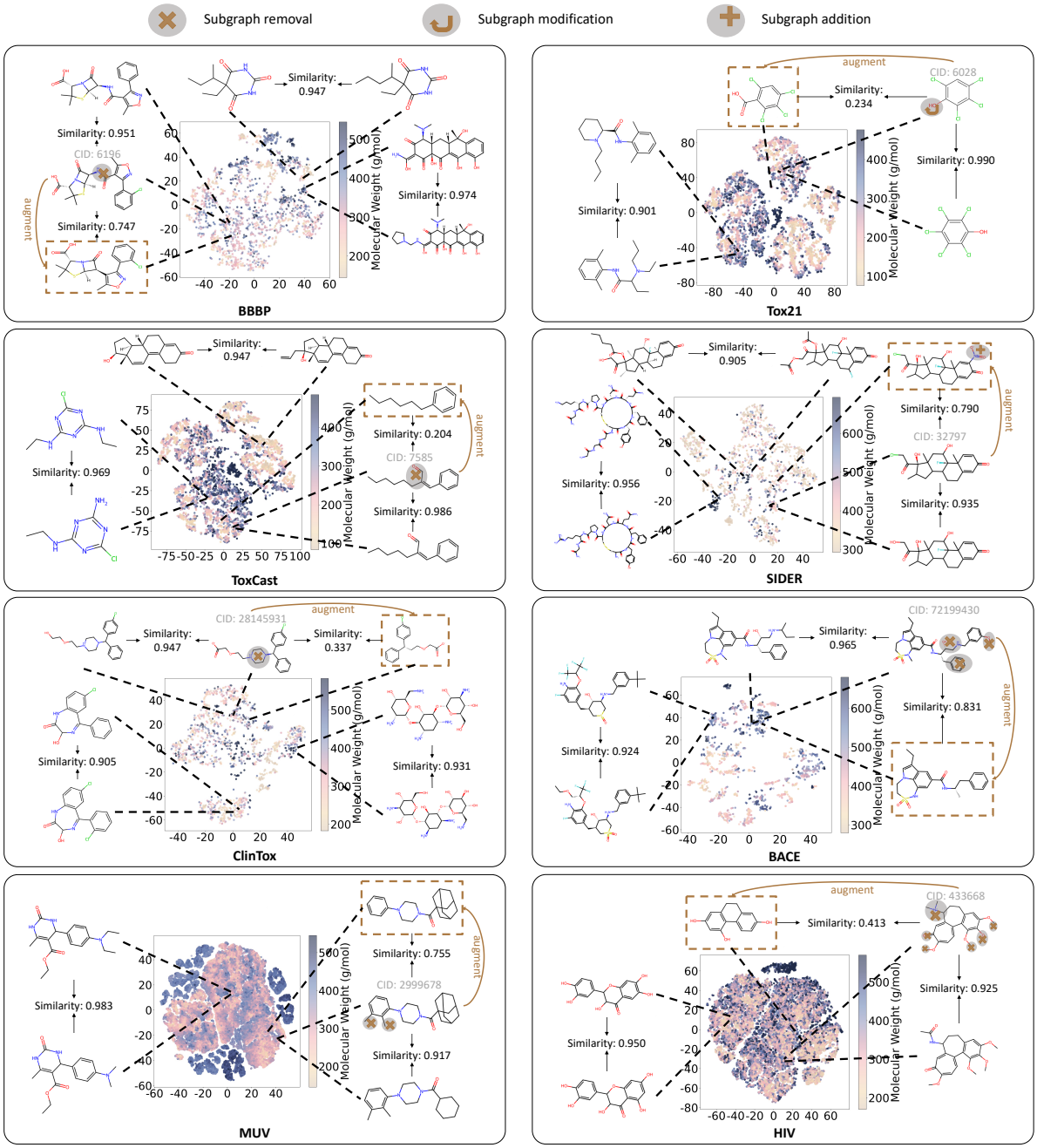
In the analysis of small molecule representations, we examine whether the molecular representations learned by MotiL are consistent with the molecular structures across different datasets. We use the t-SNE technique [69] to visualize these molecular representations in a two-dimensional space, as shown in Supplementary Fig. 3. In the plotted figures, black squares are the randomly selected positive molecule pairs, while brown squares represent the augmented molecules generated through molecular graph augmentation, serving as negative molecules for comparison. For the specific molecular graph augmentation strategies, we use subgraph modification for Tox 21, subgraph addition for SIDER, and subgraph removal for the other datasets. By analysis, we derive the following conclusions:

- Across all datasets, the RDKFP similarity scores between the displayed positive molecule pairs are consistently high, and their representations are located close to each other in the space. For example, on Tox21, the RDKFP similarity score between the molecule with CID 6028 and its positive molecule is 0.990. Similarly, on BACE, the RDKFP similarity between the molecule with CID 72199430 and its positive molecule is 0.965. It indicates that the molecular graph representations learned by MotiL are consistent with the structural characteristics of these molecules, effectively capturing the structural similarity between positive molecule pairs. Although the bi-scaled training phase in MotiL does not explicitly construct positive molecule pairs, the diffusion priming phase ensures that MotiL remains robust to the noise perturbations in the molecular graph, allowing it to still identify which molecules are similar.
- The RDKFP similarity score between the displayed negative molecule pairs is significantly lower than that between positive ones, with some showing very low similarity. For instance, the negative molecule pairs generated on the Tox21 and ToxCast datasets have similarity scores below 0.25. Correspondingly, the representations of the negative molecule pairs generated by MotiL are much farther apart in the representation space than those of the positive ones. It indicates that, compared to current MCL methods that rely on molecular graph augmentation, the augmentation-free mechanism designed in MotiL achieves significant success in extracting molecular representations that are consistent with the molecular structure.

I Expanded discussion on benchmark limitations and real-world ADME evaluation

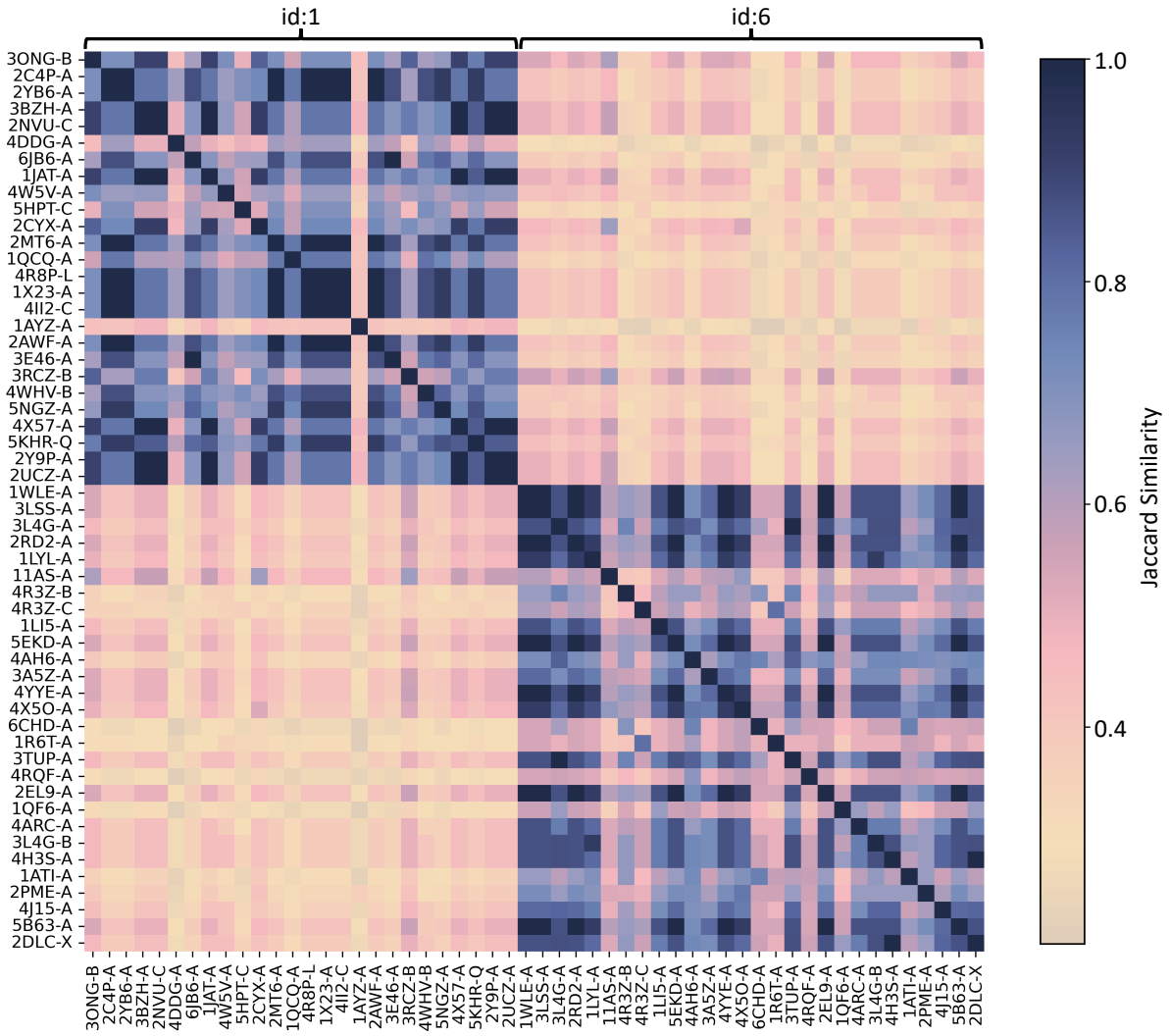
I.1 Limitations of MoleculeNet benchmarks

While MoleculeNet remains the most widely adopted benchmark suite for evaluating molecular property prediction models, its use is not without challenges. Previous studies and domain experts have highlighted issues such as label noise, duplicate molecules, and structural inconsistencies (e.g., protonation states, tautomers, stereochemistry) within its datasets [72]. These limitations can affect the reliability of comparative evaluation, especially for tasks requiring high-fidelity chemical understanding. Nevertheless,



Supplementary Figure 3 Visualization of micromolecular representations learned by MotiL via t-SNE on multiple datasets.

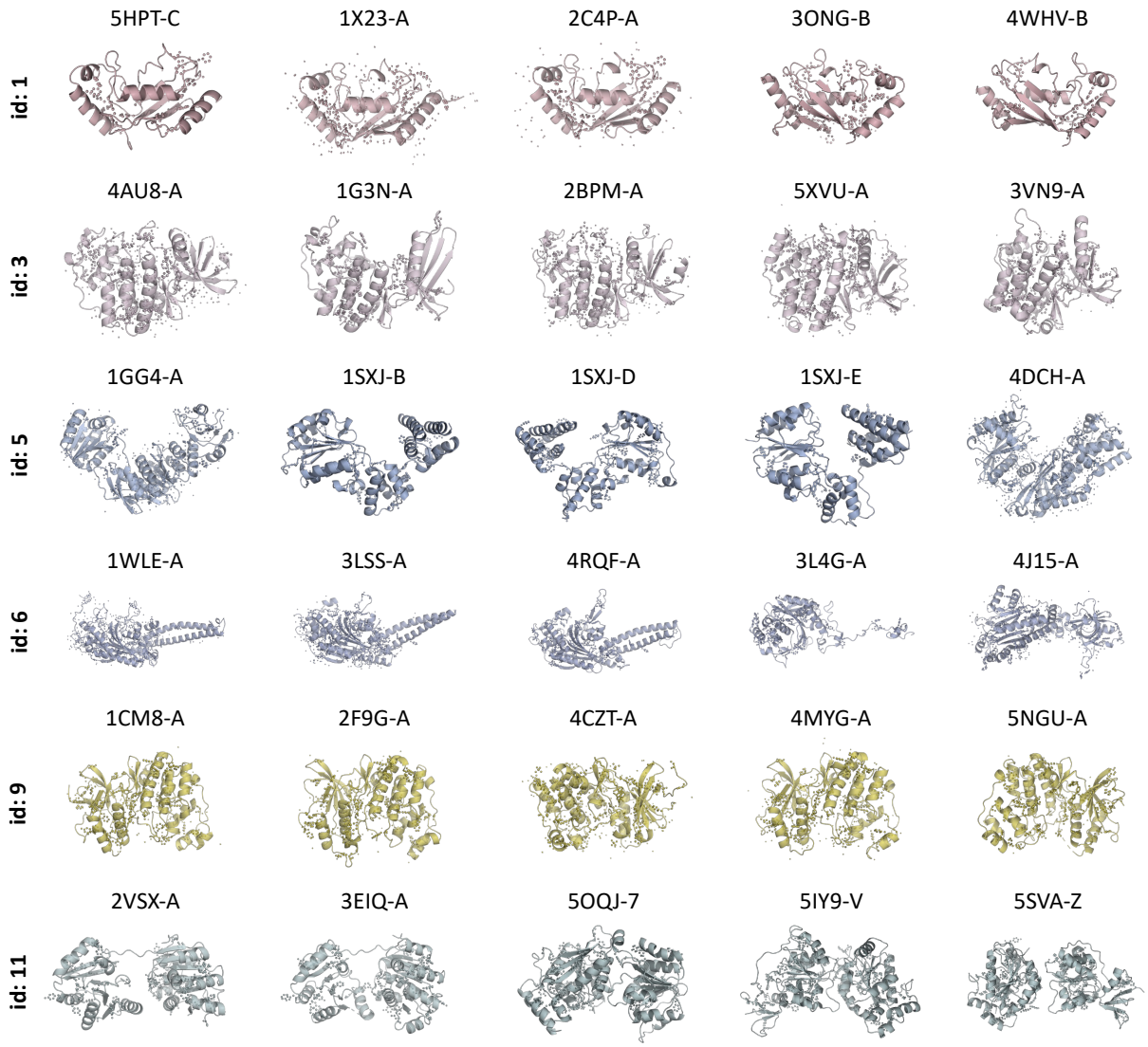
due to its established baselines and broad community adoption, MoleculeNet provides a useful foundation for measuring progress in molecular machine learning under standardized settings. To maintain comparability with prior work such as GROVER, MolCLR, and KANO, we used the unmodified MoleculeNet datasets, adhering to the same data splits and evaluation protocols. This enables fair head-to-head comparisons while acknowledging the benchmark's inherent constraints.



Supplementary Figure 4 Heatmap of the Jaccard similarity among the proteins' GO terms in Clusters 1 and 6. Source data are provided as a Source Data file.

I.2 Motivation for evaluating on ADME datasets

To move beyond synthetic benchmarks and toward more application-driven evaluation, we conducted additional experiments on six ADME-related datasets [73]. These datasets reflect core pharmacokinetic properties relevant to drug development, including metabolism (HLM, RLM), plasma protein binding (hPPB, rPPB), membrane permeability (MDR1-MDCK ER), and aqueous solubility. Unlike MoleculeNet tasks, these ADME endpoints are measured in realistic laboratory or clinical settings, often featuring complex structure-activity relationships and significant data heterogeneity. They serve as practical testbeds for evaluating a model's ability to generalize to pharmacologically meaningful and commercially relevant tasks.

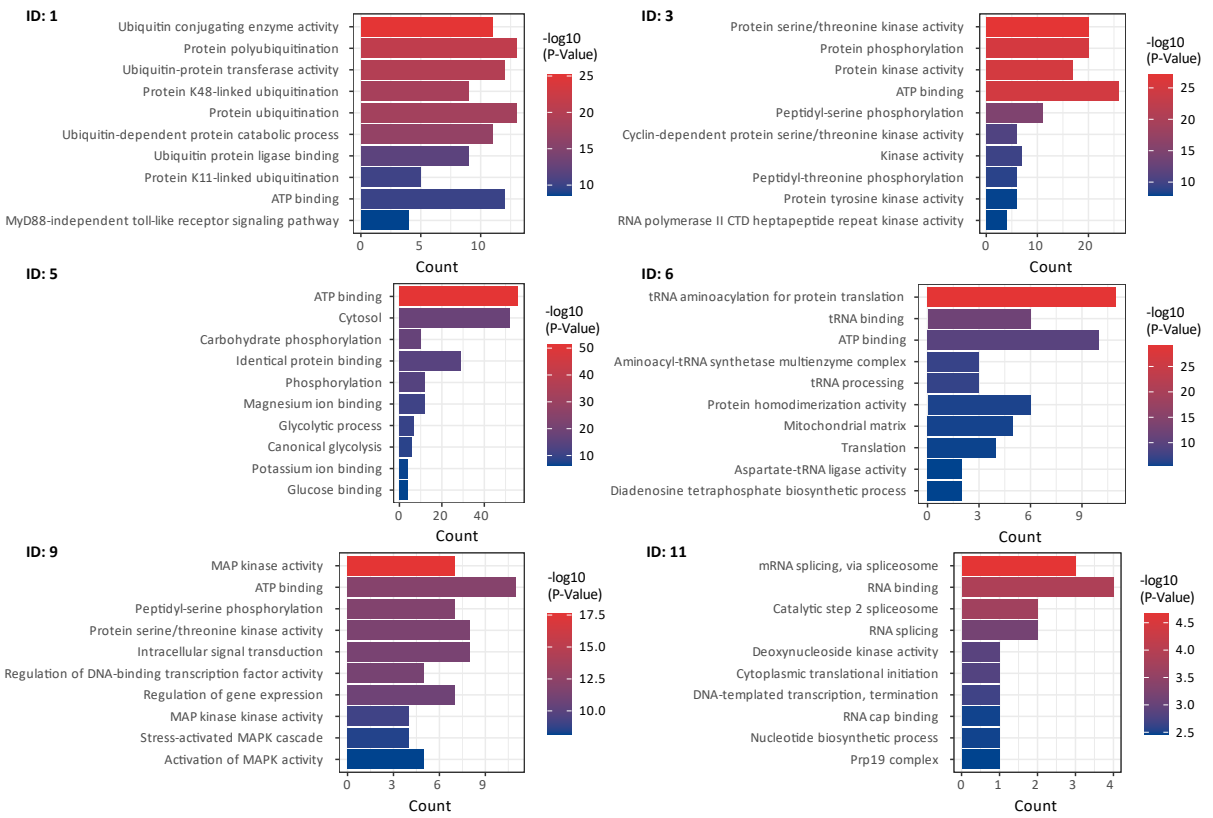


Supplementary Figure 5 3D structure comparison of the proteins in six clusters obtained by UMAP [70] and DBSCAN [71].

I.3 Comparative analysis and discussion

We report the performance of MotiL and four baseline methods (CMPNN, MGSSL, MolCLR, and KANO) on the six ADME prediction tasks, with results summarized in Supplementary Table 6. All metrics are averaged over three independent runs, and standard deviations are reported to assess stability. All baseline models are implemented using publicly available code with their default hyperparameter configurations.

The results demonstrate that MotiL achieves highly stable and competitive performance. While MolCLR and MGSSL achieve high scores on certain tasks (e.g., hPPB and MDR1-MDCK ER), their results often come with larger variances, suggesting potential inconsistency. In contrast, MotiL exhibits low standard deviations across all tasks (e.g., ± 0.0054 for HLM and ± 0.0047 for Solubility), indicating superior robustness. Although the absolute performance gap between MotiL and top-scoring models may

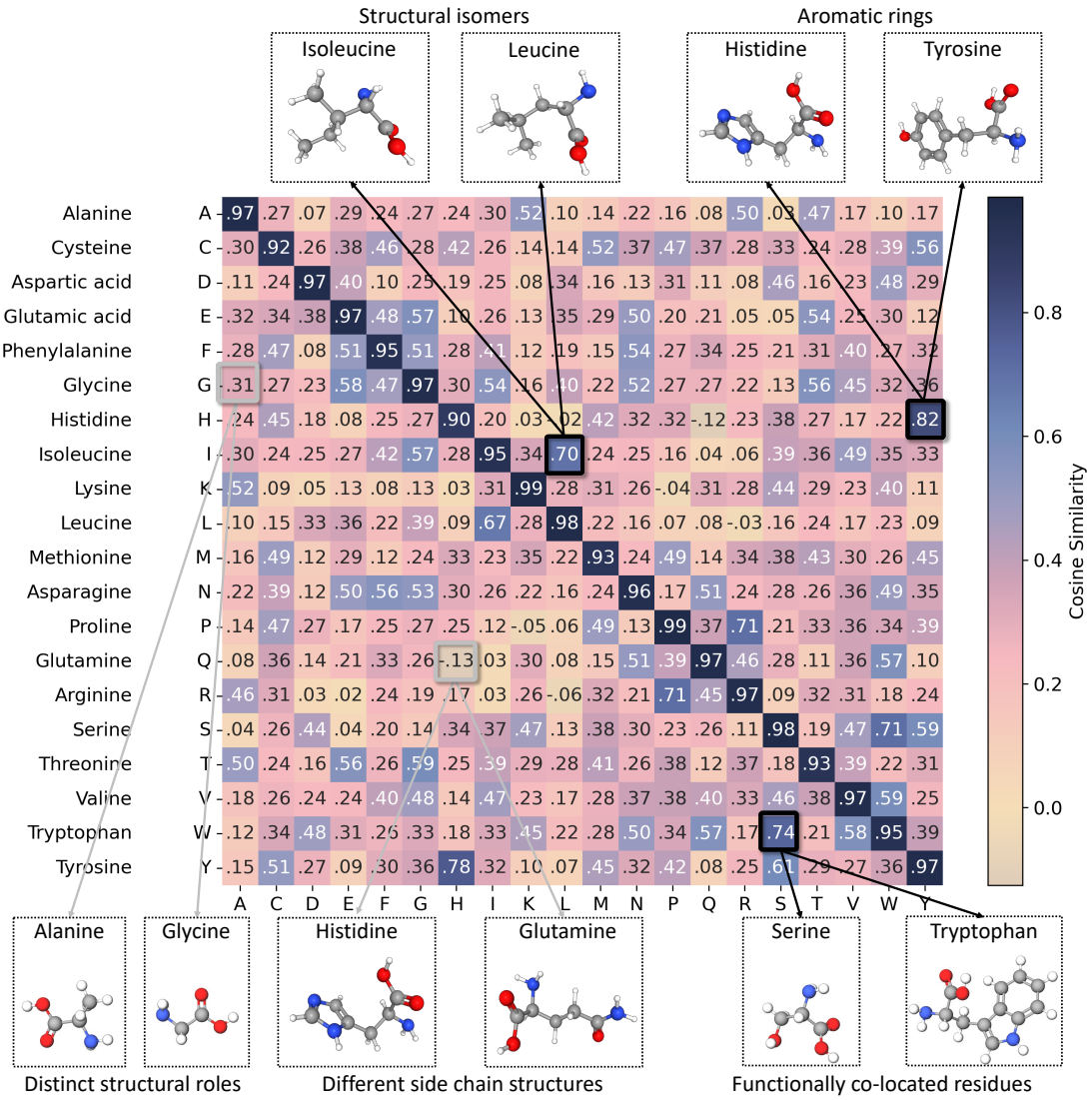


Supplementary Figure 6 Top 10 GO terms within clusters 1, 3, 4, 6, 9, and 11. Source data are provided as a Source Data file.

Supplementary Table 6 Model performance on ADME datasets. Lower RMSE values indicate better performance for all datasets. Results are reported as mean ± standard deviation over three independent runs. Statistical significance was evaluated using a two-sided t-test. Exact p-values are reported in the last column. No adjustments were made for multiple comparisons.

Dataset	CMPNN [26]	MGSSL [28]	MolCLR [22]	KANO [23]	MotiL	p-value
HLM	0.5343±0.0226	0.5030±0.0101	0.5412±0.0108	*0.4849±0.0225*	**0.4647±0.0054**	3.3e-7
hPPB	0.7566±0.0636	0.9683±0.0194	1.1117±0.0222	*0.6835±0.1349*	**0.6408±0.0100**	5.9e-3
MDR1-MDCK ER	0.5553±0.0245	0.7015±0.0140	0.6863±0.0137	*0.4869±0.0556*	**0.4493±0.0154**	1.4e-6
RLM	0.5935±0.0314	*0.5354±0.0107*	0.5884±0.0118	0.6133±0.0538	**0.5070±0.0072**	2.0e-4
rPPB	0.9115±0.1599	0.9137±0.0183	1.0883±0.0218	*0.8955±0.0352*	**0.8404±0.0294**	9.3e-5
Solubility	*0.5447±0.0415*	0.5884±0.0088	0.6023±0.0080	0.5639±0.0588	**0.5164±0.0047**	7.3e-4

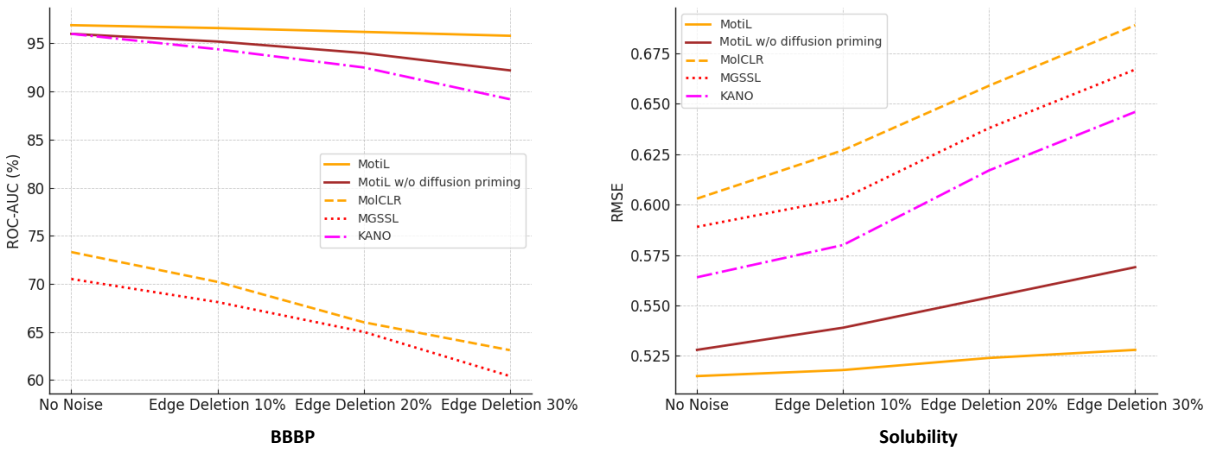
appear marginal on some tasks, MotiL consistently ranks among the best-performing methods with minimal fluctuation, highlighting its strong generalization ability. Statistical analysis further confirms that MotiL significantly outperforms the second-best method (e.g., KANO on the task of hPPB) on each task ($p < 0.05$), demonstrating its clear superiority. Moreover, MotiL's ability to handle diverse pharmacokinetic properties within a unified framework, without the need for task-specific tuning, emphasizes its practical utility for real-world applications in drug discovery.



Supplementary Figure 7 Heatmap of the amino acid's representation similarity. Source data are provided as a Source Data file.

I.4 Broader impact and positioning

By including the evaluation on real-world datasets and explicitly acknowledging the known limitations of earlier benchmarks, we aim to position MotiL as a bridge between foundational research in molecular representation learning and its applications in drug discovery pipelines. Our framework supports both small molecule and macromolecular tasks within a unified architecture and demonstrates improvements in consistency, interpretability, and practical relevance. We hope these additional experiments and clarifications provide greater context for our design choices and reinforce the broader significance of our contributions.



Supplementary Figure 8 Performance of different methods under increasing graph structural noise. Source data are provided as a Source Data file.

J Detailed results and analysis of protein representations

We apply UMAP [70] and DBSCAN clustering [71] to the adenosine triphosphate (ATP) binding proteins and select some cases to illustrate the results. As shown in Supplementary Fig. 4, we plot the names of each protein in Cluster 1 and Cluster 6 in the protein's GO terms similarity heatmap. Specifically, we calculate the Jaccard similarity among the proteins' GO terms. For example, taking proteins i and j as an illustration, if the GO term sets for i and j are respectively denoted as $\mathrm{GO}(i)$ and $\mathrm{GO}(j)$, then the Jaccard similarity score between them is given by $\frac{\mathrm{GO}(i)\cap\mathrm{GO}(j)}{\mathrm{GO}(i)\cup\mathrm{GO}(j)}$. We can see that the boundary between Cluster 1 and Cluster 6 is clear: within clusters, the similarity of GO terms is high (red areas), while between clusters, the similarity of GO terms is low (blue areas). It indicates that the learned representation of each protein is overall consistent with its molecular functions in Cluster 1 and Cluster 6.

To further verify whether the proteins' representations are consistent with their 3D structures, we randomly select six clusters detected by DBSCAN (Clusters 1, 3, 5, 6, 9, and 11) and visualize the proteins' 3D structures in these clusters in Supplementary Fig. 5. The visualization results reveal that proteins within the same cluster exhibit highly similar 3D structures, while the 3D structures of proteins from different clusters show significant differences. It indicates that the representations learned by MotiL effectively capture the structural information of proteins. Consequently, in the low-dimensional projection, proteins with similar structural features are clustered together, while those with different 3D structure characteristics are distributed in different clusters.

Furthermore, we conduct GO term analysis for the above six clusters. For each cluster, we identify the top 10 GO terms corresponding to the predominant biological themes. From Supplementary Fig. 6, we can see that clusters 1, 3, 5, 6, 9, and 11 focus on ubiquitination processes, phosphokinase activity, carbohydrate metabolism or glycolysis, tRNA regulation, MAP kinase activity, and RNA splicing, respectively. Taking cluster 9 as an example, the GO term analysis reveals a significant enrichment of proteins in the MAPK family. These proteins belong to the ATP-dependent serine/threonine kinases. Kinases hydrolyze ATP and transfer the terminal phosphate group from ATP to the hydroxyl group of amino acid residues in substrate proteins, thus performing phosphorylation modifications. The distinct feature of these kinases is their requirement for ATP binding. This uniform enrichment within the same family underscores the specific molecular function associated with each cluster. Therefore, the differentiation by the MotiL method in discerning these functional categories proves to be accurate.

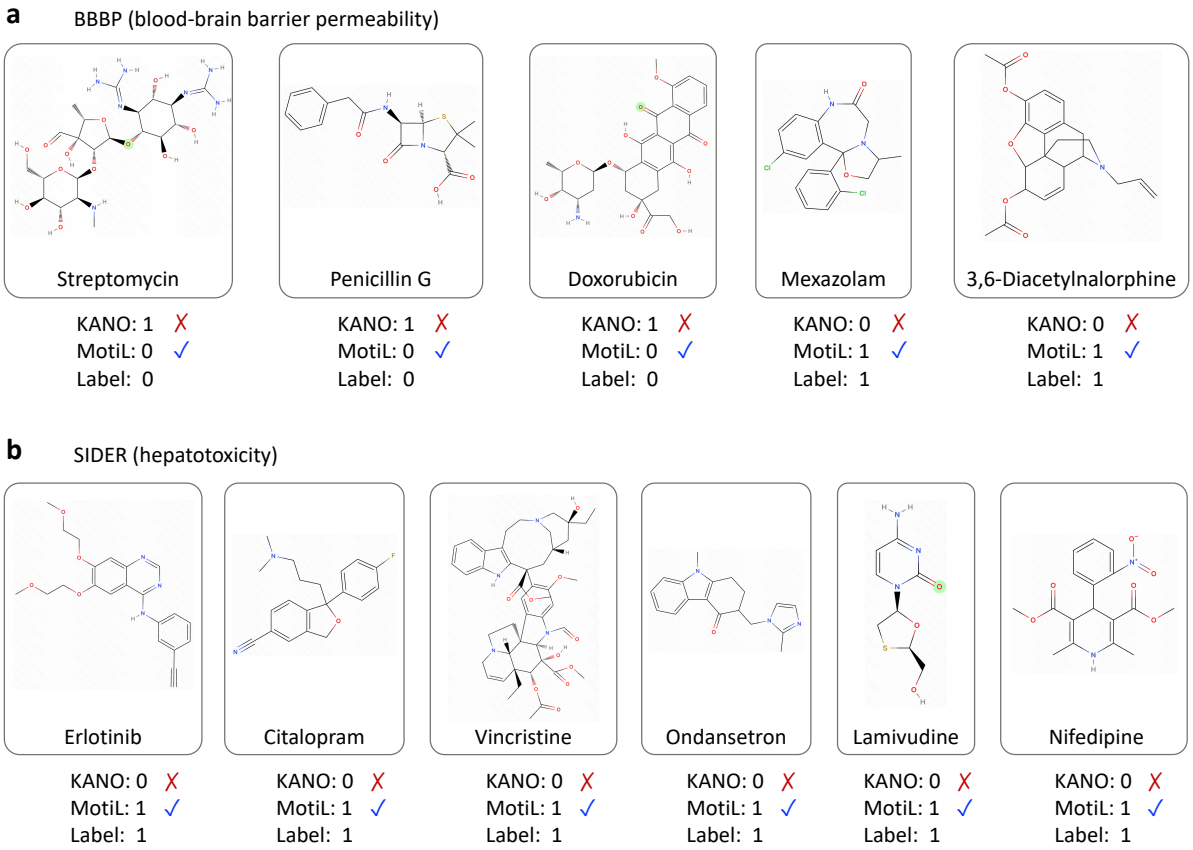
For the similarity among the representations of 20 types of amino acids, we plot the heatmap with the specific cosine similarity score on each cell in Supplementary Fig. 7. From the scores along the diagonal,

we can see that the similarity between representations of the same amino acid is very high, with all values being no less than 0.90. Also, the off-diagonal scores range within the closed interval [-0.13, 0.82]. For the off-diagonal scores, we analyze some examples of similarity scores between different amino acids. The high similarity scores are marked in black, while the low similarity scores are marked in gray.

- Example 1: The similarity between histidine and tyrosine is high, reaching 0.82. The reason for this is that both histidine and tyrosine contain aromatic rings. Although their types of aromatic rings differ, they both feature a planar structure with conjugated double bonds. Because this aromatic structure is rare among the 20 types of amino acids, histidine and tyrosine are structurally similar.
- Example 2: The similarity between isoleucine and leucine is also high, reaching 0.70. It is worth noting that these two amino acids are structural isomers, i.e., they share the same molecular formula ($C_6H_{13}NO_2$) but differ in the arrangement of their atoms. Despite this pair of structural isomers has identical molecular formulas, MotiL calculates their similarity as only 0.70, which is obviously lower than the diagonal values. It demonstrates that MotiL can capture structural differences effectively.
- Example 3: The similarity between histidine and glutamine is low, measuring -0.13. It can be attributed to the significant differences in their side chain structures. Histidine's side chain is an aromatic imidazole ring with conjugated double bonds, exhibiting complex chemical properties such as acid-base catalysis and electron delocalization. In contrast, glutamine's side chain is a simple amide group, which exhibits different physical and chemical properties from histidine. These structural differences lead to their different functions and interactions within proteins.
- Example 4: The similarity between serine and tryptophan is moderate, despite their apparent biochemical differences. Serine is a small polar residue with a hydroxyl-containing side chain, while tryptophan is a large, nonpolar aromatic amino acid. Although they differ substantially in physicochemical properties, MotiL assigns them a moderately high similarity. It can be explained by their co-occurrence in analogous structural or functional contexts, such as surface-exposed sites near active centers or domain boundaries, across the training dataset. Such context-dependent embedding behavior is characteristic of MotiL's self-supervised learning process, which encodes local neighborhood semantics rather than relying solely on static biochemical similarity.
- Example 5: The similarity between alanine and glycine is relatively low, despite their similar chemical nature. Both alanine and glycine have small, nonpolar side chains and are both weakly hydrophobic, thus they are often grouped together in conventional biochemical classifications. However, their learned similarity in MotiL is significantly lower, which reflects their distinct structural roles in proteins. Glycine, being the smallest amino acid with a hydrogen side chain, is frequently located in flexible regions like loops and turns. In contrast, alanine's methyl side chain favors α-helical structures and imparts rigidity. MotiL effectively captures these differences in structural context, leading to diverging embeddings even for chemically similar amino acids.

K MotiL corrects the erroneous prediction of the strongest baseline

KANO is the strongest baseline in the experiment of small molecular property prediction. We compare the prediction results of MotiL and KANO on BBBP and SIDER in Supplementary Fig. 9, where panel (a) presents the results on BBBP. In the displayed molecules, Streptomycin, Penicillin G, and Doxorubicin are antibiotics, while Mexazolam and 3,6-Diacetylnalorphine are chemotherapeutic drugs. Streptomycin, a highly polar aminoglycoside antibiotic with a molecular weight of 581 Da, surpasses the general upper limit for blood-brain barrier (BBB) penetration, which is typically around 500 Da. Its high polarity and large size hinder its ability to cross the lipid bilayer of the BBB without the support of specific transport systems. Penicillin G, a β-lactam antibiotic, also struggles to cross the BBB due to its polar groups, such as carboxylic acids, which reduce its diffusibility through the lipid-rich barrier and its stability in blood. Doxorubicin, though not excessively large with a molecular weight of 543.5 Da, contains multiple



Supplementary Figure 9 Representative cases comparing molecular property predictions by KANO [23] and MotiL. Each molecule is labeled with the predicted results from both models: 0 and 1 indicate predicted inability and ability to penetrate the blood-brain barrier or to induce hepatotoxicity, respectively. Predictions are marked as correct or incorrect based on ground-truth labels. a, BBBP prediction on small molecules. MotiL correctly predicts BBB permeability for central nervous system drugs, unlike KANO, which misclassifies polar compounds such as Streptomycin and Doxorubicin. **b,** SIDER hepatotoxicity prediction. MotiL better distinguishes toxic and non-toxic drugs, while KANO misclassifies clinically relevant examples like Erlotinib and Lamivudine.

polar functional groups (hydroxyl and ketone groups) that decrease its lipophilicity, limiting its ability to diffuse through the BBB. In contrast, drugs like Mexazolam and 3,6-Diacetylnalorphine exhibit characteristics conducive to BBB penetration. Mexazolam indicates for the treatment of anxiety with or without psychoneurotic conditions, which belongs to the benzodiazepine class (e.g., oxazolam and cloxazolam) and is structured to enhance lipophilicity, facilitating its passage through the BBB and other lipid membranes. Similarly, 3,6-Diacetylnalorphine, an acriflavine derivative, is a disinfectant bacteriostatic against many gram-positive bacteria. It has modifications such as acetyl groups that increase its hydrophobicity, enabling it to cross the BBB more readily. 3,6-Diacetylnalorphine is toxic and carcinogenic in mammals, so it is used only as a surface disinfectant or for treating superficial wounds. These structural adaptations highlight the importance of balancing lipophilicity and molecular size to achieve effective brain penetration. Finally, we calculate the prediction accuracy of KANO and MotiL on BBBP for blood-brain barrier permeability. The accuracy rates are 69.3% for KANO and 87.8% for MotiL, respectively.

In addition, Supplementary Fig. 9b displays the prediction results on SIDER. Erlotinib and Citalopram possess structures like quinones and fluorinated aromatics, which can form reactive metabolites through CYP450 enzyme actions, leading to oxidative stress in hepatocytes. Erlotinib also binds to the epidermal growth factor receptor tyrosine kinase in a reversible fashion at the ATP binding site of the

receptor. Vincristine, with its complex natural product structure including multiple methoxy groups and nitrogen-containing rings, can disrupt hepatocyte functions under stress. Ondansetron contains nitrogen heterocycles and sulphonamide groups and impacts liver enzyme activities, which should be attentioned for the treatment to the elderly, frail, and patients with moderate or high liver dysfunction. Lamivudine, though generally considered low in hepatotoxic potential, can still pose liver risks when used long-term or in combination therapies. Nifedipine features a nitro group in its dihydropyridine structure, which can undergo redox cycling, producing radicals and thus inducing oxidative hepatocytes damage. These molecular characteristics, coupled with the drugs' metabolization in the liver, heighten the risk of liver toxicity. Finally, we calculate the prediction accuracy of KANO and MotiL on SIDER for hepatotoxicity. The accuracy rates are 74.3% for KANO and 76.8% for MotiL, respectively.

L Activity cliff evaluation

L.1 Experimental setup

To rigorously evaluate the robustness of MotiL against small structural perturbations that cause large changes in molecular properties, we conducted an activity cliff (AC) experiment on the BBBP and SIDER datasets. The identification of activity cliff samples follows a widely accepted protocol in cheminformatics:

- Data source: We use the test set molecules from the BBBP and SIDER benchmark datasets.
- Similarity filtering: We compute pairwise Tanimoto similarity scores based on RDKit Morgan fingerprints. Only molecular pairs with a similarity score greater than 0.85 are retained, indicating high structural similarity.
- Label divergence: Among these similar pairs, we select those where the two molecules have different binary labels. These cases constitute activity cliff pairs, where slight structural differences lead to opposite bioactivities.
- Sample selection: From each AC pair, we extract individual molecules and evaluate whether the model correctly predicts their true label. We then compare MotiL and KANO on the subset of disagreement cases within this AC group.

This setting is intentionally challenging, as it tests the model's ability to distinguish subtle functional variations that significantly impact the biological activity.

L.2 Experimental results and case study

As shown in Fig. 4b of the main paper, the activity cliff experiment reveals that MotiL demonstrates stronger predictive robustness than KANO:

- On the BBBP dataset, among 15 AC disagreement samples, MotiL predicts 10 correctly (66.7%), while KANO predicts 5 (33.3%).
- On the SIDER dataset, MotiL correctly predicts 11 out of 19 (57.9%), whereas KANO predicts 8 (42.1%).

These results indicate that MotiL is more effective at resolving difficult cases involving high structural similarity but diverging functional outcomes, which are common in medicinal chemistry and toxicity analysis. To better illustrate this advantage, we present selected case studies in Supplementary Fig. 10 (BBBP) and Fig. 11 (SIDER), where each figure shows: 1. the SMILES strings of two molecules from the AC pair; 2. their binary labels (0 or 1); 3. their Tanimoto similarity score; 4. prediction outcomes from both MotiL and KANO.

Cases in BBBP (Supplementary Fig. 10): In the first three pairs, MotiL correctly predicts both molecules in each pair, whereas KANO misclassifies one molecule per pair. These examples illustrate

MotiL's stronger sensitivity to local functional group variations, such as substitutions between carboxyl and ester groups, or small changes in stereochemistry on steroid-like scaffolds (e.g., pair with similarity 0.851). In the last two pairs, both models show imperfect performance. Each model misclassifies one molecule per pair, indicating that certain structural subtleties remain challenging even for MotiL, despite its overall superior performance on activity cliffs.

Cases in SIDER (Supplementary Fig. 11): In the SIDER dataset, multiple AC pairs are shown with high similarity (e.g., 0.956 and 1.0) and opposite toxicity labels. In the first pair (Tanimoto similarity is 0.956), KANO gives identical predictions, while MotiL distinguishes the two correctly. It demonstrates that when the structural similarity is high but the relevant substructural cues are distinguishable, MotiL can succeed. In the remaining three pairs, the two models show mixed behavior: For each pair, one molecule is misclassified by KANO and the other is misclassified by MotiL. It reflects that both models struggle to some extent in resolving the subtle chemical cues that define the activity cliff boundary in these samples.

These results further support the finding that MotiL demonstrates better robustness to activity cliffs overall, while also highlighting specific cases where further improvement in motif-level discrimination is still needed.

M Fine-tuning details

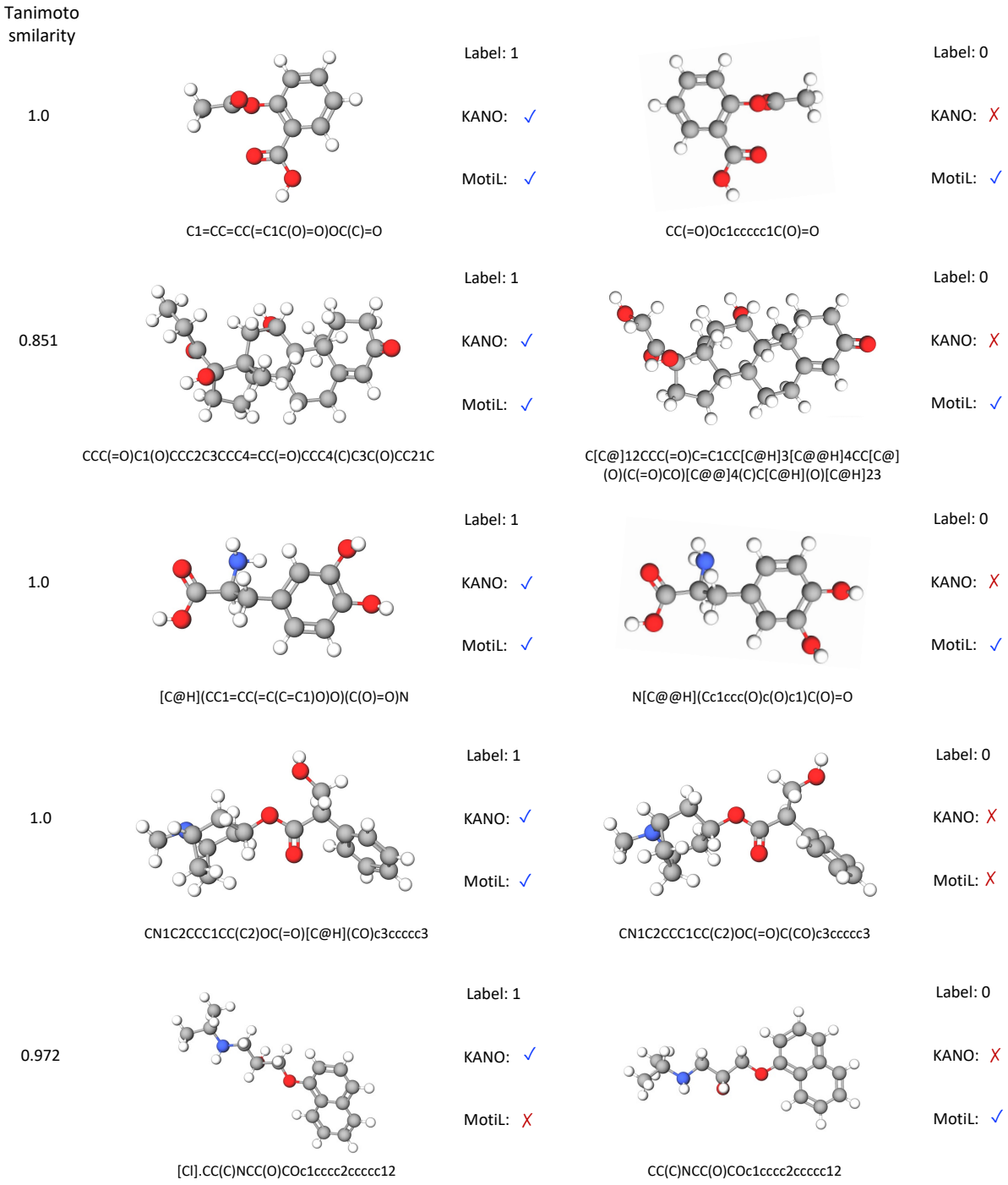
During the task-specific fine-tuning phase, we perform a random search for hyperparameters to identify the best-performing settings on the validation set, and then report the results on the test set for each dataset. The various combinations of hyperparameters used for small molecules and protein macromolecules are listed in Supplementary Table 7 and Table 8, respectively. In the small molecule experiments, we employ learning rate scheduling, specifically using step decay, where the learning rate decreases exponentially from lr_initial to lr_final over the course of the total training steps.

Name	Description	Range
epoch number	Total number of training epochs	{10, 20, 30, 40, 50, 100, 200}
lr_initial	Initial learning rate for the pre-trained GNN	{1e-5, 1e-4, 1e-3}
lr_final	Final learning rate for the pre-trained GNN	{1e-7, 1e-6, 1e-5}
batch_size	Batch size of training data	{16, 32, 64, 128, 256}
wd	Weight decay of the Adam optimizer	{1e-5, 5e-5, 1e-4}

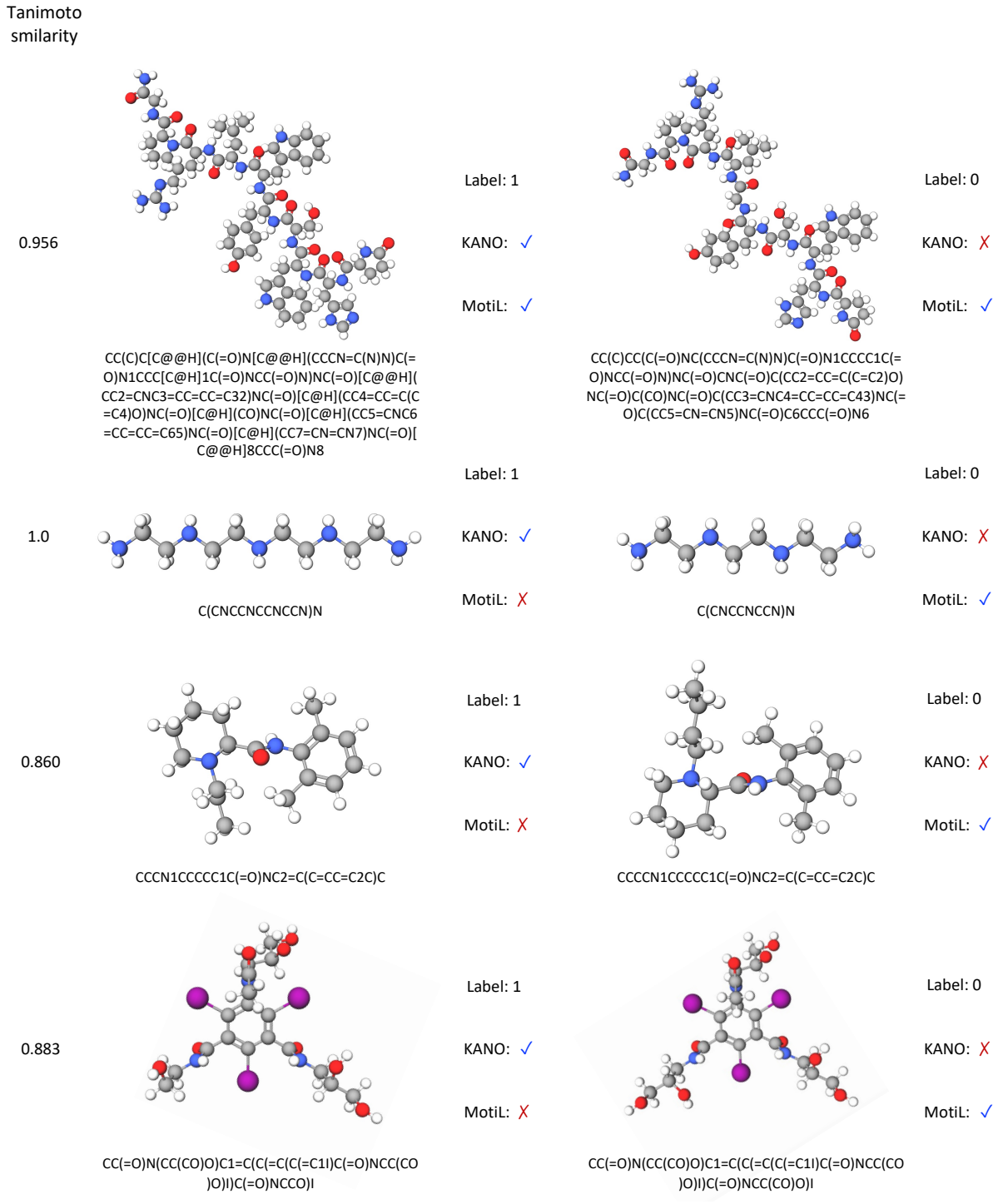
Supplementary Table 7 Hyperparameter descriptions and search ranges in the small molecule experiments.

Name	Description	Range
base_width	Bottleneck width in CDConv	{16, 32}
kernel_channels	Kernel channels in CDConv	{16, 24, 32}
batch_size	Batch size of training data	{32, 64}

Supplementary Table 8 Hyperparameter descriptions and search ranges in the protein macromolecule experiments.



Supplementary Figure 10 Prediction results for activity cliff samples on BBBP.



Supplementary Figure 11 Prediction results for activity cliff samples on SIDER.

References

[1] Zheng, G. *et al.* Layoutdiffusion: Controllable diffusion model for layout-to-image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 22490–22499 (2023).

[2] Cheng, S.-I., Chen, Y.-J., Chiu, W.-C., Tseng, H.-Y. & Lee, H.-Y. Adaptively-realistic image generation from stroke and sketch with diffusion model. *Proceedings of the IEEE/CVF winter conference on applications of computer vision* 4054–4062 (2023).

[3] Wu, T. *et al.* Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems* **36**, 39957–39974 (2023).

[4] Lin, Z. *et al.* Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. *Proceedings of the 40th International Conference on Machine Learning* 21051–21064 (2023).

[5] Oosterheert, W. *et al.* Molecular mechanism of actin filament elongation by formins. *Science* **384**, eadn9560 (2024).

[6] Heydenreich, F. M. *et al.* Molecular determinants of ligand efficacy and potency in gpcr signaling. *Science* **382**, eadh1859 (2023).

[7] Du, Y. *et al.* Machine learning-aided generative molecular design. *Nature Machine Intelligence* **6**, 589–604 (2024).

[8] Yang, L. *et al.* Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* **56**, 1–39 (2023).

[9] Du, Y. *et al.* Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. *Proceedings of the 40th International Conference on Machine Learning* 8489–8510 (2023).

[10] Huang, L. *et al.* A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications* **15**, 2657 (2024).

[11] Qiang, B. *et al.* Coarse-to-fine: A hierarchical diffusion model for molecule generation in 3d. *Proceedings of the 40th International Conference on Machine Learning* 28277–28299 (2023).

[12] Zhang, D., Tang, N. & Qu, Y. Joint motion deblurring and super-resolution for single image using diffusion model and gan. *IEEE Signal Processing Letters* **31**, 736–740 (2024).

[13] Pakornchote, T. *et al.* Diffusion probabilistic models enhance variational autoencoder for crystal structure generative modeling. *Scientific Reports* **14**, 1275 (2024).

[14] Yu, J., Zhang, X., Xu, Y. & Zhang, J. Cross: Diffusion model makes controllable, robust and secure image steganography. *Advances in Neural Information Processing Systems* **36** (2023).

[15] Chung, H., Kim, J. & Ye, J. C. Direct diffusion bridge using data consistency for inverse problems. *Advances in Neural Information Processing Systems* **36** (2023).

[16] Luo, R. *et al.* Diffusiontrack: Diffusion model for multi-object tracking. *The 38th AAAI Conference on Artificial Intelligence* 3991–3999 (2024).

[17] Zhang, J., Das, K. & Kumar, S. On the robustness of diffusion inversion in image manipulation. *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models* (2023).

[18] Zhao, H., Yang, X., Wei, K., Deng, C. & Tao, D. Unsupervised graph transformer with augmentation-free contrastive learning. *IEEE Transactions on Knowledge and Data Engineering* **36**, 7296–7307 (2024).

[19] Lee, N., Lee, J. & Park, C. Augmentation-free self-supervised learning on graphs. *The 36th AAAI Conference on Artificial Intelligence* 7372–7380 (2022).

[20] Xia, J., Wu, L., Chen, J., Hu, B. & Li, S. Z. Simgrace: A simple framework for graph contrastive learning without data augmentation. *Proceedings of the ACM Web Conference 2022* 1070–1079 (2022).

[21] Fang, X. *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* **4**, 127–134 (2022).

[22] Wang, Y., Wang, J., Cao, Z. & Farimani, A. B. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4**, 279–287 (2022).

[23] Fang, Y. *et al.* Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence* **5**, 542–553 (2023).

[24] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning* **70**, 1263–1272 (2017).

[25] Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling* **59**, 3370–3388 (2019).

[26] Song, Y. *et al.* Communicative representation learning on attributed molecular graphs. *Proceedings of the 29th International Joint Conference on Artificial Intelligence* 2831–2838 (2020).

[27] 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**, 8464–8476 (2019).

[28] Zhang, Z., Liu, Q., Wang, H., Lu, C. & Lee, C. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems* **34**, 15870–15882 (2021).

[29] Liu, S. *et al.* Pre-training molecular graph representation with 3d geometry. *The 10th International Conference on Learning Representations* (2022).

[30] Li, Y., Zhang, H. & Yuan, Y. Edge contrastive learning: An augmentation-free graph contrastive learning model. *Proceedings of the AAAI Conference on Artificial Intelligence* **39**, 18575–18583 (2025).

[31] Li, H. *et al.* Augmentation-free graph contrastive learning of invariant-discriminative representations. *IEEE Transactions on Neural Networks and Learning Systems* **35**, 11157–11167 (2024).

[32] Yan, P. *et al.* Empowering dual-level graph self-supervised pretraining with motif discovery. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 9223–9231 (2024).

[33] Zhang, R. *et al.* Motif-oriented representation learning with topology refinement for drug-drug interaction prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* **39**, 1102–1110 (2025).

[34] Yu, Z. & Gao, H. Molecular representation learning via heterogeneous motif graph neural networks. *Proceedings of the 39th International Conference on Machine Learning* 25581–25594 (2022).

[35] Ju, W. *et al.* Towards graph contrastive learning: A survey and beyond. *arXiv preprint arXiv:2405.11868* (2024).

[36] Shanehsazzadeh, A., Belanger, D. & Dohan, D. Is transfer learning necessary for protein landscape prediction? *NeurIPS 2020 Workshop on Machine Learning for Structural Biology* (2020).

[37] Rao, R. *et al.* Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems* **32**, 9686–9698 (2019).

[38] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *The 5th International Conference on Learning Representations* (2017).

[39] Velickovic, P. *et al.* Graph attention networks. *The 6th International Conference on Learning Representations* (2018).

[40] Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* **34**, 4046–4053 (2018).

[41] Baldassarre, F., Hurtado, D. M., Elofsson, A. & Azizpour, H. Graphqa: Protein model quality assessment using graph convolutional networks. *Bioinformatics* **37**, 360–366 (2021).

[42] Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. O. Learning from protein structure with geometric vector perceptrons. *The 9th International Conference on Learning Representations* (2021).

[43] Hermosilla, P. & Ropinski, T. Contrastive representation learning for 3d protein structures. *arXiv preprint arXiv:2205.15675* (2022).

[44] Zhang, Z. *et al.* Protein representation learning by geometric structure pretraining. *The 11th International Conference on Learning Representations* (2023).

[45] Fan, H., Wang, Z., Yang, Y. & Kankanhalli, M. S. Continuous-discrete convolution for geometry-sequence modeling in proteins. *The 11th International Conference on Learning Representations* (2023).

[46] Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *NeurIPS Workshop on Machine Learning for Molecules* (2020).

[47] Wang, S., Guo, Y., Wang, Y., Sun, H. & Huang, J. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* **19**, 429–436 (2019).

[48] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

[49] Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcao, A. O. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modeling* **52**, 1686–1697 (2012).

[50] Hartung, T. Toxicology for the twenty-first century. *Nature* **460**, 208–212 (2009).

[51] Richard, A. M. *et al.* Toxcast chemical landscape: Paving the road to 21st century toxicology. *Chemical Research in Toxicology* **29**, 1225–1251 (2016).

[52] Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sider database of drugs and side effects. *Nucleic Acids Research* **44**, D1075–D1079 (2016).

[53] Gayvert, K. M., Madhukar, N. S. & Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chemical Biology* **23**, 1294–1301 (2016).

[54] Subramanian, G., Ramsundar, B., Pande, V. & Denny, R. A. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of Chemical Information and Modeling* **56**, 1936–1949 (2016).

[55] Rohrer, S. G. & Baumann, K. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of Chemical Information and Modeling* **49**, 169–184 (2009).

[56] Riesen, K. & Bunke, H. Iam graph database repository for graph based pattern recognition and machine learning. *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition* 287–297 (2008).

[57] Delaney, J. S. Esol: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences* **44**, 1000–1005 (2004).

[58] Mobley, D. L. & Guthrie, J. P. Freesolv: A database of experimental and calculated hydration free energies, with input files. *Journal of Computer-aided Molecular Design* **28**, 711–720 (2014).

[59] Gaulton, A. *et al.* Chembl: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2012).

[60] Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society* **131**, 8732–8733 (2009).

[61] Ramakrishnan, R., Hartmann, M., Tapavicza, E. & Von Lilienfeld, O. A. Electronic spectra from tddft and machine learning in chemical space. *The Journal of Chemical Physics* **143**, 084111 (2015).

[62] Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling* **52**, 2864–2875 (2012).

[63] Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nature Communications* **12**, 3168 (2021).

[64] Wu, Z. *et al.* Moleculenet: A benchmark for molecular machine learning. *Chemical Science* **9**, 513–530 (2018).

[65] Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems* **33**, 12559–12571 (2020).

[66] Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry* **39**, 2887–2893 (1996).

[67] Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *The 7th International Conference on Learning Representations* (2019).

[68] Hu, W. *et al.* Strategies for pre-training graph neural networks. *The 8th International Conference on Learning Representations* (2020).

[69] van der Maaten, L. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research* **15**, 3221–3245 (2014).

[70] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[71] Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *The 2nd International Conference on Knowledge Discovery and Data Mining* **96**, 226–231 (1996).

[72] Walters, P. We need better benchmarks for machine learning in drug discovery (2023). URL https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html.

[73] Fang, C. *et al.* Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling* **63**, 3263–3274 (2023).